

Kurdistan Regional Government  
Ministry of Higher Education and Scientific Research  
University of Sulaimani  
College of Science



# **IMPROVED DETECTION TECHNIQUES FOR BREAST CANCER CLASSIFICATION USING MACHINE LEARNING**

A Dissertation

Submitted to the Council of the College of Science at the University  
of Sulaimani in Partial Fulfillment of the Requirements for the Degree  
of Doctor of Philosophy in Computer Science

**Image Processing**

By

**Srwa Hasan Abdulla**

*M.Sc. Information Technology (2011), University of JNTUH, India*

Supervisor (s)

**Prof. Dr. Ali Makki Sagheer**

**Assoc. Prof. Dr. Hadi Veisi**

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

وَعَلَّمَكَ مَا لَمْ تَكُنْ تَعْلَمُ وَكَانَ فَضْلُ اللَّهِ  
عَلَيْكَ عَظِيمًا

صدق الله العظيم

## Supervisor Certification

We certify that the preparation of this dissertation entitled “**Improved Detection Techniques for Breast Cancer Classification Using Machine Learning**” accomplished by **Srwa Hasan Abdulla**, was prepared under our supervision in the College of Computer Science and IT at the University of Anbar and the Faculty of New Sciences and Technologies at the University of Tehran in partial fulfillment of the requirements for the degree of Doctoral of Philosophy in **Computer Science (Image Processing)**.

Signature:

Supervisor: **Dr. Ali Makki Sagheer**

Title: **Professor**

Department of Computer Science,  
College of Computer Science and IT,  
University of Anbar

Date: 14 / 12 / 2022

Signature:

Co-Supervisor: **Dr. Hadi Veisi**

Title: **Associate Professor**

Department of Computer Engineering,  
Faculty of New Sciences and Technologies,  
University of Tehran

Date: 14 / 12 / 2022

---

## Certification of the Department

In view of the available recommendation, I forward this dissertation for debate by examining committee.

Signature:

Name: **Dr. Mustafa Ibrahim Khaleel**

Title: **Assistant Professor**

Department of Computer Science, College of Science, University of Sulaimani

Date: 14 / 12 / 2022

## Examining Committee Certification

We certify that we have read this dissertation entitled “**Improved Detection Techniques for Breast Cancer Classification Using Machine Learning**” prepared by **Srwa Hasan Abdulla**, and as the Examining Committee, we examined the student in its content and in what is connected with it, and in our opinion, it meets the basic requirements toward the degree Doctoral of Philosophy in **Computer Science (Image Processing)**.

Signature:

Name: **Fadhil Salman Abed**

Title: **Professor**

Affiliation: **Sulaimani Polytechnic University**

Date: 14 / 12 / 2022

Member

Signature:

Name: **Saman Mirza Abdullah**

Title: **Professor**

Affiliation: **University of Koya**

Date: 14 / 12 / 2022

Member

Signature:

Name: **Bayan Omar Mohammed**

Title: **Associate Professor**

Affiliation: **University of Human Development**

Date: 14 / 12 / 2022

Member

Signature:

Name: **Abbas Mohamad Ali**

Title: **Assistant Professor**

Affiliation: **Salahaddin University**

Date: 14 / 12 / 2022

Member

Signature:



Name: **Dr. Ali Makki Sagheer**

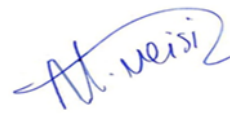
Title: **Professor**

Affiliation: **University of Anbar**

Date: 14 / 12 / 2022

Supervisor-Member

Signature:



Name: **Dr. Hadi Veisi**

Title: **Associate Professor**

Affiliation: **University of Tehran**

Date: 14 / 12 / 2022

Supervisor-Member

Signature:

Name: **Aree Ali Mohammed**

Title: **Professor**

Affiliation: **University of Sulaimani**

Date: 14 / 12 / 2022

Chairman

Approved by the Dean of the College of Science

Signature:

Name: **Dr. Soran Mohammed Mamand**

Title: **Assistant Professor**

Date: 14 / 12 / 2022

## DEDICATION

*I dedicate this dissertation to...*

*My Father's Soul*

*My Angel Mother*

*My Beloved Sisters & Brothers*

*All the people who supported me...*

## ACKNOWLEDGMENTS

First of all, I am thankful to Allah Almighty for his abundant grace and blessings that gave me strength, opportunity, and capability to complete this dissertation successfully..

I am honored to express my highest appreciation to my supervisor, **Prof. Dr. Ali Makki Sagheer**, for his marvelous supervision, invaluable support, encouragement, constant guidance, professional advice, patience, and profound understanding. I have been extremely lucky to work with such an experienced supervisor, who is also a nice and humble human being at the same time.

I would also like to extend my deepest gratitude to my co-supervisor, **Assoc. Prof. Dr. Hadi Veisi**, for his excellent supervision, great support, and encouragement, which helped me all the time of my research. He gave me great guidance, offered precious advice, and made constructive comments during my dissertation time, which made the accomplishment of this work possible.

I would like to heartily dedicate this dissertation to my father's soul, who took the lead to heaven before the completion of this work. I am extending my deepest gratitude to my family for allowing me to realize my own potential. All the support they have provided me over the years was the greatest gift anyone has ever given me for their consistent prayers, unconditional love, care, understanding, patience, and cooperation to follow my dreams. Without their support, I would not have made it this far in my PhD dream.

Additionally, I wish to express my special thanks to **Asst. Prof. Dr. Karzan Tofiq Mahmood**, the Dean of the College of Agricultural Engineering Sciences, for his collaboration and continued support. I would also like to acknowledge **Ammar Hamid Ali** for his help and support during this process.

I would also like to express my unique appreciation to all my beloved friends, especially to **Dalia Mohammad Toufiq**, who have always been there when I needed them. Finally, my thanks go to all the people who supported me in completing this dissertation, directly or indirectly.

*Srwa Hasan Abdulla*

## ABSTRACT

Breast cancer is one of the leading causes of mortality among women worldwide. Therefore, detecting breast cancer at an earlier stage helps to increase the cure rate and reduce mortality. Among the different methods for detecting and diagnosing breast cancer, mammography is a significant and popular technique. Manually reviewing mammograms takes time and is accompanied by human error. Therefore, computer-aided detection/diagnosis (CAD) systems are proposed to support radiologists in making accurate diagnosis decisions. In this dissertation, two methods are proposed for breast cancer detection using mammogram image processing and machine learning:

- The first method proposes a technique for detecting micro-calcifications in mammography images. In this method, an optimized region-growing (ORG) technique is presented to detect calcifications in more than one region of the breast, which cannot be determined by using the standard region-growing algorithm. Then, the Haralick-based features and support vector machine (SVM) classifier are used to distinguish between benign and malignant tissues. The experimental results on the curated breast imaging subset of the digital database for screening mammography (CBIS-DDSM) dataset obtained a sensitivity of 98.2%, a specificity of 100%, and an accuracy of 98.82%, which exceeds the performance of the reference methods.
- The second proposed method addresses two challenges in breast cancer detection. The first challenge is the existence of pectoral muscles, which are misidentified and result in a high percentage of false positives. To handle this challenge, a region-growing method is performed to isolate the pectoral muscles, then a suspicious region of interest (ROI) is segmented, and after that, the features are extracted from the segmented ROI. The second challenge

is the imbalance problem in the dataset. To ease this challenge, the synthetic minority oversampling technique (SMOTE) is used to obtain a balanced dataset. A random forest (RF) classifier is finally used to categorize the segmented regions as benign or malignant. The accuracy, sensitivity, and specificity of the experimental results on the mammographic image analysis society (Mini-MIAS) dataset are 97.1%, 95.8%, and 98.4%, respectively.



# TABLE OF CONTENTS

<b>ABSTRACT</b> .....	i
<b>LIST OF TABLES</b> .....	vi
<b>LIST OF FIGURES</b> .....	vii
<b>LIST OF ABBRIVIATIONS</b> .....	x
<b>1. Chapter One: Introduction</b> .....	1
1.1 Introduction .....	2
1.2 Literature Review .....	3
1.3 Problem Statement.....	13
1.4 Objective of the Study .....	14
1.5 Contributions .....	15
1.6 Dissertation Structure .....	16
<b>2. Chapter Two: Theoretical Background</b> .....	18
2.1 Introduction .....	19
2.2 Breast Anatomy .....	20
2.3 Breast Cancer.....	21
2.3.1 Breast Cancer Signs and Symptoms .....	22
2.3.2 Types of Breast Cancer .....	23
2.3.3 Breast Cancer Lesions .....	24
2.3.3.1 Calcifications.....	24
2.3.3.2 Mass .....	25
2.3.3.3 Architectural Distortions .....	27
2.3.3.4 Bilateral Asymmetry .....	27
2.3.4 Early Detection of Breast Cancer .....	28
2.3.5 Breast Cancer Stages .....	29
2.4 Medical Image Modalities .....	29
2.4.1 Mammography.....	30
2.4.1.1 Mammogram Projections .....	31
2.4.2 Other Modalities .....	32
2.4.2.1 Digital Breast Tomosynthesis (DBT).....	32
2.4.2.2 Contrast-Enhanced Mammography (CEM) .....	33

2.4.2.3	Breast Ultrasound (BUS) .....	34
2.4.2.4	Magnetic Resonance Imaging (MRI) .....	34
2.5	Digital Image Processing .....	36
2.5.1	Computer-Aided Detection/ Diagnosis (CAD) .....	38
2.5.2	Techniques of Digital Image Processing .....	38
2.5.2.1	Image Acquisition .....	40
2.5.2.2	Pre-processing .....	40
2.5.2.3	Segmentation .....	42
2.5.2.4	Features Extraction .....	46
2.5.2.5	Feature Selection .....	48
2.5.2.6	Classification .....	48
2.6	Evaluation Metrics .....	52
2.6.1	Confusion Matrix .....	53
2.6.2	Receiver Operating Characteristic (ROC) .....	54
2.7	Synthetic Minority Oversampling Technique (SMOTE) .....	55
2.8	Mammographic Databases .....	57
2.8.1	Mammographic Image Analysis Society (MIAS) .....	57
2.8.2	Digital Database for Screening Mammography (DDSM) .....	58
<b>3.</b>	<b>Chapter Three: Methodology</b> .....	<b>59</b>
3.1	Introduction .....	60
3.2	Image Processing Based CAD Systems .....	60
3.3	Proposed Method 1: Segmentation using K-means Clustering and Optimized Region-growing Technique .....	62
3.3.1	Data Acquisition .....	64
3.3.2	Pre-processing .....	64
3.3.3	Pre-segmentation (Denoising) .....	66
3.3.3.1	K-means Clustering .....	67
3.3.3.2	Eliminate Noise Objects .....	68
3.3.3.3	Recover the Important Area .....	69
3.3.4	Segmentation .....	70
3.3.5	Features Extraction .....	73
3.3.6	Support Vector Machine Classifier .....	78
3.4	Proposed Method 2: Pectoral Muscle Removal and Solving Data Imbalance Problem	79
3.4.1	Data Acquisition .....	82

3.4.2	Pre-processing.....	82
3.4.2.1	ROI Segmentation.....	83
3.4.2.2	Pectoral Muscle Removal.....	83
3.4.3	Segmentation .....	86
3.4.4	Features Extraction .....	86
3.4.5	Balancing the Data.....	88
3.4.6	Random Forest (RF) Classifier .....	90
<b>4.</b>	<b>Chapter Four: Results and Discussion .....</b>	<b>93</b>
4.1	Introduction .....	94
4.2	Evaluation Results of the Proposed Models .....	94
4.3	Proposed Model 1: Evaluation Results of the Segmentation using K-means Clustering and Optimized Region-growing Technique .....	95
4.3.1	Dataset .....	95
4.3.2	Preprocessing Results .....	96
4.3.3	Segmentation Results.....	96
4.3.4	Classification Results.....	99
4.4	Proposed Model 2: Evaluation Results of the Pectoral Muscle Removal and Solving the Data Imbalance Problem.....	104
4.4.1	Dataset .....	104
4.4.2	Pre-processing Results .....	105
4.4.3	ROIs Segmentation Results .....	105
4.4.4	Segmentation Results.....	106
4.4.5	K-Fold Cross-Validation.....	109
4.4.6	Classification Results.....	109
<b>5.</b>	<b>Chapter Five: Conclusions and Future Works.....</b>	<b>115</b>
5.1	Conclusions .....	116
5.2	Future Works .....	119
	<b>Publications .....</b>	<b>120</b>
	<b>References.....</b>	<b>121</b>
	<b>Appendices .....</b>	<b>133</b>
	Appendix A .....	133
	Appendix B.....	139

## LIST OF TABLES

Table 2. 1 Confusion matrix .....	53
Table 4. 1 The selection and distribution of samples from the CBIS-DDSM dataset.....	95
Table 4. 2 Comparison of experimental results for the proposed method.....	98
Table 4. 3 Confusion matrix for the proposed method.....	99
Table 4. 4 Comparison of the existing techniques with the proposed system.....	101
Table 4. 5 Mini-MIAS dataset description .....	105
Table 4. 6 Mini-MIAS dataset samples distribution before and after applying the SMOTE	110
Table 4. 7 Confusion matrix for the balanced dataset of the proposed method .....	110
Table 4. 8 Comparison of evaluation metrics with and without SMOTE.....	111
Table 4. 9 The performance measures of the proposed algorithm for the Mini-MIAS dataset .....	112
Table 4. 10 Comparison of the proposed method with the existing techniques.....	113

## LIST OF FIGURES

Figure 2. 1 The anatomy of the breast .....	21
Figure 2. 2 Sign and symptoms of breast cancer .....	22
Figure 2. 3 Types of breast cancer .....	24
Figure 2. 4 The calcifications in the breast .....	25
Figure 2. 5 Speculated mass on mammogram .....	26
Figure 2. 6 Screening mammogram .....	31
Figure 2. 7 MLO and CC projections .....	32
Figure 2. 8 Breast imaging .....	35
Figure 2. 9 Block diagram for the classification process .....	39
Figure 2. 10 Distribution of the Gaussian function values .....	41
Figure 2. 11 Machine learning approaches.....	50
Figure 2. 12 SVM schemes (A) Linear (B) Non-linear .....	51
Figure 2. 13 The mechanism of RF .....	52
Figure 2. 14 Schema of ROC curve and AUC: red line: a perfect classifier, blue curve: a great classifier, yellow line: a random classifier, and shaded area: AUC for the random classifier [129] .....	55
Figure 3. 1 System architecture for breast cancer detection.....	61
Figure 3. 2 Flowchart of the proposed method.....	63
Figure 3. 3 The preprocessing and pre-segmentation steps: (a) original image (b) Gaussian filter (c) applying k-means (d) erosion filter and (e) breast area retrieval .....	66
Figure 3. 4 The k-means clustering process: (a) original image (b) applying k-means .....	68
Figure 3. 5 The eliminate noise object process: (a) applying k-means (b) erosion filter (breast area mask).....	69
Figure 3. 6 Recover the important area process: (a) erosion filter (breast area mask) (b) breast area retrieval .....	70
Figure 3. 7 The segmentation process: (a) breast area retrieval (b) segmented area and (c) ROI .....	71
Figure 3. 8 (a) Angular second moment values of malignant and benign samples .....	77
Figure 3. 9 Flowchart of the proposed method.....	81
Figure 3. 10 Left pectoral muscle removal process: (a) subtracted image (b) defines the truncating triangle (c) pectoral muscle removal .....	85

Figure 3. 11 Right pectoral muscle removal process: (a) subtracted image (b) defines the truncating triangle (c) pectoral muscle removal .....	85
Figure 4. 1 The pre-processing and segmentation Process: a) Original image, b) Gaussian filter, c) Applying k-means, d) Erosion filter, e) Breast area retrieval f) Segmented area, g) Micro-calcifications (ROI) .....	97
Figure 4. 2 Comparison of experimental results for the proposed method .....	98
Figure 4. 3 ROC curve of the classification results .....	100
Figure 4. 4.a Sensitivity comparison of the existing techniques with the proposed system ..	102
Figure 4. 5 Breast region identification results: a) Original mammogram, b) Darkening the background, c) Erosion operation.....	105
Figure 4. 6 Breast region segmentation process: a) Original revised image, b) (step 1) breast segmentation, c) (step 2) breast segmentation.....	106
Figure 4. 7 Invert and subtract process: (a) Inverted image (b) Subtracted image.....	107
Figure 4. 8 Segmentation process: (a) Pectoral muscle removal process (b) Segmented image .....	107
Figure 4. 9 Segmentation of suspicious regions outputs: a) Original image , b) Binarized image , c) Morphological erosion, d) First step of the breast region segmentation, e) Second step of the breast region segmentation, f) Inverted image, g) Subtracted image, h) Pectoral muscle removal process, i) Segmented image.....	108
Figure 4. 10 A 5-fold cross-validation.....	109
Figure 4. 11 ROC curve of the classification result .....	111
Figure 4. 12 Comparison of evaluation metrics with and without SMOTE.....	112
Figure 4. 13 Comparison of the existing techniques with the proposed model.....	114
Figure A. 1 The main form of the first model .....	133
Figure A. 2 Open a mammogram image .....	133
Figure A. 3 Pre-processing step.....	134
Figure A. 4 Create a mask for erosion step .....	134
Figure A. 5 Erosion step .....	135
Figure A. 6 Pre-segmentation step (breast area retrieval) .....	135
Figure A. 7 Micro-calcification segmentation step .....	136
Figure A. 8 Feature extraction step for one sample.....	136
Figure A. 9 loading the dataset.....	137
Figure A. 10 Feature extraction step for the loaded dataset .....	137
Figure A. 11 Saving the extracted features.....	138

Figure A. 12 Classification step .....	138
Figure B. 1 The main form of the second model.....	139
Figure B. 2 Open a mammogram image step for one sample .....	139
Figure B. 3 Pre-processing (special thresholding) step .....	140
Figure B. 4 Noise-removing step.....	140
Figure B. 5 Left-Right image cropping step.....	141
Figure B. 6 Up-Down image cropping step.....	141
Figure B. 7 Inverting step .....	142
Figure B. 8 Subtracting step .....	142
Figure B. 9 Pectoral muscle removal step .....	143
Figure B. 10 Pre-segmentation (ROI) step .....	143
Figure B. 11 Segmentation step.....	144
Figure B. 12 Dataset loading step.....	144
Figure B. 13 Features extraction step for the loaded dataset.....	145
Figure B. 14 Minority splitting step .....	145
Figure B. 15 Applying SMOTE on minority samples step .....	146
Figure B. 16 Classification step.....	146

## LIST OF ABBRIVIATIONS

Abbreviation	Description
AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under the Curve
BI-RADS	Breast Imaging–Reporting and Data System
BSE	Breast Self-Exam
BUS	Breast Ultrasound
CAD	Computer-Aided Detection and Diagnosis Systems
CBE	Clinical Breast-Exam
CBIS-DDSM	Curated Breast Imaging Subset of DDSM
CC	Cranio-Caudal
CEM	Contrast-Enhanced Mammography
CESM	Contrast-Enhanced Spectral Mammography
CNN	Convolutional Neural Network
CPU	Central Processing Unit
DBT	Digital Breast Tomosynthesis
DCIS	Ductal Carcinoma In Situ
DDSM	Digital Database for Screening Mammography
DFO	Dragon Fly Optimization
DICOM	Digital Imaging and Communications in Medicine
DIP	Digital Image Processing
DL	Deep Learning
DT	Decision Tree
DWT	Discrete Wavelet Transform
ELM	Extreme Learning Machine
ERR	Error Rate
FD	Fractal Dimension
FFDM	Full-Field Digital Mammography
FFNN	Feed Forward Neural Network
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GA	Genetic Algorithm
GLCM	Gray-Level Co-Occurrence Matrix
GLRLM	Gray Level Run Length Matrix.



HE	High-Energy
KNN	K- Nearest Neighbour
LCIS	Lobular Carcinoma In Situ
LDA	Linear Discriminant Analysis
LE	Low-Energy
MC	Micro-calcification
MIAS	Mammographic Image Analysis Society
ML	Machine Learning
MLO	Mediolateral Oblique
MRI	Magnetic Resonance Imaging
MSE	Mean Square Error
MSVM	Multi-Class Support Vector Machine
NB	Naïve Bayes
NN	Neural Network
ORG	Optimized Region Growing
ORGSMN	Optimal Region Growing Segmentation with MobileNet
PCA	Principle Component Analysis
PM	Pectoral Muscle
PPV	Positive Predictive Value
PR	Positive Rate
PRC	Pixel Range Calculation
RBFNN	Radial Basis Functions Neural Network
RF	Random Forest
RGB	Red Green Blue
ROC	Receiver Operating Characteristics
ROI	Region of Interest
SL	Supervised Learning
SMOTE	Synthetic Minority Over-sampling Technique
SSL	Semi Supervised Learning
SURF	Speeded Up Robust Features
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
TPR	True Positive Rate
USL	Unsupervised Learning
WBCD	Wisconsin Breast Cancer Dataset
WDBC	Wisconsin Diagnostic Breast Cancer

# *Chapter One*

## *Introduction*

## 1.1 Introduction

Cancer is a group of disorders in which cells in the body alter and grow uncontrollably. These cells expand, push out healthy cells, and eventually form a lump or mass known as a tumor that may be either benign or malignant. Benign tumors are not cancerous because their cells seem normal, they grow slowly, and they do not invade neighboring tissues or spread to other regions of the body. Malignant tumors are cancerous which spread swiftly to other regions of the body. Most types of cancers that cause a death are breast cancer among women who have the disease [1]. Although males are equally susceptible to developing breast cancer, women over the age of 50 have the greatest risk as well as the highest incidence of the disease. The cause of the disease is unknown, and the reasons for the rise in incidence are also unknown. Furthermore, there has yet to be established a means for preventing its occurrence. As a result, lowering breast cancer mortality requires early identification and treatment. Different screening approaches and procedures are utilized in clinical settings to obtain physiological and functional medical images of the human body.

Medical imaging is one of the new medical treatment standards for illnesses including cancer, trauma, and many others [2]. Recent imaging technologies concentrate on combining user-friendliness with a high level of precision, enabling information to be retrieved quickly and accurately while increasing throughput. These innovative methods are cost-efficient and applicable to a wide range of therapeutic applications [3]. Digital breast tomosynthesis (DBT) scanning and mammography have introduced new features that enable new diagnostic capabilities and clinical applications. The growth of digital imaging resulted in a new generation of performance and speed, since it provided new data access and transmission possibilities and generated new amounts of data. Mammography is the standard method for breast cancer screening. Both analogue and digital mammography have been shown to lower breast cancer

mortality [4]. Mass lesions, asymmetries between images of the two breasts, and architectural deformation are some of the radiologic indicators of breast cancer. To correctly diagnose breast cancer at the earliest possible stage, all aspects affecting the collection, display, and interpretation of the mammogram must be adjusted and maintained throughout time [5]. Mammograms are breast radiography images that are used to detect early signs of breast cancer. These radiographic images reduce human error in detection, reduce diagnosis time, and increase diagnosis accuracy. However, this task could lead to mistakes as it relies on human vision and radiologists' experience to make the right decision. As a result of these limitations, the CAD was developed. Computer science, pattern recognition, artificial intelligence, and image processing technologies are all used in CAD [6].

This dissertation provides an overview of the techniques and methods that are used for breast cancer detection and classification, which may be separated into five stages: data acquisition, pre-processing, segmentation, feature extraction, and classification.

## 1.2 Literature Review

Machine learning (ML) is one of the primary subfields of artificial intelligence (AI) study that has seen fast development in recent years. There is a substantial amount of interest being shown in the use of ML in the medical field. From the very beginning, medical datasets have been analyzed using machine-learning algorithms that have been conceived and developed from an initial concept [7]. This section discusses a previous related works with respect to this dissertation.

**C. Varela et al. (2007)** proposed a computer-aided detection system for malignant breast masses in digital mammograms. Through use of the iris filter, suspicious regions were segmented using an adaptive threshold. Suspect regions were classified using characteristics derived from the iris filter output as well as

contour-related, texture, gray level, and image morphological features. A backpropagation neural network classifier was trained to decrease the amount of false positives. A test set of 66 malignant and 49 normal cases produced sensitivity of 88% and 94% for lesion-based and case-based evaluation, respectively, using free-response receiver operating characteristic analysis [8].

**Bhagwati and G. R. Sinha (2014)** provided a unique technique for performing mammographic feature analysis through tumor detection in terms of size and shape. The goal is to find out the abnormality in tumor tissues through three stages: preprocessing, segmentation, and post-processing. The first stage remove a noise with preprocessing, then the segmentation is used to detect the mass, and post processing is utilized to determine the benign and malignant tissue with the impacted region in the cancerous breast image. In addition, these processes determine the size of the tumor. This system achieved 96.5% of sensitivity, 89% of specificity, and 95.6% of accuracy [9].

**A. Jaleel et al. (2014)** proposed a model for detection of breast masses. The research describes an effective discrete wavelet transform (DWT) technique and a modified gray level co-occurrence matrix (GLCM) method for extracting textural features from segmented mammography. For classification, each tissue pattern is classified into benign and malignant masses. Using supervised classifiers, 148 mammography images from the mini mammographic image analysis society (Mini-MIAS) database were classified as benign or malignant masses. The radial basis functions neural network (RBFNN) was employed as a classifier. The proposed system has an accuracy of 94.6% for cancer detection from digitized screening mammograms [10].

**K. Kashyap et al. (2015)** proposed a model to enhance, segment, and classify abnormalities present in the mammogram. This study demonstrates a way for automatically identifying abnormalities in mammograms. A variety of image processing techniques were used prior to the discovery of abnormalities. Un-

sharp masking has been used to improve mammography. Discrete wavelet transform were applied before segmentation to ensure the filtered image produced an accurate result. Fuzzy-c-means with thresholding were used to segregate the suspect ROI. Tamura features, shape-based features, and moment invariants are used to detect abnormalities in mammograms by analyzing the segmented ROI. The Mini-MIAS dataset has been utilized to test the suggested methodology. 96.92% accuracy, 97.14% sensitivity and 96.67% specificity has been achieved using combination of fractal dimension with moments features [11].

**Arden S. et al. (2015)** suggested classifying mammogram images using Law's texture energy measure (LAWS) as a technique for extracting texture features. Images that are normal, benign and malignant are classified using artificial neural networks (ANN). The MIAS database is used to train information for the mammography classification model. The results demonstrate that LAWS offers more accuracy than other comparable methods like GLCM. While GLCM only offers 72% of accuracy for normal-abnormal classification and 53% of accuracy for benign-malignant classification, LAWS offers 93% of accuracy for normal-abnormal and 83% of accuracy for both benign and malignant [12].

**R. Rouhi et al. (2015)** proposed two automated methods for identifying benign and malignant masses in mammograms. The first proposed method uses an automatic region growing whose threshold is determined by a trained artificial neural network. In the second proposed method, a genetic algorithm (GA) accomplishes segmentation by determining the parameters of a cellular neural network (CNN). From segmented tumors, intensity, textural, and shape features are retrieved. GA is used to select appropriate features from a set of extracted features. ANNs are then utilized to classify the mammograms as benign or malignant in the following stage. Different classifiers (random forest,

naïve bayes, SVM, and KNN) are utilized to evaluate the performance of the proposed approaches. The proposed techniques were evaluated using MIAS and digital database for screening mammography (DDSM) databases. Sensitivity, specificity, and accuracy rates were calculated to be 96.87%, 95.94%, and 96.47%, respectively [13].

**W. Xie et al. (2015)** proposed a novel computer-aided diagnosis (CAD) system based on extreme learning machine (ELM) for the diagnosis of breast cancer. In the case of a mammographic image, interference is first eliminated during the preprocessing stages. The preprocessed images are then segmented using the proposed level set model. After that, a model of multidimensional feature vectors is created. Because not every feature vector contributes to performance improvement, feature selection is performed using a combination of extreme learning machine and support vector machine. An optimal subset of feature vectors is fed into the classifiers to distinguish between malignant and benign masses. All of the images utilized in this article are from publically available digital mammographic image databases; the first is MIAS database, and the second is DDSM database. The results show that the proposed CAD system outperforms SVM and support vector machine with particle swarm optimization in terms of sensitivity, specificity and accuracy. In the end, the proposed method has an average accuracy of 96.02% [14].

**F. Last et al. (2017)** proposed a straightforward and efficient approach of oversampling that was based on k-means clustering and SMOTE oversampling. The method prevents the formation of noise and successfully resolves imbalances both across and within classes. The suggested strategy improved classification outcomes when applied to training data that has been oversampled using the empirical findings of extensive tests including 71 different datasets. In addition, k-means and SMOTE routinely achieves better results than other

widely used oversampling algorithms. Python, which is a programming language, has been made available as a possible implementation [15].

**S. Punitha et al. (2018)** utilized Gaussian filtering for pre-processing tumor images. The initial seed points and thresholds are ideally created utilizing a swarm optimization method known as dragon fly optimization (DFO) in the automated detection method for breast masses that is suggested. GLCM and grey level run length matrix (GLRLM) techniques were used to extract features from the segmented images. Images are classified into two classes (i.e., benign and malignant) based on a feed forward neural network (FFNN) classifier. DDSM dataset for 300 images was used to evaluate the proposed system and the results achieved a sensitivity of 98.1%, and a specificity of 97.8% [16].

**L. Shen et al. (2019)** proposed a deep learning algorithm for detecting breast cancer using an "end-to-end" training technique that makes efficient use of training datasets with either full clinical annotation or with only the cancer labeling of the image. This method requires lesion annotations only during the initial training step, while subsequent phases just require image-level labels. The best AUC obtained on the DDSM are 0.88. On full-field digital mammography images from the INbreast, the best AUC was 0.95 [17].

**Prabhpreet k. et al. (2019)** presented the Mini-MIAS dataset of 322 images and used k-means based on speed-up robust features (SURF) to extract features. For classification, the deep neural network and multiclass support vector machine (MSVM) are utilized with ratio of 70% for training and 30% for testing. The result showed that the suggested automated deep learning (DL) technique employing K-mean clustering with MSVM has a higher accuracy rate than using a decision tree (DT) model. According to experimental findings, the suggested method's average accuracy rates for the three types of cancer (i.e., normal, benign, and malignant) are 95%, 94%, and 98% respectively [18].



**M. Mabrouk et al. (2019)** proposed an improved CAD system based on supervised classification that can be more accurate and faster than standard examination programs by using image-processing techniques such as preprocessing, segmentation, feature extraction, and classification stage. This work is based on the integration of shape, texture, and invariant moment features. Instead of employing only one type of feature in breast cancer classification, this integration generated a great result in terms of sensitivity and specificity. The integration system's accuracy reached 96% in the automated mode of ANN, with the best accuracy achieved by features based on invariant moments reaching 97% [19].

**V. Viswanath et al. (2019)** presented a CAD approach to identify and classify breast cancer tumors into three categories which are malignant, benign and normal from medical mammogram images. The authors utilized three machine learning algorithms for classifying breast tumors which are k-nearest neighbors (K-NN), random forest (RF), and support vector machine (SVM) with accuracy above 95%. This work also studied the effect of pre-processing stage before applying the selected classifiers that enhancing the performance of training and predicting the different classes [20].

**H. Azary and M. Abdoos (2020)** utilized the MIAS dataset, which was downloaded from the website <http://peipa.essex.ac.uk/info/mias.html>. The dataset consists of images for breast mammography and labeled images, with the labeled images used as a reference for evaluation. GLRLM features and static features were used for feature extraction. The extracted features are mean, variance, standard deviation, and absolute deviation. Finally, the breast tumor was segmented using a semi-supervised method. The training was applied based on two classifiers: SVM and Bayes. The results showed that the proposed system performed with a higher accuracy of 94.04% compared with supervised methods of 43.17% and 87.52%, respectively [21].

**M. Kamil and A. Jassam (2020)** proposed a feature extraction method for breast mass detection using mammography images. The Mini-MIAS database was used, and the features were extracted based on GLCM from the region of interest. The k-nearest neighbor (KNN) classification method was utilized, the distance metric was the Euclidean method, and the number of neighbors was equal to 10. It was shown that the optimal angle for distinguishing between abnormal and normal tissues is  $135^\circ$ . The best accuracy was observed with a KNN classifier equal to 86.1%, and the sensitivity was 92% [22].

**Ichrak K. et al. (2021)** presented a unique method for removing pectoral muscle (PM) from mammography. This method was built on the concepts of clustering, region, and edge. The suggested approach was evaluated using the Mini-MIAS database's digital mammography for implementation, testing, and verification. The DICE coefficient and the structural similarity measure were used to determine the quality of segmentation between segmented areas and the ground truth. When compared to other current strategies in the same context, the proposed strategy has proven to be more successful and superior with accuracy reached to 92.47% [23].

**M. Elsadig et al. (2021)** proposed a framework for early detection of breast cancer. The study utilized the Wisconsin breast cancer datasets, which contain 30 features with 699 numerical instances. The dataset has been improved by selecting the most relevant features and removing the redundant ones. The result features represent the input for seven classifiers: random forest, logistic regression, NB, ANN, SVM, KNN, and DT. The finding showed very high performance for the SVM classifier, which outperformed the others with accuracy above 97.4%. [24].

**Yassir A. et al. (2022)** improved the segmentation and classification processes for breast cancer diagnosis. In addition, the proposed approach enhanced the quality of the mammogram images and their pectoral muscle.

Several steps involved for image enhancement: firstly, the mammography images processed in three levels which are red, green, and blue. The consistency of the underlying information on morphological operations forms the basis of the second stage. Thirdly, PCA method was used to reduce the size of the images. The excision of the pectoral muscle employing a region growth approach constitutes the fourth phase. The final step involves enhancing the coherence of the image's different areas with a second order Gaussian Laplacian and an orientated diffusion filter to produce a substantially enhanced contrast image. The proposed approach tested using dataset contained 11,194 images and 700 images used to validate the proposed approach. This performance of the suggested technique showed that it can boost the computerized breast cancer detection method's diagnostic performance with accuracy reached to 97.9% [25].

**Nada F. et al. (2022)** proposed this study to enhance breast cancer diagnostic detection performance using CC and MLO view analysis. Instead of using single-view, an image processing framework for multi-view screening was applied to enhance the diagnostic outcomes. The framework presented in this paper included feature extraction, segmentation, and image enhancement. The steps of image quality enhancement are crucial because mammographic images' poor contrast frequently causes overlaps between cancerous and healthy tissue. The images were segmented using a texture-based method called first-order local entropy. The results of feature extraction were used to compute the radius and the area of likely malignancy. The accuracy of breast cancer identification utilizing CC and MLO images was 88% and 80.5% respectively. The suggested framework proved beneficial in the diagnosis of breast cancer, with detection outcomes and characteristics assisting doctors in treatment decisions [26].

**L. Singh and A. Alam (2022)** proposed a successful hybrid technique for locating and detecting troublesome mass regions in digital mammograms. The suggested hybrid methodology is created by combining a faster region-based

convolution neural network (Faster R-CNN) with an efficient pixel-based low level pre-processing technology. The R-CNN model has been showed as a powerful resource for health image analysis because of its speed. Faster R-CNN, on the other hand, has substantial difficulties when it comes to identifying breast cancers since the mass regions are partially masked by normal breast cells and pectoral muscles as well as noise. An effective mass identification strategy based on low-level preprocessing and a Faster R-CNN approach is proposed to handle the aforementioned challenge. Multiple performance metrics, including as sensitivities, precision, specificity, and area under the curve (AUC), are used to evaluate the proposed technique. Overall, the result had 95.2% sensitivity, 94.2% accuracy, 93.5% specificity, and a high 0.983 AUC [27].

**S. Sakib et al. (2022)** proposed an automated disease detection system that uses machine learning and deep learning techniques to assist medical professionals in diagnosing a disease, provide an efficient, reliable, and faster response in order to reducing the risk of death. The purpose of the study was to compare machine learning and deep learning approaches for breast cancer detection and diagnosis. For classification, 10-fold cross-validation was employed with, decision tree, random forest, k-nearest neighbor, support vector machine, logistic regression and deep learning technique. As a training dataset, the breast cancer wisconsin (diagnostic) dataset was used to assess and compare the effectiveness and efficiency of each algorithm in classification performance. According to the experimental results, random forest outperformed all other models with accuracy and F1-scores of 96.66 % and 0.963, respectively [28].

**D. Rose et al. (2022)** proposed a novel CAD for breast cancer identification and classification using optimal region growing segmentation with a MobileNet (CAD-ORGSMN) model. Pre-processing, segmentation, feature extraction, and classification are all steps of operations in the proposed CAD-ORGSMN model. To eliminate noise from mammography images, the suggested model employs a

wiener filtering based pre-processing technique. The CAD-ORGSMN model employs a glowworm swarm optimization (GSO) based region growing technique for image segmentation, with the glowworm swarm optimization algorithm creating the initial seed points and threshold values optimally. Furthermore, a MobileNet-based feature extractor is used, in which the MobileNet model's hyper parameters are optimally determined using a swallow swarm optimization (SSO) approach. Finally, a variation auto encoder is used as a classifier to assign class labels to the input mammography images. The Mini-MIAS dataset is used to assess the results. The proposed CAD-ORGSMN model produced results with an accuracy of 86.26% [29].

**Z. Sarvestani et al. (2022)** proposed two methods for image enhancement and highlighting of breast tissue micro-calcifications for the desired locations by regional ROI based on fuzzy system and Gabor filtering method to study the effectiveness and accuracy of automatic separation of images of breast tissue micro calcifications. The decision tree classification algorithm is used to classify the clusters of breast tissue micro calcifications. Then, for segmentation, samples suspected of micro-calcification are highlighted and masked, and tissue characteristics are extracted in the final stage. Following that, the benign and malignant types of segmented ROI clusters were determined using an artificial neural network. The suggested system is trained using DDSM database, and simulations are performed using the MATLAB software. The results of this training reveal an accuracy of 93% and sensitivity of more than 95% [30].

**Y. Almalki et al. (2022)** proposed three steps method, the database classification was the first step, while the second eliminated the pectoral muscle from the mammography image. To diagnose breast cancer, to find abnormal regions in a well-enhanced image, the third step used new image enhancing methods and a new segmentation module. The data contained 2892 images. In addition, the proposed approach is evaluated on 322 images from

MIAS database. For the Qassim health cluster dataset, the proposed method achieves an accuracy of 92%. The proposed method gives accuracy approximately 97% on the mammographic image analysis society database [31].

### 1.3 Problem Statement

Breast cancer is the most frequent type of cancer in women and the main cause of mortality among non-preventable cancers. Digital mammography is a common screening modality for breast cancer and is an effective method of detecting breast cancer even at an early stage [32]. Due to the morphological characteristics and ambiguity of the boundaries of the masses, it is a difficult task to recognize the normal or suspicious areas. Analyzing mammography during the diagnosis of a breast tumor has become a difficult process due to complexities such as low contrast, and the varied types of abnormalities could prevent radiologists from taking the right decisions [33]. Therefore, computer-aided detection/diagnosis (CAD) systems are proposed to help radiologists make accurate diagnoses. These systems commonly use digital image processing and machine learning techniques to reach this goal. In this dissertation, we have used these techniques for automatic breast cancer detection. Although several techniques are proposed for this aim up to now, still various challenges and limitations exist. This dissertation addresses some of them.

- Breast calcifications are calcium deposits within the breast tissue. They appear bright with hazy edges, varied forms and distributions, and low contrast on mammography screening. Because of their small size and random scattering, they are difficult to detect. Furthermore, their closeness to the surrounding dense tissues adds to the difficulty of identification due to the significant overlapping between normal and suspect tissues. The accurate identification of micro-calcifications is critical for the early detection of most breast cancer cases since their presence is significantly connected with breast cancer, especially when they arise in clusters. Therefore, the first problem that

this dissertation focuses on is detecting micro-calcifications in mammography images, especially in more than one region of the breast.

- The pectoral muscle is a high-density area seen on mammograms in the mediolateral oblique (MLO) views. However, its existence has an impact on the segmentation, feature extraction, and classification procedures that result in a high percentage of false positives. Due to the variations in size, shape, intensity, and contrast of pectoral muscles, it is difficult to correctly eliminate muscle areas from mammograms. Therefore, the second problem of this work is to isolate the pectoral muscles to reduce false positives.
- The third problem that this dissertation seeks to solve is the imbalance problem in the data sample, which is a common problem in medical machine learning. This is also a challenge in breast cancer detection because the datasets are mostly benign cases with a few malignant cases that should be carefully handled since class imbalance has a negative influence on classification accuracy.

#### **1.4 Objective of the Study**

Developing an efficient approach that helps in early detection is an important issue, and achieving this goal is based on using advanced computing techniques to improve diagnostics for breast cancer. Therefore, this dissertation aims to create an automated system that can classify digital mammogram images into benign and malignant. The objectives are ordered as follows:

- The dissertation presents a new detection technique for micro-calcifications in mammogram images. This work utilizes mammography for automated segmentation and classification processes to produce two types of classes, which are benign and malignant. Machine learning algorithms have been used in various medical fields; breast cancer detection is one of these fields. K-means is used for the segmentation process while SVM is used for classification.

- The work presents a diagnosis method to detect an abnormality in mammograms automatically, and then image-processing techniques are used to correctly segment the suspicious ROI with the removal of pectoral muscles (PMs) in order to improve the segmentation process and diagnostic accuracy. Thus, successful removal of PMs is vital to avoid false detection.
- To enhance the classification result, a random forest (RF) is applied and the SMOTE algorithm is used to accomplish better classifier efficiency, which presents new samples from the minority classes to get a balanced dataset. The Mini-MIAS dataset is used to evaluate the proposed method and is predominately composed of benign samples, with only a tiny percent of malignant samples.

Thus, the aim of this dissertation is to build an automatic system that is able to diagnose breast cancer with high accuracy and that assists the radiologists in making the right decision, which leads to saving the patient's life and reducing the mortality rate.

## 1.5 Contributions

The following are the main contributions of this dissertation, which are the proposed methods for dealing with the issues mentioned in the problem definition section:

1. An optimized region growing (ORG) method is proposed for identifying breast micro-calcifications (MCs) by utilizing two-level segmentation processes. To this end, the breast area from the image is isolated using k-means clustering, and then, multi-points (seeds) and thresholds are generated optimally depending on the color values of the image pixels. The first method is implemented, in which Haralick texture features and SVM classifier are utilized to identify benign and malignant tissues.
2. A segmentation process is proposed based on performing different operations to solve the pectoral muscle problem. To increase efficiency in feature



extraction and classification, a region-growing method is performed to isolate the pectoral muscles. After that, a suspicious ROI is segmented, and then the features are extracted from the segmented ROI. As another contribution to the second proposed method, the method also found a solution to the imbalance problem in the Mini-MIAS dataset by using the SMOTE technique to provide new samples from the minority classes to obtain a balanced dataset and achieve better classification efficiency using a RF classifier.

## 1.6 Dissertation Structure

This dissertation is structured as follows:

- **Chapter One:** reviewed the latest studies that have been published by the authors in the field of detecting and classifying breast cancer types. Based on these studies, this dissertation specifies the main problems that are facing the research community and radiologists. Therefore, the dissertation proposes a framework which is explained in the objective section and presents its contributions.
- **Chapter Two:** introduces machine learning techniques and tools that are required for better breast cancer diagnosis. Furthermore, this chapter presents the main components of digital image processing. Also, the chapter gives a background about the breast cancer, its types and its symptoms. Finally, the chapter presents a review regarding accuracy, confusion matrix, specificity, and sensitivity.
- **Chapter Three:** presents the proposed methodology for this dissertation which contains various steps to implement breast detection and diagnosis. These steps include image acquisition, pre-processing to remove unrelated noise, segmenting and clustering the tumor regions, and finally applying machine learning algorithms to train and testing models.

- **Chapter Four:** shows the results that are obtained by carrying out this work using the evaluation metrics such as accuracy, sensitivity, specificity, and the confusion matrix.
- **Chapter Five:** this chapter emphasizes the main conclusions of the research, highlighting the key achievements and limitations. In addition, the chapter discusses future development and research.

# *Chapter Two*

## *Theoretical Background*

## 2.1 Introduction

Cancer occurs when cells start growing in an improper manner; this may lead to breast cancer. Cells that have this disease often form a tumour, which may frequently be seen on an x-ray or mammography modalities. Breast cancer represents one of the foremost factors behind the death of women worldwide. Although it is more common in women, males may still have the disease. Hence, early diagnosis and detection increase the probability of recovery and reduce the mortality rate [34]. Currently, different modalities have been used for screening, detection, and diagnosis of breast cancer such as ultrasound, digital breast tomosynthesis (DBT), contrast enhanced mammography (CEM) magnetic resonance imaging (MRI), and mammography. Mammography is considered to be one of the most effective and important methods for early breast cancer detection. It has been verified as a reliable and essential screening technique by obtaining visual images of the internal structure of breasts using a low-energy procedure. However, examining large numbers of mammographic images manually takes time and there is an error rate of between 10% and 30% due to human factors [35]. In view of this, computer-aided detection and diagnosis (CAD) systems have been developed to aid radiologists in detecting mammographic lesions that may indicate the presence of breast cancer. These systems act only as a second reader, and the final decision is made by the radiologist. The CAD system based on image processing has become an essential technique for the early diagnosis of breast cancer. Images obtained during mammography always include both the breast and the non-breast regions of the body. As a consequence of this, there are usually aspects in the background, particularly those related to the area of the pectoral muscles, noises, and artifacts that can influence the performance of a CAD system. Therefore, different preprocessing approaches have been suggested to enhance the quality of mammograms, allowing for them to be analyzed with a higher degree of

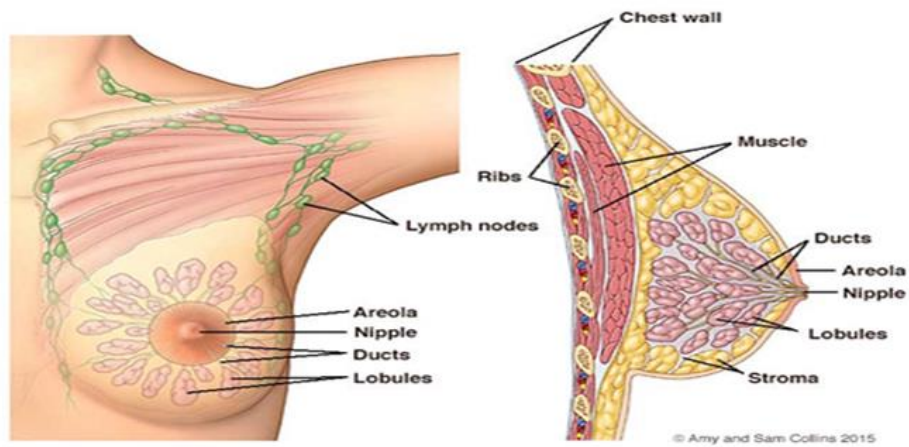
precision. Image processing requires consequent steps to accurate detection of breast cancer. These steps are as follows: image pre-processing to remove any noise, marker, background, and breast segmentation. The next step is determining the region of interest of the infected area based on some algorithms, and afterwards, the most distinct features are extracted from the target image. Finally, these images are classified according to various machine learning methods into different classes [36]. Machine learning algorithms such as SVM and RF are used to classify medical images into malignant and benign.

This chapter provides an introduction to the various machine learning strategies and technologies that are necessary for improved breast cancer detection. In addition, the chapter shows the background information on breast cancer, including its forms as well as its symptoms. Moreover, the primary aspects of digital image processing are discussed in detail throughout this chapter. Afterwards the chapter presents a review regarding accuracy, confusion matrix, specificity, and sensitivity. In the end, the chapter contained main datasets that are available publicly on the web.

## 2.2 Breast Anatomy

The breast is the tissue that lies on top of the chest (pectoral muscles), extends between the second and sixth ribs [37]. The breast is composed of glandular tissue, which is responsible for the production of milk, as well as fatty tissue. Milk is generated in lobules, which are arranged in lobes having 15-20 sections. Milk then travels via a network of small tubes known as ducts. These tubes connect to form bigger ducts that finally leave the skin in the nipple, where the milk is transported through the duct tubes. The darkened region of skin that surrounds the nipple is referred to as the areola. Additionally, the breast includes nerves that are responsible for feeling, blood arteries, lymphatic vessels, and lymph nodes [38]. Cancer can originate from any section of the breast; however, most breast cancers in females initiate from the ducts or lobules and

subsequently expand to other body parts via blood and lymph channels [39]. Figure 2.1 illustrates the anatomy of a breast.



**Figure 2. 1** The anatomy of the breast [40]

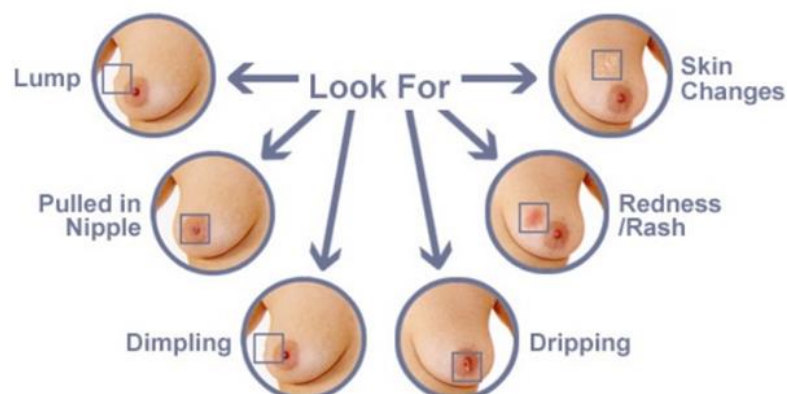
### 2.3 Breast Cancer

Cancer starts originating from cells, which are the smallest unit of all tissues and body organs, including the breasts. It is caused due to cell mutations, anarchic segmentation, and replication, or aberrant cell alterations. Usually, when cells undergo aging or destruction, they die and are replaced by new ones. In certain conditions, when this mechanism goes wrong, the cell develops the ability to continue dividing without being controlled or ordered, resulting in the formation of a mass of tissue known as a lump or a tumor [41]. Malignant tumours and benign tumours are the two categories that fall under the term of breast cancer. Malignant tumours are the most harmful kind of tumours since they include cancer cells that have the ability to infect the tissues that are located nearby. On the other hand, benign tumours develop regionally, they are larger than malignant ones, and they do not infiltrate the tissues of surrounding organs. Breast cancer is a kind of cancerous tumour that develops from cells that are found in the breast [42]. Cells that are not malignant are referred to as normal cells or benign cells. The malignant tumour is made up of cancerous cells, and this kind of cancer is referred to as in situ cancer [43]. Glandular tissue, which is

more prominent in the upper outer quadrant of the breast, is often where breast cancer first begins to form in a patient. Once it has spread into the tissue, a tumour might be classified as malignant. The site of the first proliferation of cancer cells is what establishes the subtype of breast cancer, which may be ductal or lobular [44].

### 2.3.1 Breast Cancer Signs and Symptoms

Symptoms lead doctors to diagnose breast cancer when screening tests show negative. Breast cancer presents signs and symptoms on the human body, which enables it to be identified at an early stage [42]. Most cases of breast cancer are identified by females in the form of a lump in the breast, which is generally painless. Breast pain has been recorded as the initial symptom only in a small percentage of individuals. Additional symptoms that occur less often include thickening, swelling, redness, nipple discharge, skin irritation, and distortion in shape as shown in Figure 2.2. Because many of these symptoms are also frequently observed in benign breast diseases, the predictive value of symptoms for diagnosing breast cancer is limited, and additional diagnostic investigations are typically required [45].



**Figure 2. 2** Sign and symptoms of breast cancer [46]

### 2.3.2 Types of Breast Cancer

Types of breast cancer depend on the part of the breast that is affected and the kind of cell. Ductal carcinoma in situ (DCIS) and lobular carcinoma in situ (LCIS) are the most common types. Carcinomas are tumors that start in the epithelial cells that line organs and tissues all over the body.

#### 1. In situ cancers

- **Ductal carcinoma in situ (DCIS)** this type of breast cancer begins inside the milk ducts and it classifies as a non-invasive or pre-invasive breast cancer.
- **Lobular carcinoma in situ (LCIS)** is a region of abnormal cell development in the lobules, which are glands that produce milk and are located at the end of the breast ducts.
- **Invasive (infiltrating)** it is a type of cancer that expands into neighboring breast tissue.

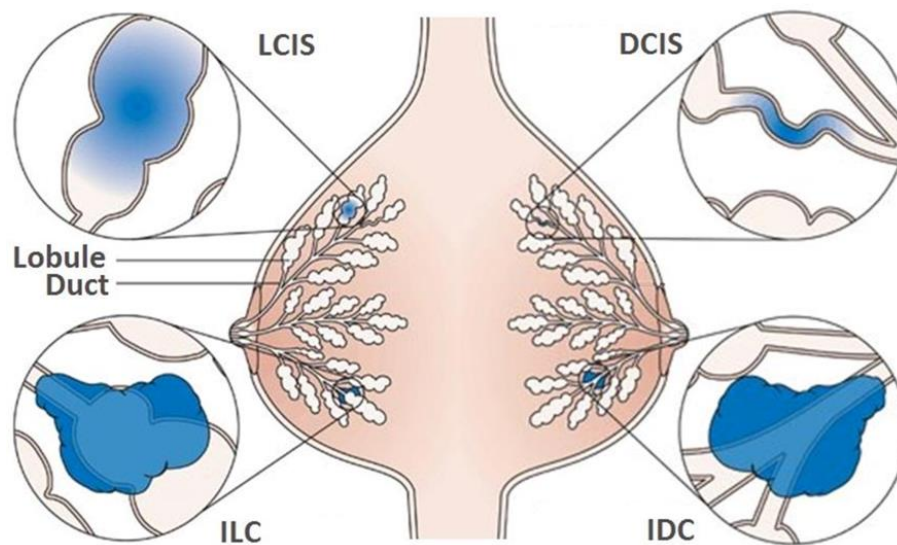
#### 2. Less common types of breast cancer

- **Inflammatory breast cancer** this type of cancer is categorized as an aggressive form of breast cancer that grows spreads rapidly.
- **Paget disease of the nipple** it is a rare form of breast cancer. It starts in the breast ducts and spreads to the skin of the nipple and then to the areola.
- **Phyllodes tumor** this type is rarely found that develops in the connective tissue (stroma) of the breast. They are classified most as benign, except some types which are malignant (cancer) [41].

It is currently unclear what causes breast cancer, and there is no reliable strategy to prevent the disease from developing in the first place. Breast cancer has a high mortality rate and a high incidence rate. The environmental and genetic factors that contribute to the development of this illness have both been the focus of research. However, there is inadequate evidence to support the ideas that connect family history, alcohol, poor eating, genetic mutations, pollution,



and other factors in its development. Breast cancer is a progressive illness that develops across the many phases of cellular development; thus, the timing of breast cancer detection is very important. Cancers are treatable by a variety of methods including chemotherapy, radiation treatment, and surgery. Every one of these therapies is determined by the kind of cancer and the area of the body in which the cancer is located [44]. Figure 2.3 illustrate the types of breast cancer.



**Figure 2. 3** Types of breast cancer [47]

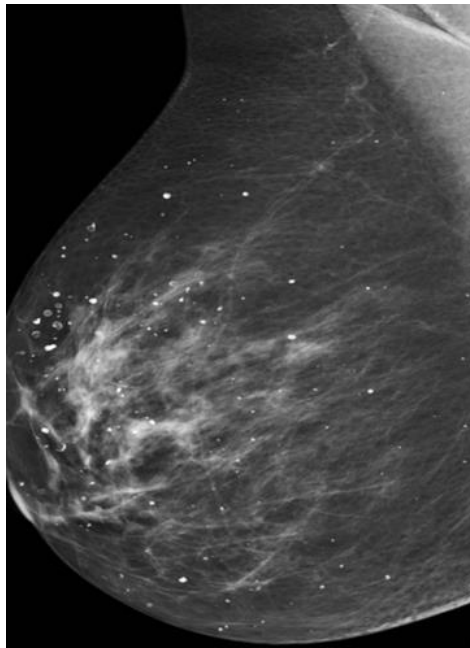
### 2.3.3 Breast Cancer Lesions

There are distinctive lesions that characterize breast cancer that include calcifications, masses, architectural distortions and bilateral asymmetry.

#### 2.3.3.1 Calcifications

Macro-calcifications and micro-calcifications are two different types of calcifications, which are calcium deposits in the breast. On a mammography, macro-calcifications show as large white spots dispersed irregularly across the breast; they are non-cancerous cells. Micro-calcifications are clusters of calcium deposits in the breast tissue that appear as a bright white spots of varied sizes and forms. Typically, micro-calcification is regarded as the main indicator of

early breast cancer or a marker of pre-cancerous cells already present [48]. Malignant micro-calcifications develop into masses that are angular, irregularly formed, tiny, grouped, varied in size, and oriented in a branching fashion. As opposed to malignant instances, benign cases' micro-calcifications are often larger, more rounded, fewer in number, more evenly distributed, and more uniform in size and shape [49]. Figure 2.4 shows the calcifications in the breast.

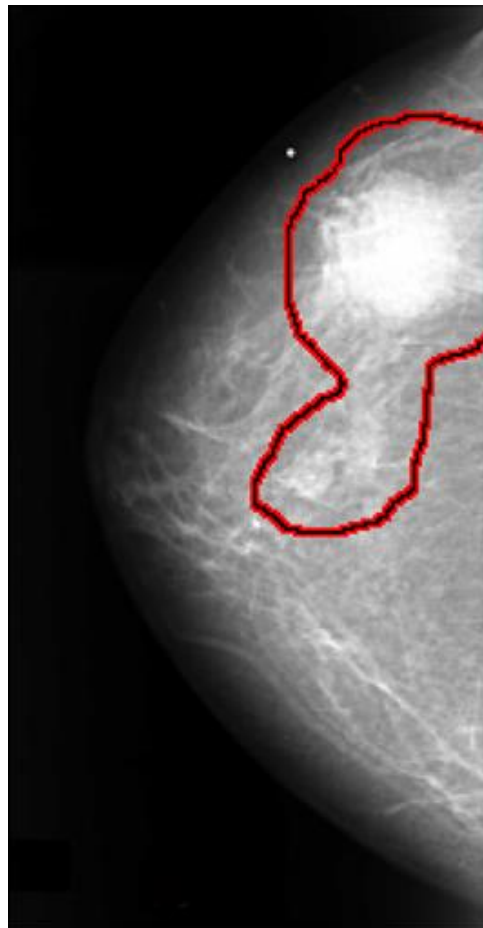


**Figure 2. 4** The calcifications in the breast [49]

### 2.3.3.2 Mass

A mass is a lump that forms in the breast. The majority of the tumours are non-cancerous (or benign), but a few may be cancerous (or malignant) and pose a life-threatening risk if left untreated. A breast mass may be detected via breast self-examination or a standard clinical evaluation. Lumps may or may not be uncomfortable and may be accompanied by nipple discharge or skin changes [50]. In general, mass form might be circular, oval, lobular, or irregular, with confined to speculative edges. Figure 2.5 shows the speculated mass in mammogram.

When a mass is discovered, it may be difficult to determine if it is benign or malignant; nonetheless, there are distinctions in the characteristics of both benign and malignant masses regarding their form and texture. The texture of benign masses is usually smooth and distinct, and their forms are frequently close to those of spheres. Malignant masses, on the other hand, are asymmetrical, and the lines delineating their borders are often hazy. A mass with a regular shape is more likely to be benign, while a mass with an irregular shape is more likely to be malignant [51].



**Figure 2. 5** Speculated mass on mammogram [52]

### 2.3.3.3 Architectural Distortions

Architectural distortion is defined as the uneven appearance of breast tissues in a random manner with no mass discoveries. Architectural distortion can be benign or malignant, depending on the look and shape of the distortion [53]. Nearly 6% of abnormalities seen on screening mammograms are attributable to architectural distortion, making it the third most prevalent mammographic appearance of non-palpable breast cancer after mass and calcification. Architectural distortion may be subtle and presents in a variety of ways, making it more difficult to detect than calcification or an obvious lump on mammography. Architectural distortion may be the initial sign of breast cancer and is often seen in retrospective analyses of false-negative mammography results [54].

### 2.3.3.4 Bilateral Asymmetry

Bilateral asymmetry occurs when the left and right breasts vary in their external look in the respective mammographic scans. Asymmetry is defined as the existence of increased breast tissue volume or density without a distinguishable mass, which might include the presence of more notable ducts in one breast in comparison to the relative portion of the other breast [42].

Asymmetry in the breast tissue is a finding made in relation to the identical location on the opposite breast. There is no focal mass, no deformed architecture, no central density, and no related breast calcifications. Asymmetrical breast tissue is seen in around 3% of results of breast screening mammography. Only a limited number of women having asymmetry in their breast tissue will be recommended for a biopsy, out of which merely a small percentage of them will report a diagnosis of breast cancer [55].

### 2.3.4 Early Detection of Breast Cancer

The best way to treat breast cancer is via early diagnosis. Breast cancer does not initially show any signs. As a result, it is advised to use screening techniques to find breast cancer early. The recognizable physical symptom that may be felt as the growth of the breast tumour increases is a lump that is not painful. Some less frequent signs of cancer development include secretion from the nipple and any changes in the breast's size, shape, or colour. A medical professional should check the breast for the existence of these alterations [43].

Treatment is more successful if the cancer is detected early and kept confined, thus, the five-year life expectancy rate reaches 99%. Therefore, early identification is essential for reducing mortality. Mammograms, clinical breast examinations, and breast self-exams are early detection tools. Comparative to imaging, breast examinations are safer and less expensive. A breast self-exam (BSE) is conducted by the patient, while a clinical breast exam (CBE) is performed by a medical professional. In less-developed nations where imaging and CBEs are less accessible, BSEs are often the sole option for early breast cancer identification. Even in the United States, BSEs are beneficial for women who are younger than 40 or 50 years of age, the age at which mammography is suggested (according on the recommendation) and who have a greater risk of breast cancer death owing to race or family history. However, the current standards regulating CBEs and BSEs are problematic. The American Cancer Society does not advise CBEs or BSEs to women exposed to the average risk of breast cancer anymore since research has not shown obvious advantages in the circumstances where mammography screening is accessible and awareness is high. Further, they also advise women to stay aware of their breasts and report the alterations to their healthcare practitioners [56].

### 2.3.5 Breast Cancer Stages

Breast cancers evolve through stages over time. In the localized stage of breast cancer, a tumour grows inside the breast. Eventually, the cancer cells move via the lymphatic system to the lymph nodes, and the disease enters its regional stage. In the distant phase, cancer cells move through the circulation to distant organs, such as the bones, lungs, brain, and liver. In clinical settings, the staging technique utilized by clinicians to choose the most suitable therapy is far more comprehensive. Classical anatomic staging is based on the TNM classification, where T represents tumour size, N represents lymph node metastases, and M represents distant metastasis. Each of the three classifiers is assigned a score, which is then summed to generate a comprehensive breast cancer stage from 0-IV (including subgroups) (A, B, C) [57].

When the T, N, and M categories are combined, five distinct phases may be identified. Stage 0 refers to in situ carcinomas and paget disease, stage I (IA, IB) is associated with localized tumours, stages II (IIA, IIB), and III (IIIA, IIIB, IIIC) are associated with regional metastasis and/or large tumours, and stage IV refers to tumours with distant metastasis. Women whose breast cancer has not spread beyond the breast have a greater chance of surviving the illness. Survival chances are better for women whose breast cancer has been confined [58].

### 2.4 Medical Image Modalities

Several imaging modalities, including mammography, ultrasonography, MRI, DBT, and CEM, may be employed to make medical images. In addition to examining the anomalies that indicate the presence of breast cancer, these imaging modalities identify a variety of other breast problems. All methods have distinct benefits and drawbacks that must be considered by the healthcare provider when deciding on a course of therapy [59].

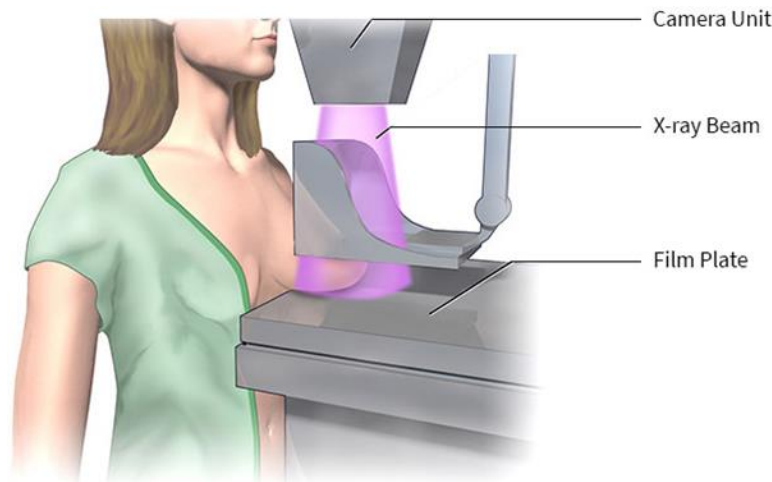
In the following subsections, a review of imaging modalities that are used to detect breast cancer is presented in more details.

### 2.4.1 Mammography

Mammography is the most significant technique for diagnosing and identifying breast cancer. It is an imaging technique for breast evaluation that provides information about breast shape, anatomy, and diseases. Early identification of breast cancer is crucial for the successful treatment of this illness. This process is identical to conventional X-rays, with the exception that minimal dosages are employed, resulting in great contrast, resolution, and low noise. The breast is susceptible to ionizing radiation, hence it is preferable to utilize the lowest dosage of radiation compatible with high image quality [60].

During an X-ray mammogram, each breast is normally compressed between two plates while X-ray images are captured in two separate planes. This is done while the treatment is being performed. In between each set of images, the patient will be requested to switch positions. Both craniocaudal (CC) and mediolateral oblique (MLO) views are the usual views that are often taken [48]. When passing through the breast, the X-rays encounter a variety of tissues that absorb the energy in a unique manner. The final image will have a different level of brightness as a direct consequence of this discrepancy [61].

Additionally, the only diagnostic technique that can detect micro-calcifications and masses which are the most crucial signs of malignant cancer is mammography. However, there are certain circumstances in which the radiologists are unsure of whether the abnormality is benign or malignant, leading to the utilization of a second call-back mammogram or other types of exams like ultrasound or MRI [48]. Figure 2.6 show the screening mammogram.



**Figure 2. 6** Screening mammogram [62]

### 2.4.1.1 Mammogram Projections

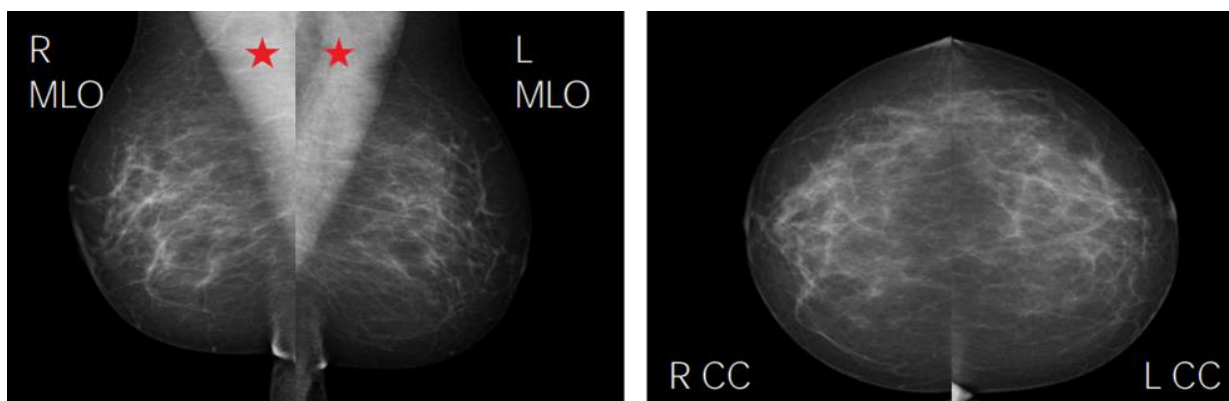
Mammography is a technique that is well recognized and used by radiologists for the purpose of detecting and diagnosing breast cancer at an earlier stage [63]. During this examination, a total of four images are acquired. Two of these images relate to the right breast, while the other two correspond to the left breast. These images are produced using the projections craniocaudal (CC) and mediolateral oblique (MLO) views. The combination of CC and MLO images helps to enhance breast tissue visibility and raises the chance of determining the presence of non-palpable breast cancer earlier in the diagnostic process. During the examination, the radiologist will combine the information obtained from these two views in order to increase the chance of identifying a priori regions with abnormalities [64].

The CC view is used to examine information from top to bottom, whereas the MLO view examines data from the side view. The MLO image of the mammography might be challenging to interpret because to the increased area of the pectoral muscle mass tissue, its complicated shape, and its structural volume. On the other hand, the pectoral muscle is an area that is thick and stands out in mammograms. It does not supply any information that is useful in any way. In



addition to this, it has an effect on the process of segmentation, feature extraction, and classification, which ultimately results in a high proportion of false positives [65].

Pectoral muscles are the triangle-shaped area that resides on one side, either the left or right upper corner of a mammogram. This part of the image includes the pixels that are the brightest overall. It seems as if the pectoral muscles have almost the same density as the dense tissues that are of interest in the image. Therefore, the identification of the pectoral muscles and their subsequent removal play a significant role in the elimination of the tumour cells [66]. Figure 2.7 represents the MLO and CC projections.



**Figure 2. 7** MLO and CC projections [67]

## 2.4.2 Other Modalities

### 2.4.2.1 Digital Breast Tomosynthesis (DBT)

This technique is a subset of mammography in which the X-ray tube spins at a restricted angular angle ( $15-60^\circ$ ) from the compressed breast tissue in order to obtain 3D breast information. DBT images are created by repeatedly exposing the breast tissue to light from a variety of directions, and then the resulting data is reconstructed into half-millimetre slices. This strategy has been found in a number of trials to result in an increase in the patient's radiation dosage of 20%.

However, the cancer detection rate is increased by around 15–30%, and the recall rate is decreased by approximately 15–20%. The primary benefit of tomosynthesis is the identification of masses and lesions that may not be detected in traditional mammography owing to overlap with thick breast tissue. This is one of the reasons why tomosynthesis is preferred over conventional mammography. Tomosynthesis has a high sensitivity and a low rate of producing false-positive detections at the same time. Consequently, the staging of breast cancer is the most approach. Although DBT is more effective in detecting non-calcified lesions than conventional mammography, classified lesions also have outcomes that are comparable to or even better in DBT than in conventional mammography. In conclusion, the most significant drawback of this method is being less sensitivity when it comes to the identification of micro-calcifications [68].

#### **2.4.2.2 Contrast-Enhanced Mammography (CEM)**

This innovative imaging method, also known as contrast-enhanced spectral mammography (CESM), uses a dual-energy methodology to identify breast cancer. After the injection of iodine contrast medium, contrast-enhanced spectral mammography is carried out using high-energy (HE) and low-energy (LE) acquisitions to acquire the recombined images of bilateral breasts. While the HE image shows post-contrast improved mammograms by employing the K-edge effect of iodine to assess tumor neovascularity, the LE image reveals morphological information comparable to two-dimensional (2D) digital mammography. These advantages have made CESM an effective alternative imaging method for the early diagnosis of breast cancer. Previous studies have demonstrated that the use of CESM, in comparison to conventional 2D mammography, significantly increased the positive rate, accuracy, and sensitivity for breast cancer detection [69]. In addition, this technique decreased the mortality of breast cancer because this recombined image could eliminate

the fibrous glands of the breasts and overcome the overlapping of normal breast tissues. For the detection of breast cancer in dense breasts, CESM has been shown to be more accurate than mammography [70].

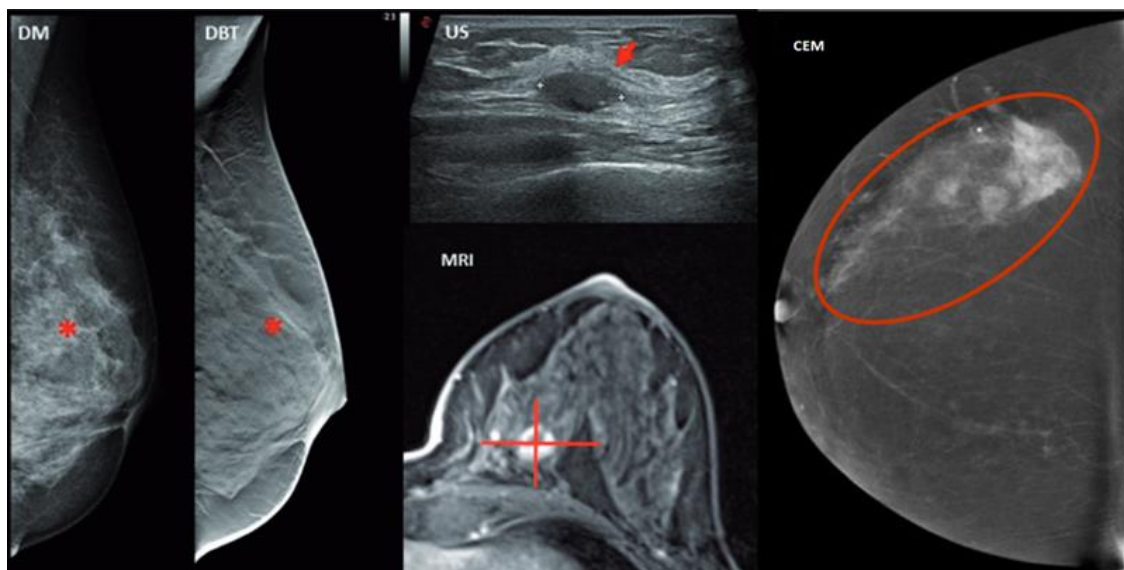
### **2.4.2.3 Breast Ultrasound (BUS)**

Ultrasound, commonly known as sonography, is an imaging technique that transmits and transforms high-frequency sound waves sent through the breast into images. There is no radiation involved in the image capture process. Because young women's breasts are often thick, mammograms may not detect cancer in them. In such situations, ultrasonography may be useful. However, ultrasonography is often used in conjunction with mammography. If a mammogram or physical examination reveals an abnormality, ultrasonography is the best approach to determine whether the abnormality is solid (such as a benign fibroadenoma or cancer) or fluid-filled (such as a benign cyst). It is unable to identify calcifications or determine whether a solid mass is malignant. A cyst cannot be adequately identified with a physical examination alone. Ultrasound of the breast may also be used to guide a biopsy needle into some breast lesions [71]. Ultrasound screening has a limited sensitivity for identifying impalpable tumours, which is one of its disadvantages. The recorded images must be tagged by the radiologist according to breast location and orientation. Any little variation in beam direction may cause a portion of the breast to be missed or imaged twice, making breast ultrasonography an operator-dependent technique [72].

### **2.4.2.4 Magnetic Resonance Imaging (MRI)**

MRI is a non-invasive medical technology used for diagnosing and monitoring the treatment of medical disorders. The MRI technique employs strong and powerful magnets, radio frequency pulses, and a computer to create high-quality images of soft tissues and interior body components. The images generated by the MRI enable clinicians to analyse various bodily components

and detect the presence or absence of illness in an organ [73]. MRI has been used to diagnose breast cancer because it is the most sensitive technique for detecting breast diseases and has excellent soft tissue imaging capabilities. MRI has also been used to measure the size of the cancer and look for metastasized tumours in women who have previously been diagnosed with breast cancer. MRI has been used to precisely identify and quantify tumors that are less than or equal to 2 cm in size [74]. However, MRI is time-consuming and expensive, with a cost that is more than 10 times that of mammography. Its limited resolution restricts its use to micro-calcifications and extremely tiny lesions [75]. The selection of these techniques is depends on the patient's condition, stage, age, and breast tissue density. Hybrid imaging approaches seem to be a viable method for enhancing breast cancer detection. Figure 2.8 illustrates the breast imaging.



**Figure 2. 8** Breast imaging [76]

## 2.5 Digital Image Processing

The most popular and practical method of communicating or sharing information is via images. They display spatial information that enables one to identify them as objects and convey information about the locations, sizes, and connections between items. Approximately 70% of the information acquired by humans is in the form of images [77]. An image refers to a two-dimensional function, where  $x$  and  $y$  are spatial coordinates, and the intensity or gray level of the image defines as an amplitude of ( $f$ ) at any pair of coordinates ( $x, y$ ). The image is known as digital image when  $x, y$ , and the intensity values of ( $f$ ) are all finite, discrete quantities. When the computer processes these types of images, the field refers to digital image processing. The digital image consists of finite number of elements called pixels and each one has a value and a location in the image [78]. There are four types of digital images:

- **Binary Images:** are the most basic sorts of images that can have two values, usually black and white, or "0" and "1". Because only one binary digit is required to accurately represent each pixel in a binary image, this kind of image is also known as a 1-bit per pixel image. The most common usage for these kinds of images is in computer vision applications.
- **Gray-Scale Images:** Images that only include one colour are said to be monochrome and refers to the grayscale. They simply provide information about the brightness and do not reveal any details about color. The number of various brightness levels that may be achieved is directly proportional to the number of bits that are employed for each individual pixel. The standard image has 8 bits of data per pixel, which gives us the ability to have 256 distinct degrees of brightness (gray) between 0 and 255.
- **Color Images:** Color images can be represented as three-band monochrome visual data, with each band representing a distinct color. The brightness details of every spectral band are the real information included in data

obtained from digital images. Typical color images are expressed as RGB images in red, green, and blue colors. Keeping the 8-bit monochrome format as a guide, the comparable color image would contain 24-bits per pixel (bpp) with 8 bits each reserved for all three RGB color bands [79].

- **Spectral Imaging:** Multiple bands from the electromagnetic spectrum are used in spectral imaging. While a typical camera only records the visible spectrum's red, green, and blue (RGB) wavelength bands, spectral imaging includes a wide range of methods that go beyond RGB. The infrared spectrum, visible spectrum, ultraviolet spectrum, x-rays, or some combination of the aforementioned may be used for spectral imaging.

The process of acquiring an image, performing any necessary pre-processing steps, isolating (segmenting) the particular region and defining characteristics in a format that is understandable by a computer is referred to as digital image processing [78].

Applications for digital image processing have grown dramatically in a variety of sectors, including pattern recognition, biometrics, medical image processing, etc. Digital image processing includes acquisition, preprocessing, extracting (segmenting) the individual areas, describing features in a form suitable for a computer, and recognizing those features. These steps are called digital image processing. This growth is due to the exponential expansion of internet usage and the fast progress of digital communication technology. As a result, methods for digital image processing are made easier in order to provide information more quickly and accurately. The healthcare sector has been searching for cutting-edge medical practices and procedures that will integrate with technology in terms of computing and advancement in hardware resources since health is so important. Digital image processing methods might aid radiologists in this effort by assisting them in identifying and detecting illnesses. Thus, the processing of medical images and the identification of anomalies from

those images has grown more dependent on digital image processing methods. Accordingly, image-processing-based CAD systems have emerged as an intriguing research issue in the field of medical image processing. CAD systems use computers to assist medical practitioners in making diagnoses from medical images [80].

### **2.5.1 Computer-Aided Detection/ Diagnosis (CAD)**

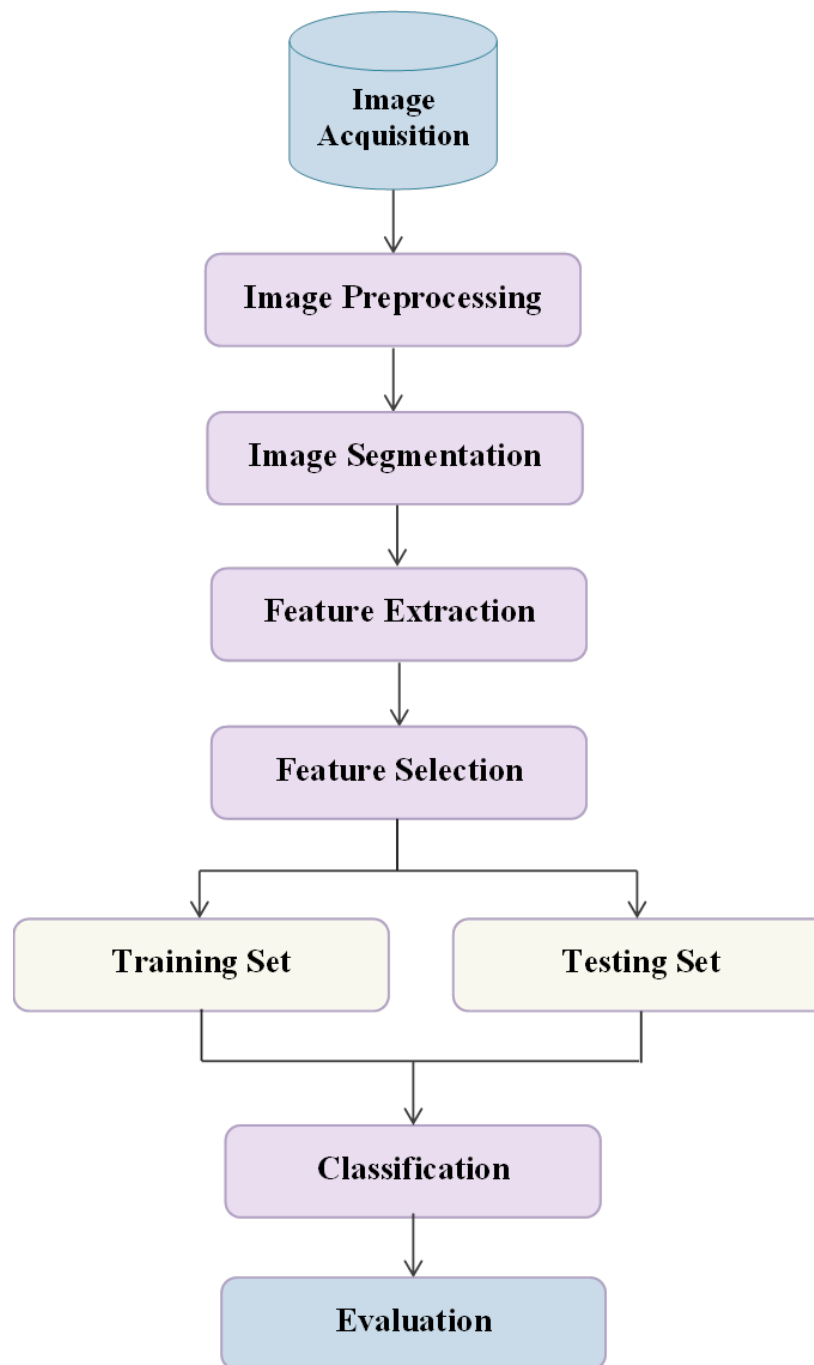
CAD systems are used to enhance the image quality, which aids in appropriately interpreting medical imaging and processing the images to emphasize the portions that desired. A variety of principles from artificial intelligence (AI), computer vision, and medical image processing are all included in the technology known as CAD. Finding abnormalities in the human body is the primary use of CAD technology. Cancer affecting multiple organs like breasts, lungs, prostate, and colon, bone metastases, congenital heart defect, abnormal brain diagnosis, and Alzheimer's disease are all diagnosed with CAD systems [36].

CAD techniques are classified into two types, namely CADe (computer-aided detection) and CADx (computer-aided diagnosis). CADe involves the use of computer analysis to determine whether or not a certain instance has a specific illness (the target disease), such as breast cancer. On the other hand, CADx aids in the evaluation of the detection process. The doctor or other medical professional manages the patient in both cases and makes the final diagnosis using techniques of digital image processing [81].

### **2.5.2 Techniques of Digital Image Processing**

Digital image processing has been developed to automate breast cancer classification into benign and malignant. This system can help doctors find and discriminate tumors in the tissues. Selecting the appropriate system for breast cancer diagnosis requires a better understanding of the contents of cancer

images [82]. The block diagram for the classification process is presented in Figure 2.9. The figure illustrates the five main processing stages for classifying breast cancer into benign and malignant. Complete descriptions of these steps are introduced in detail in the next subsections.



**Figure 2.9** Block diagram for the classification process [82]



### 2.5.2.1 Image Acquisition

It is the fundamental step of digital image processing. Under image acquisition the image is given in digital format [83]. There are two sources for breast cancer imaging: The first source is imaging modalities such as mammography, digital breast tomosynthesis (DBT), contrast-enhanced mammography (CEM), breast ultrasound (BUS) and magnetic resonance imaging (MRI) [84]. More details about these types of modalities are described in a section 2.4. In the contrast, the second type of sources is datasets. There are various datasets for breast cancer tumours that are available globally, but the most popular ones are digital database for screening mammography (DDSM) [85], mammographic image analysis society (MIAS), Wisconsin breast cancer dataset (WBCD) [86], and Wisconsin diagnostic breast cancer (WDBC). More details about DDSM and Mini-MIAS datasets are described in a section 2.8. [87].

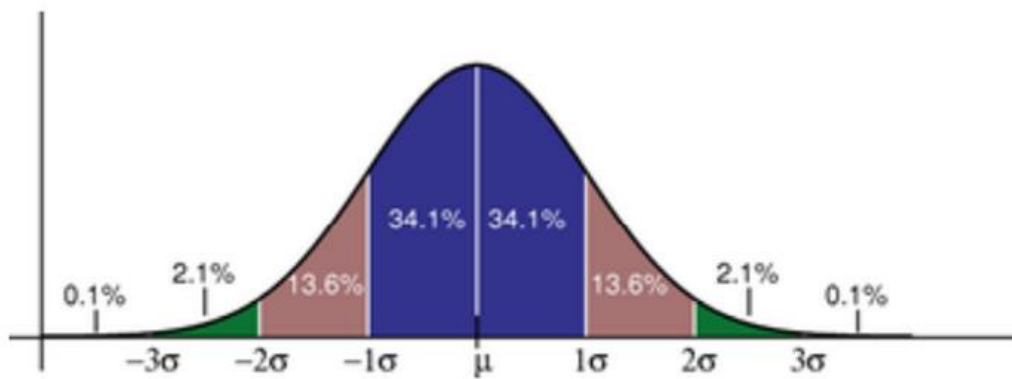
### 2.5.2.2 Pre-processing

It is essential and valuable phase, employed prior to any method of image processing to achieve the correct and reliable accuracy [88]. In the field of image processing, a pre-processing step is a procedure that is conducted before any additional processing. The purpose of this phase is to either enhance an image features or improve image quality by removing undesired distortions. The precision of the preprocessing that comes before image processing stages like segmentation, feature extraction, feature selection, and classification is very important to the overall outcome of those following steps. Pre-processing is often required for clinical images due to the presence of undesirable qualities such as inhomogeneity, poor contrast, and noise that cannot be recognized. When these difficulties are present in medical images and have an impact on analysis, pre-processing may assist in suppressing them. In pre-processing, a wide variety of methods, including human correction and mathematical

operations, augmentation, and the elimination of noise, are used [89]. Filtering is a process that must be applied to the image so that the effects of noise, also known as interference, may be removed. This allows for an improvement in the image's overall visual quality [90]. One of these filters is Gaussian filtering, which can be utilized to blur and remove any noise in the images. For one dimension, the Gaussian function is expressed as follows:

$$G(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \quad (2.1)$$

Where  $\sigma$  is the standard deviation of the distribution [91]. The Gaussian function's standard deviation has a significant impact on its behaviour. The values between +/- make up 68% of the total, whereas two standard deviations from the mean (blue and brown) make up 95%, and three standard deviations (blue, brown, and green) make up 99.7%. This is critical for creating a Gaussian kernel with a fixed length as shown in Figure 2.10.



**Figure 2.10** Distribution of the Gaussian function values [91]

When dealing with images, the two-dimensional Gaussian function is required [92]. This is a sum of two 1D Gaussian functions (one for each direction) and may be calculated as follows:

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (2.2)$$

### 2.5.2.3 Segmentation

Image segmentation is a critical and difficult image processing technique in which the inputs are images and the outputs are the properties retrieved from those images [93]. Image segmentation refers to the process of splitting an image into comparable structural components, which includes the detection and partition of target regions. Segmentation is a key function and the first critical stage in image processing that must be completed effectively before moving on to other tasks such as feature extraction and classification. Segmentation plays a significant function in medical imaging because it allows the identification of anatomical structures and other areas of interest [94]. There are various techniques that are used to apply for segmentation on the selected image such as threshold-based, region-based, and edge-based techniques.

- **Threshold-based Segmentation**

Image thresholding is a straightforward type of image segmentation. It is a method for converting a grayscale or full-color image to a binary image. It is useful to be able to distinguish between the image regions that correspond to the interested objects and the image regions that correspond to the background in many vision applications. Thresholding is frequently used to accomplish segmentation based on the varying intensities or colors in the foreground and background areas of an image. A grayscale image is often used as the input to a thresholding procedure, and the result is a binary image with only black and white pixels indicating the segmentation. Normally, black pixels correlate to the background and white pixels to the foreground, although this can be inverted. A single variable known as the intensity threshold determines segmentation in basic applications. This threshold is applied to each pixel in the image. If its value exceeds the threshold value, the pixel is set to white in the result; if it is less than or equal to the threshold, the pixel is turned to black in the output [95].

A number of thresholding techniques have been previously proposed using global and local techniques. In local thresholding methods, various threshold values are applicable to distinct parts of the image, whereas, in global approaches, only a single threshold value applies to the whole image. The value is determined by the location of the pixel to which the thresholding is performed.

In image segmentation, the threshold method is considered to be one of the most essential procedures. It can be expressed as a following formula:

$$T = T[x, y, p(x, y), f(x, y)] \quad (2.3)$$

Where  $T$  refers to the value that defines the threshold, the  $x$  and  $y$  are the coordinates of the location at which the threshold value is determined. The points of  $p(x, y)$  and  $f(x, y)$  on the grayscale image are the pixels [96]. Threshold image  $g(x, y)$  is defined as follow:

$$g(x, y) = \begin{cases} 1 & \text{if } f(x, y) > T \\ 0 & \text{if } f(x, y) \leq T \end{cases} \quad (2.4)$$

- **Region-based Segmentation**

Region-based techniques divide neighboring pixels with the same or comparable properties into separate areas. In contrast to edge-based segmentation techniques, regions are composed of connected pixels that are produced depending on preset criteria (such as color, texture, or intensity). Region growth, region splitting, and region merging procedures are the three categories into which region-based approaches are divided. Growing regions require manually chosen seed points. As long as a pixel meets the requirements for that region class and no boundaries are identified, a region is expanded by looking at and adding nearby pixels to that region class. Region splitting and merging of algorithms work in the opposite direction of region growth, which divides the entire image into regions in the top-down pattern. These isolated

regions are disconnected at random, but the elements within them are more homogeneous than the entire image. Regions may be divided into sub-areas, which can then be combined to build new regions with more appropriate regions [97].

- **Edge Based Segmentation**

Edge-based segmentation techniques are a structural method for locating the edges between diverse areas, with a sudden change in intensity. The edge-based segmentation approach is effective in images with high contrast and no noise [98]. Edge detection is the method most often employed in digital image processing. In image processing, particularly computer vision, edge detection deals with the localisation of significant gray level tones in the image and the recognition of the physical and geometrical features of scene objects. It is a key technique that identifies the contours and borders of an image's objects and backdrop. Edge detection is used for object detection, which has several applications including medical image processing and biometrics, among others. Edge detection is a hot topic in the study because it allows for more sophisticated image analysis [99]. Edge detection is very beneficial for image segmentation, reconstruction, and interpretation, tracking, and data extraction challenges [100].

- **Clustering (K- means Clustering)**

Clustering techniques attempt to identify similarities within the data and divide it into a set number of groups or clusters. Clustering is an unsupervised learning technique in which a clustering algorithm learns from existing data without being instructed. K-means clustering is the most well recognized clustering algorithm and is one of the most straightforward and widely used unsupervised learning methods, particularly in data mining and statistics [97]. As a partitioning method, its objective is to generate groupings of data points depending on the variable  $k$ , which represents the number of clusters.  $K$  must be

specified before execution begins. K-means employs an iterative refinement technique to generate its final grouping depending on the number of clusters specified by the user and the data set. Initially, k-means randomly selects  $k$  as the mean values of  $k$  clusters, also known as centroids, and finds the closest data points of the selected centroids to build  $k$  clusters. Afterwards, the method repeatedly recalculates the updated centroids for each cluster until it converges on a single optimal value. K-means clustering would be appropriate for low-dimensionality numerical data since numerical data is utilized to calculate the mean value [101]. The K-means clustering technique is a prominent algorithm that operates on a variety of data kinds, including medical images, text, etc. [102].

- **Morphological Operation**

Morphological image processing (or morphology) refers to a class of image processing algorithms that focus on the geometry of image characteristics. Typically, morphological treatments are used to eliminate flaws produced during segmentation [103]. During this stage, regions, and neighborhoods are compared, while each region is indexed with a binary image. To analyze a tiny form or template, a morphological approach known as structural pixels or structuring elements is applied. The structuring pixels are placed in various spots in the image and then matched with the pixels in a surrounding area. The procedure then begins by inspecting the pixel to determine if it "fits" within the neighborhood or whether it "hits" or overlaps the neighbourhood [104].

The two principal morphological operations are dilation and erosion. Dilation causes things to grow, potentially filling minor gaps and linking disparate items. Erosion reduces the size of items by eroding their boundaries. These operations may be tailored to a specific application by selecting the integral component (a rectangular array of pixels with the values 1 or 0), which controls the pattern of dilation or erosion in the objects. The erosion of  $A$  by  $B$  is defined as:

$$A \ominus B = \{a \in A \mid a + b \in A, b \in B\} \quad (2.5)$$

Where  $A$  is the image with pixels  $a$ , and  $B$  is the structuring element with elements  $b$ . And the dilation of  $A$  by  $B$  is defined as follows:

$$A \oplus B = \{c \in A \mid c = a + b, a \in A, b \in B\} \quad (2.6)$$

Where  $A$  is the image with pixels  $a$ ,  $B$  is the structuring element with elements  $b$ , and  $c$  represents the new pixels in  $A$  after dilation [105].

### 2.5.2.4 Features Extraction

Feature extraction refers to a technique that minimizes in number of features through generating new set of attributes with the same information of the old ones. Working with a huge dataset with hundreds or thousands of features without extracting the most ones that represent the actual observations about given variables could lead to overfitting in a model of machine learning (ML). Therefore, applying this technique reduces the risk of overfitting and increases the performance of the ML model. In other words, the purpose of features extraction is to discard the original features by proposing new ones that summarize most of the information of the old characteristics in the dataset. In addition, this technique increases the speed of training, accuracy, and enables visualization [106]. Examples of feature extraction technique are textural features. Texture is an important feature that can be utilized to identify objects or regions of interest in an image. The most common texture descriptors in image analysis are haralick texture features. Haralick's textural features are based on gray-level co-occurrence matrices (GLCM) [107]. The texture of an image can be measured using the co-occurrence matrix, which can take into account the image's intensity, grayscale values, or various color dimensions. Because co-occurrence matrices are often large and sparse, multiple matrix metrics are often used to obtain a more useful set of features [108].

**Cross correlation coefficient:** It is a method for objectively comparing two time series and determining where the best match occurs. It may also identify any repeated patterns in the data. A correlation coefficient is calculated to determine how well the values of one series predict the values of another. The sequences are then exchanged and the procedure is repeated. The range for the correlation coefficient of the time series value is from -1.0 to +1.0 [109]. The formula can be illustrated down as follows:

$$\text{Cross Correlation Coefficient} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \quad (2.7)$$

**Pearson correlation coefficient:** It is usually referred to as the pearson coefficient. For the purpose of analysing the correlation, the two variables, which are denoted by the labels  $x$  and  $y$ , are laid out on a scatter plot. The Pearson coefficient may be stated numerically in the same way as a linear regression correlation coefficient, with values ranging from (-1 to 1). The result of a full positive association between two or more variables is a value of +1. This may occur in a number of different ways. If there is a positive correlation between two variables, it is likely that those variables will continue to move in the same direction. A value of -1, on the other hand, denotes the existence of the perfect negative link. If one variable increases, then the other variable should decrease as a result of the variables being negatively related, as shown by negative correlations. A score of 0 indicates that there is no relationship [110]. The following is the formula for calculating the Pearson correlation coefficient:

$$\text{Pearson Correlation Coefficient} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (2.8)$$



### 2.5.2.5 Feature Selection

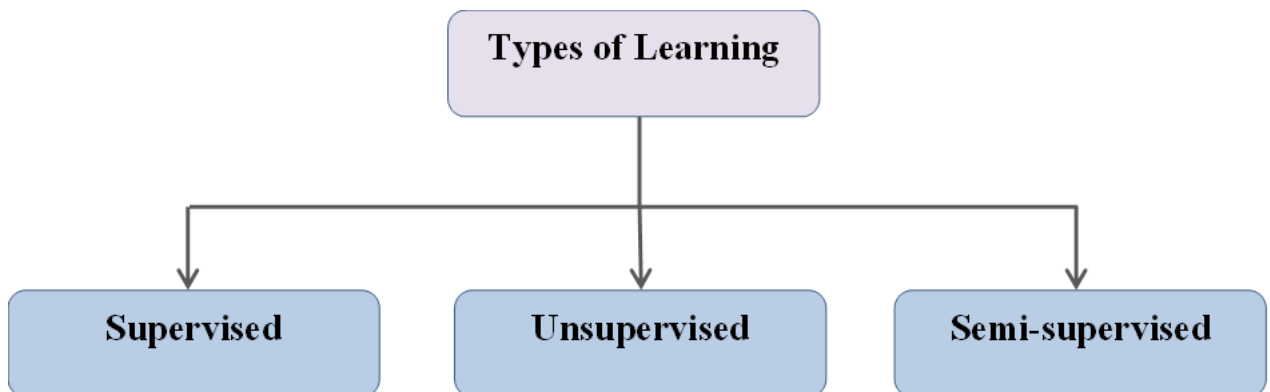
Feature selection is one of the methods for dimensionality reduction. Real-world datasets include a dense of information. They often have high-dimensional characteristics. However, not all characteristics are helpful for clustering methods. In addition high-dimensional features not only lengthen the computing time for machine learning, but also raise the chance of overfitting. Dimensionality reduction seeks to minimize the dimensions of data by collecting a collection of primary data or deleting duplicate and dependent characteristics. It converts the characteristics of a high-dimensional space to a low-dimensional space. It might be used to minimize complexity, prevent overfitting, and lessen the impact of outliers. Selecting valuable features from the original features or removing unnecessary or superfluous characteristics from the original dataset is the purpose of feature selection. In the categorization of medical images, reducing the number of characteristics decreases diagnostic costs and time [111]. There are several methods that can be utilized to apply feature selection (i.e., filter method, wrapper method, and embedded method). The filter method includes selecting a subset from the dataset that contains only relevant features by using filtering methods such as Pearson correlation. The wrapper method is more accurate than the previous one as it uses machine learning algorithms for evaluating features, but it needs more processing time. The embedded technique involves adding and removing features based on the performance of the model [106].

### 2.5.2.6 Classification

The critical step in the processing of images is classification. It is the last step of an image's feature detection process [112]. It is used to categorize features into many classes according to various properties and is crucial for medical imaging, particularly for the detection and identification of tumours, as well as for many other applications including robotics and voice recognition [113].

Machine learning algorithms can be classified mainly into three types, supervised learning (SL), unsupervised learning (USL), and semi supervised learning (SSL). The first type requires training through a labelled dataset that contains inputs and output as a target. In this type, there are three phases, the training phase, validation phase and the testing phase. In the training phase, the model in SL is first built through data that are labelled manually by human intervention, in the validation phase, the trained model is examined to determine the best model parameters, and then testing phase, new data that are not seen by the model will be used to test it [114]. Cross-validation is often used as part of the process of assessing different machine learning methods. It is an assessment of how well the model will be able to forecast data that will occur in the future. The procedure known as cross-validation begins by separating a dataset into two distinct groups. The first portion is used for the purpose of instructing an algorithm, and the second portion is utilized for assessing the efficiency of the method. One of a typical technique for cross validation is k-fold, where data is randomly divided into k folds (usually k is 10). Nine parts of k are used to train a model while one part is utilized for evaluating the model. The second type of machine learning algorithm, which is USL, there is no need to train a model. The data samples are classified based on common features of these samples; this type is suitable when there is no labelled data. There is another type of machine learning that is placed between the two aforementioned types, this type called Semi-Supervised Learning (SSL) that needs a few samples of labelled data that are used to label unlabelled samples. The SL algorithms can be used to solve problems of classification and regression. Classification is a learning process that deals with discrete data and classifies them into classes [115]. In the USL, clustering is used to solve the problem of identification of samples based on common characteristics. There are many algorithms that are utilized by researchers or developers to solve problems of classification and identification of samples automatically and rapidly. Support vector machine (SVM), K-

nearest neighbour (K-NN), naïve bayes (NB), random forest (RF) are the most famous algorithms that are used to identify breast cancer tumors [116]. Figure 2.11 shows the main types of machine learning approaches. In the below, we will briefly explain random forest and support vector machine.



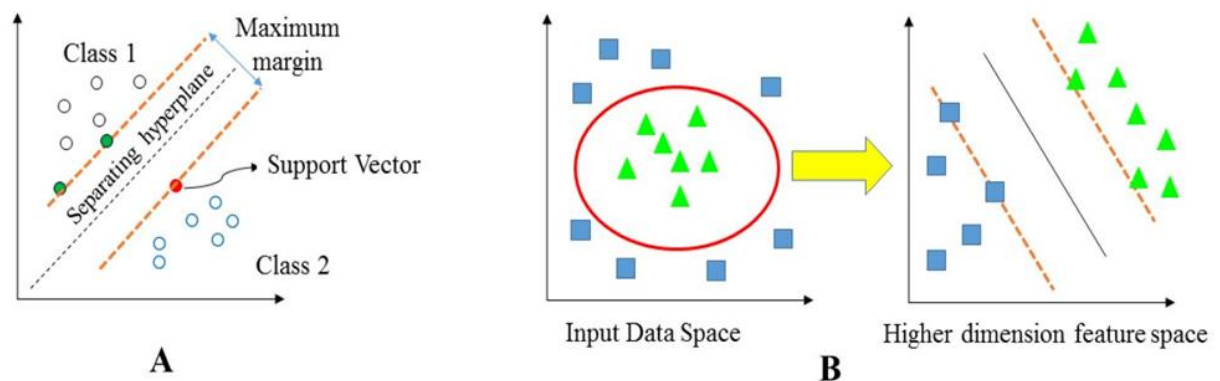
**Figure 2. 11** Machine learning approaches

- **Support Vector Machine (SVM)**

Support vector machine is a supervised machine learning approach used for data categorization. SVM is based on the concept of a hyperplane that separates a dataset into two groups in the best way possible. SVM has been used to solve a variety of issues, including pattern identification in bioinformatics and cancer detection. The separating hyperplane can be linear or non-linear. In the case of a linear function, SVM calculates the linear decision function in the gap that lies in the center of the two classes by categorizing all of the training data points and positioning the decision function as far as possible from the provided data points. In addition, SVM employs a non-linear mapping approach to translate the original training data into a higher dimensional form. To reduce classification mistakes, the classification must be done separately for each class [117].

SVM works by projecting data points from a given two-class training set onto a higher-dimensional space and finding the best hyperplane. The one that

isolates the data with the greatest margin is the ideal one. Support vectors are data points identified by SVM that are close to the ideal separation hyperplane. The margin of the SVM classifier is defined as the separation between the closest positive and negative data points and the separating hyperplane [118]. SVM schemes are shown in Figure 2.12:



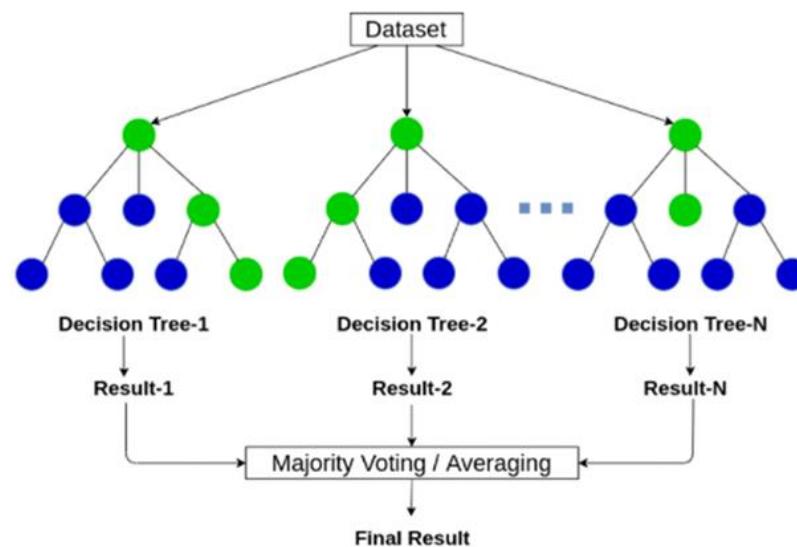
**Figure 2. 12** SVM schemes (A) Linear (B) Non-linear [119]

- **Random Forest (RF)**

Random forest is a well-known machine learning algorithm that uses the supervised learning method, it may be utilized for both classification and regression issues. It is based on ensemble learning, which is a method of integrating several classifiers to solve a complicated issue and increase the model's performance. RF is a technique that used widely in large datasets efficiently and quickly [120]. The technique relies on principle of ensemble learning that creates various classifiers and merges their results [121]. The performance of a single classifier is less than of multiple weak classifiers that both using the same dataset. RF is a type of ensemble approaches that classifies new instances through constructing many decision trees and majority voting. In this algorithm, an entire set of features is divided into many subsets and each one represents a decision tree that selected randomly [122].

RF prevents overfitting by calculating the average of all projections and cancelling out the biases. It can also manage missing values by replacing continuous variables with median values and determining their proximity-weighted average. Furthermore, it encourages the adoption of the most contributory characteristics for the classifier. RF is stable in high-dimensional environments and with a large amount of training data [123].

This technique gives more accuracy than a single decision tree, can handle datasets including a large number of predictor variables, and it may be used for variable selection. It is important to highlight the efficiency of random forest in a variety of fields, including the categorization and detection of the most relevant variables in ecology, breast cancer diagnosis and prognosis, and genomic data applications [124]. Figure 2.13 represents the RF mechanism:



**Figure 2. 13** The mechanism of RF [125]

## 2.6 Evaluation Metrics

Evaluation metrics are statistics that are used to assess the performance of each machine learning algorithm, some algorithms can provide a good result on a certain dataset or a class while performing poorly on the other group of data or other classes, and thus there is not a particular way to select the optimal

algorithm. The primary objective of estimating the classification model is to provide an accurate assessment of the quality of the model's performance [126]. The confusion matrix, accuracy, error rate, receiver-operating curve (ROC), area under the ROC curve (AUC), sensitivity, specificity, precision, and the F1 score are some of the methods used to evaluate classifiers.

### 2.6.1 Confusion Matrix

The confusion matrix is a specialized table that depicts the classifier's performance. In the machine learning approach, a confusion matrix is often referred to as the error matrix. Depending on the data type, an image area is either positive or negative. Moreover, a choice based on a detected result may be valid (true) or invalid (false). As a result, the decision will thus fall into one of four categories, which are defined as below [127]:

- True Positive (TP): is a positive instance correctly diagnosed as positive.
- True Negative (TN): is a negative instance correctly diagnose as a negative.
- False Positive (FP): is a negative instance incorrectly detected as a positive.
- False Negative (FN): is a positive instance incorrectly detected as a negative.

The relation between positive class and negative class predictions can be depicted as a 2×2 confusion matrix in Table 2.1, which tabulates whether the obtained prediction falls into one of four categories. The right choice is represented by the diagonal of the confusion matrix [128].

**Table 2. 1** Confusion matrix

		<i>Predicted Classes</i>	
		<b>Positive</b>	<b>Negative</b>
<i>Actual Classes</i>	<b>Positive</b>	TP	FN
	<b>Negative</b>	FP	TN

**Accuracy:** is a metric that determines the probability that how many samples (positive and negative) are correctly identified. The best accuracy is 1, whereas the worst is 0. With the  $2 \times 2$  confusion matrix, the formula of prediction accuracy is shown in Eq. 2.9.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.9)$$

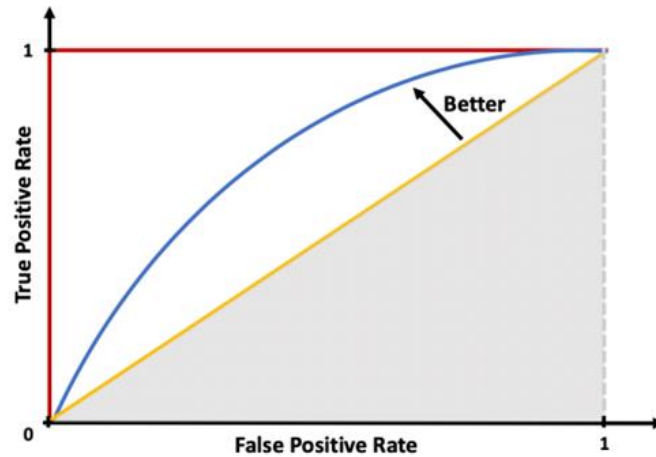
### 2.6.2 Receiver Operating Characteristic (ROC)

The ROC curve is a visual representation of a plot that demonstrates how well a binary classification model performs. It is produced by graphing the true positive rate (TPR) versus the false positive rate (FPR) at a variety of discriminating thresholds. The TPR is also known as the sensitivity or recall, and the FPR may be computed as the product of the specificity and the sensitivity. If the discriminating threshold of the model is altered, the confusion matrix may take on a new appearance, and a new point may be shown on the ROC curve.

A random binary classification model's ROC curve would look as a diagonal line extending from (0, 0) to the point where the model stopped making predictions (1, 1). Any curve that is located above the diagonal line is indicative of a good classification model that is superior to random, but any curve that is located below the diagonal line indicates that the model is inferior to random. A perfect binary classification model would provide a straight line from (0, 1) to (1, 1) as the output. This would indicate that the model has a sensitivity of 100% and a specificity of 100%.

The area under the ROC curve is referred to as the AUC and its value is always between 0 and 1. When examining binary classification models using the ROC curve approach, one might reach the conclusion that a greater AUC indicates a superior model. The AUC for a random classifier is 0.5, whereas an excellent model has an AUC that is near to 1. In situations when the AUC is less

than 0.5, the model has a tendency to reverse the classes [129]. Figure 2.14 depicts a schematic representation of a ROC curve and AUC.



**Figure 2. 14** Schema of ROC curve and AUC: red line: a perfect classifier, blue curve: a great classifier, yellow line: a random classifier, and shaded area: AUC for the random classifier [129]

**Sensitivity (Recall):** measures the rate of true positives that are correctly identified as positive. Its formula is in Eq. 2.10 [130].

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (2.10)$$

**Specificity:** measures the rate of true negative cases that are correctly identified as negative and its formula is as in Eq. 2.11. The range of these parameters is from 0 to 1, when the values close to value 1, they are being more desirable [131].

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (2.11)$$

## 2.7 Synthetic Minority Oversampling Technique (SMOTE)

In many different domains, like credit card fraud, medical diagnosis, and others, consists of imbalance data. The unbalanced data is caused by an unequal balance in the amount of positive and negative cases, or binary class labels, in a dataset. It is one of the most difficult aspects of classification using machine



learning since the classification is biased toward the majority class and performs poorly according to the receiver operator characteristic curve (ROC). Imbalanced data can provide high accuracy of testing results since the testing sample will simply be categorized to the majority class, which inhabited a high proportion of the entire population. It is a significant problem because the primary focus in dataset is on the minority class. As a result, any errors in data classification or inability to detect the true negative in even one of the data may provide a difficulty for the case study [132].

An unbalanced dataset has considerably lower number of training cases for one class than for another. In accordance with this, the former is referred to as the minority class and the latter as the majority class. Most common learning or classification algorithms have a tendency to classify the majority class with a high accuracy rate and the minority class with a low accuracy rate for an unbalanced dataset. Due to these disparities in accuracy rates, the classifier performs poorly when identifying minority class samples. As a result, one of the major challenges in classification research is classifying an unbalanced dataset. Minority class samples are more crucial to identify than majority ones in many applications.

Re-sampling, also known as dataset reconstruction, is the practice of modifying the distribution of training set samples via data processing in order to enhance classification performance by lowering the dataset's imbalance. This strategy incorporates oversampling, under sampling, and other mixed sampling strategies [133]. In under sampling, the number of occurrences for a majority class is decreased such that it is equal to or comparable to that of the minority class. In oversampling, the number of instances in the minority class is raised such that it is equivalent to or comparable to the majority class [134]. Given that minority class samples of the original data are replicated, over-sampling may lead to border or noisy data, increase processing time, and result in inefficient

over-fitting [133]. In this context, overfitting refers to the machine learning algorithm's inability to generalize to a new dataset due to its best adaptation to the training set (e.g., test set). An easy technique to detect overfitting is to compare the results on the training and testing sets. If the accuracy on the training set is much greater than the accuracy on the testing set, then the algorithm is overfit [134]. Chawla introduced SMOTE, which stands for synthetic minority oversampling technique and is a robust over-sampling method that is most commonly used to balance skewed data in machine learning [135]. The SMOTE approach randomly generates new instances of a minority class from the minority classes nearest neighbors. In addition, these instances are utilized to examine the different characteristics of the original dataset and are regarded as the original instances of the minority class [136]. The hybrid-sampling methods combine more than one sampling methodology to overcome the drawbacks of every sampling technique [137].

## 2.8 Mammographic Databases

A variety of well-known databases for mammographic image analysis have been developed. Some datasets have been made publicly available, while others have been kept private by a group of academics. The MIAS database and the DDSM database are the most commonly used datasets. The following is a more detailed database description.

### 2.8.1 Mammographic Image Analysis Society (MIAS)

Mammographic image analysis society (MIAS) has created a database that is made up of digitized versions of mammograms. Prior to the process of digitalization, the United Kingdom National Breast Screening Programme provided us with the original X-ray film that we gathered. The database is accessible at <http://www.wiau.man.ac.uk/services/MIAS/MIASweb.html>. The Mini-MIAS is a compact version of the MIAS dataset, in which the original images have been cropped. The dimensions of each image in the collection have

been adjusted to  $1024 \times 1024$  pixels (8-bit resolution). There are a total of 322 mammography images in the database, including 208 normal, 63 benign, and 51 malignant mammograms. The radiologist's ground truth is also included in the database, but it may be accessed as a distinct text file. This file includes information on abnormality, severity (benign or malignant), and tissue density (fatty, fibroglandular, heterogeneously dense or extremely dense). A great number of academics in the past have made use of the database for the purpose of creating segmentation or classification algorithms [53].

### **2.8.2 Digital Database for Screening Mammography (DDSM)**

The digital database for screening mammography (DDSM) is a resource available at <http://marathon.csee.usf.edu/Mammography/Database.html> for the mammographic image analysis research community. A group of specialists from the University of South Florida produced the DDSM database. The DDSM database has 2,620 breast cases separated into 43 volumes. In each case, each breast is imaged four times using two distinct slide perspectives MLO and CC. A typical DDSM database is  $3000 \times 4800$  pixels in size. This comprises the kind of cancer (benign or malignant) and the wound location. The curated breast imaging subset of DDSM (CBIS-DDSM) database includes pixel-by-pixel annotations for ROI, which consists of tumours, calcifications, and disease pathology (benign or malignant). In addition, each ROI is identified as calcification or mass. The majority of mammograms have just one ROI [138].

# *Chapter Three*

## *Methodology*

### 3.1 Introduction

As described in chapter two, segmentation and classification of mammography images are vital steps in the detection and diagnosis of breast cancer. In order to recognize the abnormalities of breast cancer, numerous researchers have suggested various algorithms. In fact, computers are capable of separating different parts of an image and extracting suspicious areas ROI from medical images using a wide range of methods.

This chapter can be divided into two approaches: The first approach discusses the methodology of the new detection technique for micro-calcifications in mammogram images using the proposed multi-points (seeds) region growing method, and the second approach argues the methodology of improving breast cancer classification using the SMOTE technique and pectoral muscle removal in the mammogram images of the Mini-MIAS dataset.

### 3.2 Image Processing Based CAD Systems

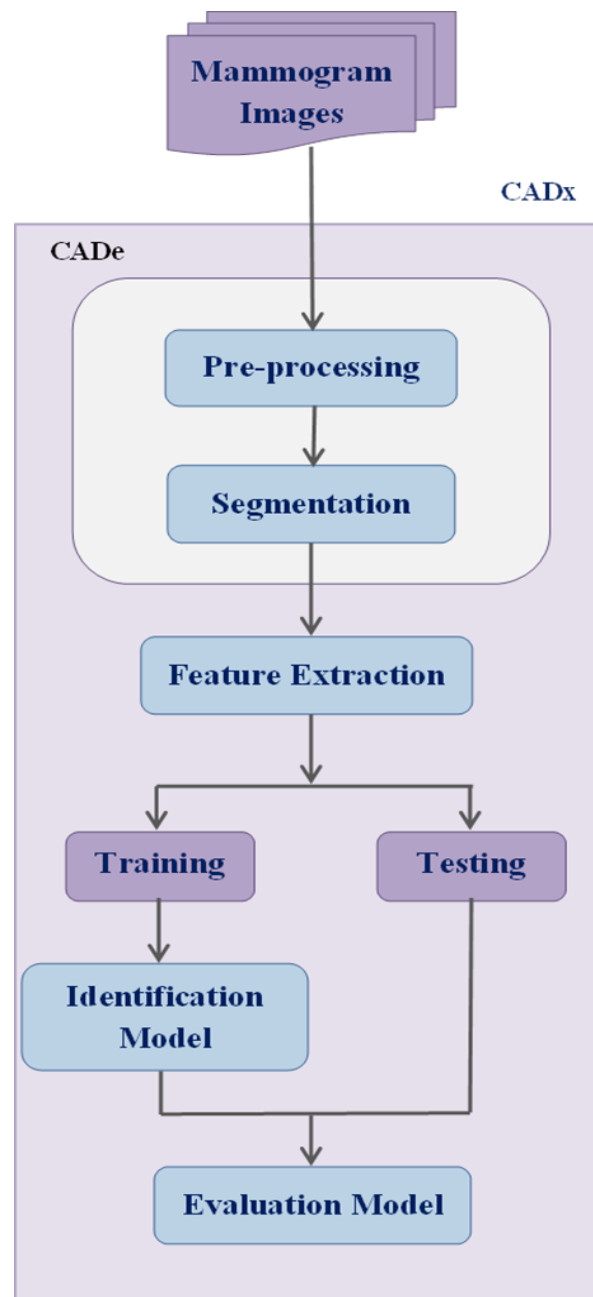
Computer-Aided Detection/Diagnosis (CAD) is defined as the ability to utilize a computer system in medical diagnosis. Most CAD systems are intended to help detect breast cancer using mammograms. However, the images must be in the appropriate digital format first, in order to be useful as an input to a CAD system. Thus, the first role of image processing is often simply the digitization of an existing mammogram.

However, this is frequently the first step, as subsequent image processing is used to improve the image's quality before identifying, separating, or otherwise marking the image elements or features of interest.

The process of separating the abnormal from the normal part is known as segmentation. Each recognized region reflects the information to which it belongs as well as structuring features that distinguish the abnormality. The

primary goal of segmentation in this CAD model is to separate mass or micro calcification from breast tissue.

Radiologists depict masses or micro-calcifications by their texture properties. In CAD, classification is used to predict the type of mass as benign or malignant based on the extracted set of features. Figure 3.1 shows the system architecture for breast cancer detection.

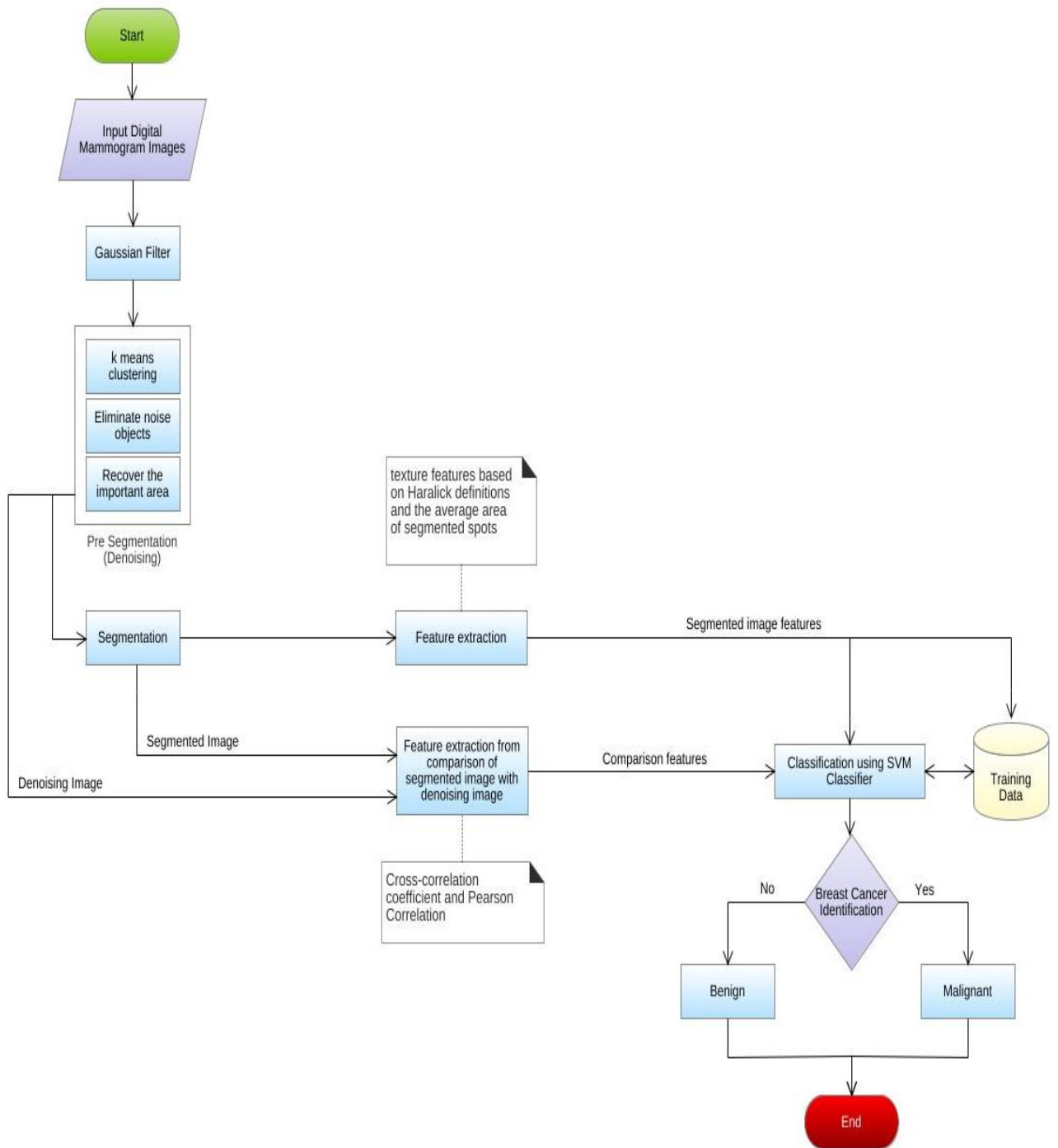


**Figure 3. 1** System architecture for breast cancer detection

### 3.3 Proposed Method 1: Segmentation using K-means Clustering and Optimized Region-growing Technique

This method introduces a useful pixel-based technique for region-growing. Due to the nature of the calcifications, which can be sporadic, multi-points have been utilized to detect calcifications in more than one region of the breast, which cannot be determined by using the standard region growing algorithm. On the other hand, using an initial seed point increases the computing cost and execution time.

The proposed methodology involves multiple steps, including image processing techniques. The first step is the image acquisition from the dataset collected in the curated breast imaging subset of the digital database for screening mammography (CBIS-DDSM), where regular and irregular mammograms are collected. The optical mammograms are then preprocessed by Gaussian filters and pre-segmented using k-means clustering for noise reduction. The images are further processed using the proposed multi-points (seeds) region growing method to extract the ROI, which targets breast micro-calcifications (MCs). For feature extraction, the ROIs are then processed where a collection of texture features are extracted using Haralick texture characteristics. Then the extracted textures are further fed into the SVM classifier. In Figure 3.2, the detailed processes of the proposed model are demonstrated.



**Figure 3. 2** Flowchart of the proposed method



### 3.3.1 Data Acquisition

The curated breast imaging subset of DDSM (CBIS-DDSM) dataset is available at [139], which includes the digital images fed to the proposed method. The DDSM database has 2,620 breast cases separated into 43 volumes. In each case, each breast is imaged four times using two distinct slide perspectives MLO and CC. A typical DDSM database is  $3000 \times 4800$  pixels in size. This comprises the kind of cancer (benign or malignant) and the wound location. The images were decompressed and saved in DICOM format.

In this work, 440 digital mammograms of breast micro-calcification containing benign and malignant severity were used in craniocaudal (CC) views for training and testing purposes. This collection comprises 329 benign cases and 111 malignancy cases in both left and right breasts.

### 3.3.2 Pre-processing

The preprocessing consists of two steps: the first step is resizing, and the second step is filtering the mammogram image.

- **First step (Resizing):** The mammography images are high resolution images that include small features of interest and need to be scaled down to enable better transport and processing. In order to effectively reduce the size of the mammography without affecting its quality or regions of interest, the bilinear interpolation approach is applied. Bilinear interpolation is determined by taking a weighted average of the four neighboring pixels to calculate its final interpolated value. The final image is significantly smoother than the original. When all known pixel distances are equal, then the interpolated value is simply their sum divided by four. This technique interpolates in both horizontal and vertical directions. The interpolation kernel for bilinear interpolation is given by [140]:

$$u(x) = \begin{cases} 0 & |x| > 1 \\ 1 - |x| & |x| < 1 \end{cases} \quad (3.1)$$

Where  $x$  = distance between interpolated point and grid point.

- **Second Step (Filtering):** The noise is an unwanted phenomenon that is inherent to many image acquisition and transmission sources. Due to various interferences, noise has a negative impact on the quality of the image. In order to suppress unwanted noise, smoothing filters are often applied to the image to enhance its quality while attempting to preserve as many important details as possible in the image.

In this step, the pre-processing method utilizes a Gaussian filter to remove noise and soften the images. Figure 3.3.a is the original image, and Figure 3.3.b shows the effect of applying a Gaussian filter to the image.

A Gaussian filter is a low pass filter that is used to reduce noise and blur image regions. To get the desired effect, the filter is built as an odd sized symmetric kernel (DIP version of a matrix) that is passed through each pixel of the image. The filter uses a kernel to perform a convolution filter, which is calculated using the kernel 2D method and then transformed to an integer sharpening kernel. First, the integer kernel from kernel 2D is calculated by dividing all elements by the element with the smallest value. A Gaussian filter can be thought of as an approximation to the Gaussian function (mathematics). When applying the Gaussian filter to an image, we first define the size of the kernel ( $11 \times 11$ ) that will be used to blur the image. Because the sizes are generally odd, the overall results can be computed on the middle pixel. Also, because the kernels are symmetric, they have the same number of rows and columns. The Gaussian function is used to compute the values within the kernel, which is as follows:

$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (3.2)$$

Where,

$x$ :  $x$  coordinate value

$y$ :  $y$  coordinate value

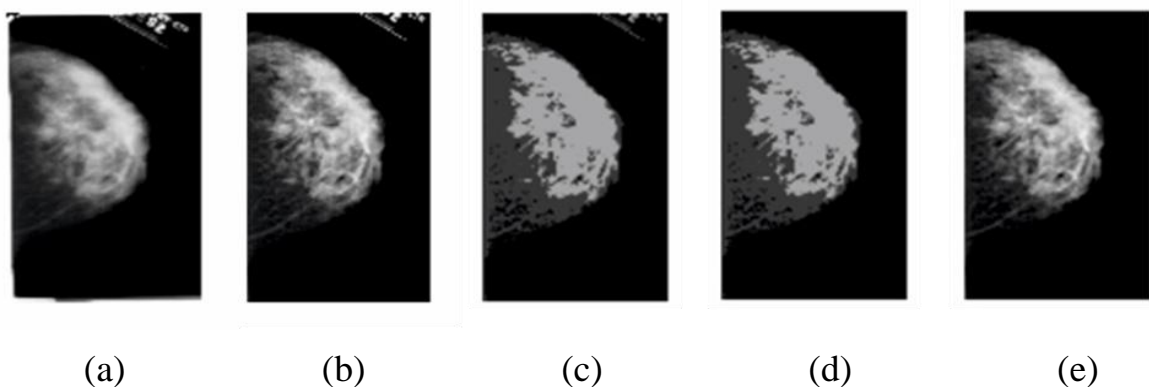
$\pi$ : Mathematical Constant  $PI$  (value = 3.14)

$\sigma$ : Standard Deviation (*sigma*)

The role of *sigma* in the Gaussian filter is to control the variation around its mean value. Therefore, as the *sigma* becomes larger, the more variance is allowed around the mean and as the *sigma* becomes smaller, the less variance is allowed around the mean. We adopted the sigma value of 4.

### 3.3.3 Pre-segmentation (Denoising)

Usually, medical images contain some symbols, words, or letters that show the type or some of the medical-physical characteristics of the image. This is generally considered image noise and may affect classification accuracy. To overcome these problems denoising is used. This process goes through three stages: 1- K-means clustering, 2- Eliminate noise objects and 3- Recover the important area. Figure 3.3 (c, d, e) illustrates the process of applying these steps.



**Figure 3. 3** The preprocessing and pre-segmentation steps: (a) Original image (b) Gaussian filter (c) Applying k-means (d) Erosion filter and (e) Breast area retrieval

### 3.3.3.1 K-means Clustering

K-means is a simple learning technique for solving the well-known clustering problem. K-means split a set of data into  $k$  number groups. The k-means algorithm is divided into two stages. In the first stage, it computes the  $k$  centroid, and in the second phase, each data point is assigned to the cluster that has the closest centroid to it. Once the grouping is complete, it recalculates the new centroid of each cluster and, based on that centroid, calculates a new Euclidean distance between each center and each data point and assigns the cluster points with the lowest Euclidean distance. Each partition's cluster is defined by its member objects and centroid. The centroid of each cluster is the point at which the total of distances between all the objects in that cluster is minimized. Finally, the goal of this algorithm is to minimize the sum of squared distances between all points and the cluster center. The objective function is as follows [101]:

$$d = ||p(x, y) - c_j||^2 \quad (3.3)$$

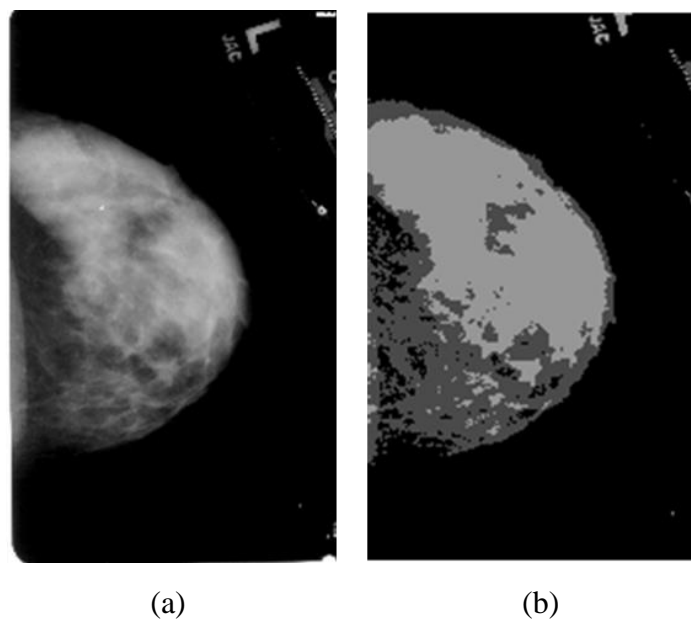
$$(c_j) = \frac{1}{k} \sum_{x=1}^{c_j} \sum_{y=1}^{c_j} p(x, y) \quad (3.4)$$

Let us consider an image with resolution of  $x \times y$  and the image has to be cluster into  $k$  number of cluster. Let  $p(x, y)$  be an input pixels to be cluster and  $c_j$  be the cluster centers. The algorithm is composed of the following steps:

- Step 1** Place 3 points into the space represented by the objects that are being clustered. These points represent initial group centroids  $k = 3$ .
- Step 2** For each pixel of an image, calculate the Euclidean distance  $d$  between the center and each pixel using the relation given in Eq. 3.3.
- Step 3** Assign each object to the group that has the closest centroid.
- Step 4** When all objects have been assigned, recalculate the positions of the K centroids using Eq. 3.4.

**Step 5** Repeat steps 2 to 4 until the centroids no longer move. This produces a separation of the objects into groups, from which the metric to be minimized can be calculated.

At this stage, the image is divided into a group of converged areas in color intensity, and as a result, the background of the image will be isolated from its components more clearly. Also, the noise in the images will be isolated from the rest of the image components, which facilitates the process of cutting them later. Figure 3.4 illustrates the process of applying k-means clustering.

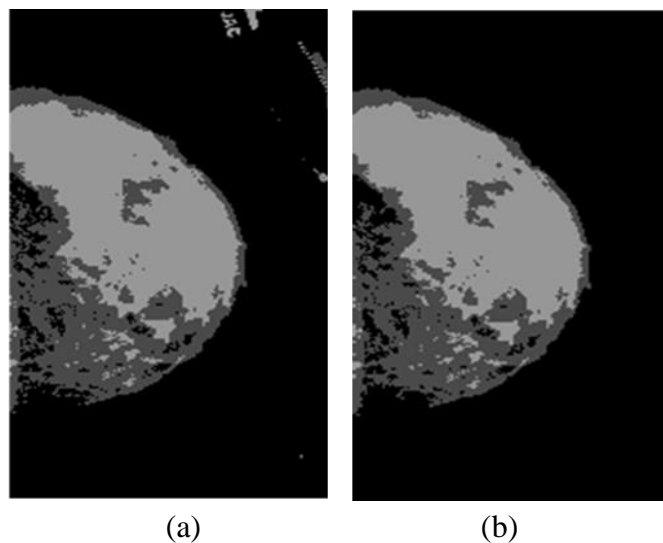


**Figure 3. 4** The k-means clustering process: (a) Original image (b) Applying k-means

### 3.3.3.2 Eliminate Noise Objects

Morphological image processing is a collection of non-linear operations related to the shape or morphology of features in an image. A morphological operation is used to rearrange the order of the pixel values, which operates on structuring elements and input images. Structuring elements are attributes that probe features of interest. Erosion is an essential operation used here, and during erosion, the rock bottom value is chosen by comparing all the pixel values in the region of the input image with kernel (3×3).

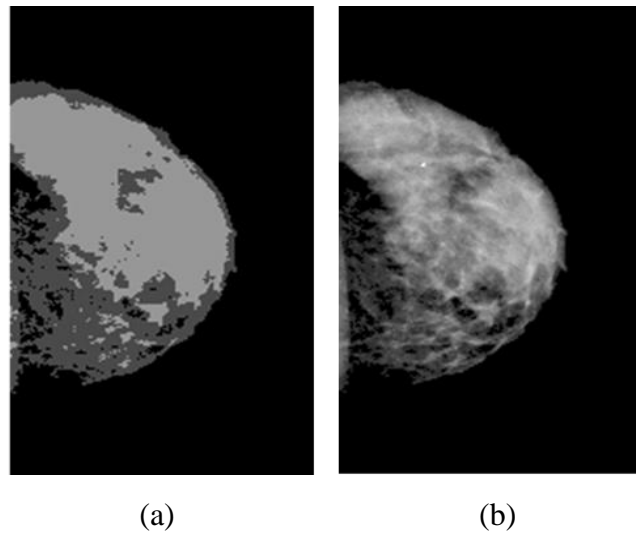
After dividing the image in the previous step into specific areas of converging intensity in color and isolating them from the background, the extra foreign objects that cause noise in the image are cut off by erosion. Thus, we have an image that contains only the breast area without any noise as shown in Figure 3.5, this area will be used to restore the breast area from the original image, which we call the breast area mask.



**Figure 3. 5** The eliminate noise object process: (a) Applying k-means (b) Erosion filter (breast area mask)

### 3.3.3.3 Recover the Important Area

It is the last step in the pre-segmentation stage. The mask that was produced in the previous step is relied on and applied to the original image to restore the equivalent area of the mask, then neglect the components of the image and consider it as background for the area resulting from the retrieval process. Figure 3.6 demonstrates the area of the retrieved breast.

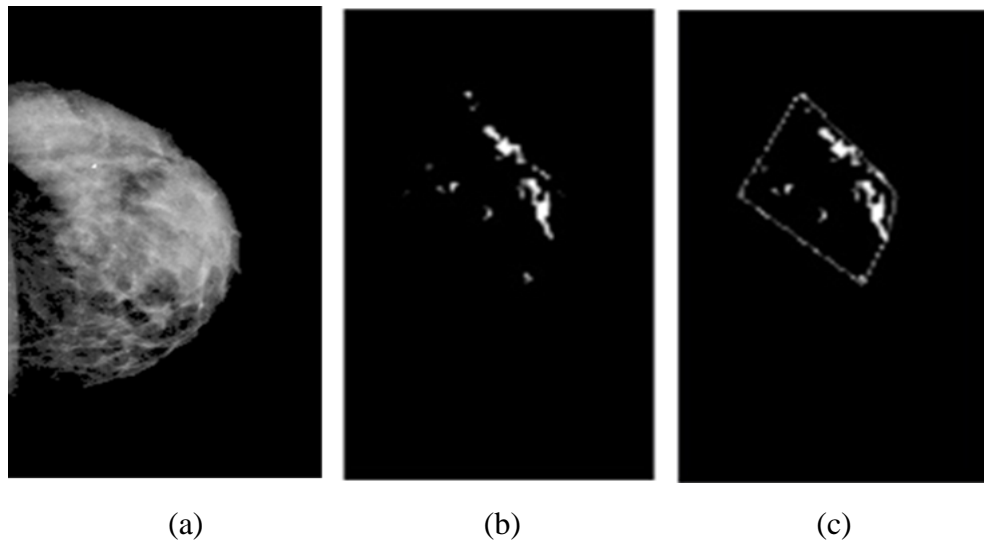


**Figure 3. 6** Recover the important area process: (a) Erosion filter (breast area mask)  
(b) Breast area retrieval

### 3.3.4 Segmentation

The segmentation method separates benign and malignant areas of digital mammograms into non-overlapping segments from the background portions. The region-based strategies find a seed point and growing regions until a criterion of homogeneity is reached.

This method presents an effective variant of the region-growing pixel-based technique that produces optimal seeds and thresholds. Due to the nature of the calcifications, which may be sporadic, multi-points (seeds) have been used to determine the calcifications if they are in more than one region of the breast, which cannot be determined when using the traditional region-growing algorithm. Figure 3.7 depicts the segmented image of micro-calcifications.



**Figure 3. 7** The segmentation process: (a) Breast area retrieval (b) Segmented area and (c) ROI

In the following the main steps for generating optimal seeds using multi-points (seeds) region-growing segmentation is presented:

- Let mammogram values be a list of mammogram pixel values (it will contain all the pixel values in the image without duplicating).
- Sort mammogram pixel values in descending order.
- Determine the segmentation threshold, which will determine the approved values from the list and which will be adopted in the segmentation process. This is done by dividing the sorted mammogram values into 10 sections and adopting the first section of it (higher values), which will represent the list of segmentation values or seed points.
- For each pixel in the mammogram image, if the pixel value belongs to the segmentation values, then the pixel will be adopted, otherwise, the pixel will be discarded.
- For each adopted pixels generated from the previous step:
  - Check the neighboring pixels and add them to the region if they are equal or upper to the seed intensity value.



- Repeat the previous step for each of the newly added pixels; stop if no more pixels can be added.

The algorithm for the multi-points region-growing method is presented below:

### Algorithm 3.1 Optimized Region Growing

---

**Input:** Digital Mammogram Image (M)

**Output:** Segmented Image (S)

#### Local Variables

W = Digital Mammogram Image width

H = Digital Mammogram Image height

MammValues<>: list of Mammogram pixel values

SegmentationValues<>: list of segmentation threshold values

#### Begin

##### Step 1

Aggregation (M)

**For** i ← 1 to W **do**

**For** j ← 1 to H **do**

**If** (! MammValues.Contains(M<sub>(i,j)</sub>))

      MammValues.Add(M<sub>(i,j)</sub>)

**End**

**End**

**End**

##### Step 2

Sort (MammValues)

##### Step 3

Determine thresholds (MammValues)

**For** i ← 1 to (MammValues.Count / 10) **do**

  SegmentationValues.Add(MammValues[i])

**End**

**Step 4**

Segmentation (M, SegmentationValues)

**For**  $t \leftarrow 1$  to SegmentationValues.Count **do**

**Apply region growing**

**End**

**Return S**

**End**

---

### 3.3.5 Features Extraction

Features of the image show the current attributes and characteristics. The extracted features utilized for classification should be identifiable, effective, and autonomous.

In the first step of features extraction, statistical textural analysis-features including cross-correlation coefficient and Pearson correlation information from the comparison of the denoising image with the segmented image intensities extracted.

- **Cross-Correlation Coefficient:** It is a measure of similarity of two series as a function of the displacement from one another. The cross-correlation coefficients are more robust to changes in illumination than the mean square error (MSE).

$$\text{Cross-Correlation Coefficient} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \quad (3.5)$$

Where,

$n$ : is sample size

$x_i, y_i$ : are the individual sample points indexed with  $i$

$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ : (the sample mean); and analogously for  $\bar{y}$

- **Pearson Correlation Coefficient:** Pearson correlation in statistics is defined as a measure of the strength of a relationship between two variables and their correlation with one another. Pearson's correlation coefficient computes the effect of a change in one variable when the other variable changes.

$$\text{Pearson Correlation Coefficient} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (3.6)$$

Where,

$n$ : is sample size

$x_i, y_i$ : are the individual sample points indexed with  $i$

$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ : (the sample mean); and analogously for  $\bar{y}$

Then, the average area of segmented spots was obtained. In this extraction process, the average area is calculated which represents the average of the infected areas. Suppose that  $B = \{B_1, B_2, \dots, B_N\}$  is the set of segmented blobs, where  $N$  is the number of the blob segmented from the mammograms.

$$\text{Average Area} = \frac{\sum_{i=1}^N \text{Area}(B_i)}{N} \quad (3.7)$$

In the second step of feature extraction, the proposed method utilizes a collection of texture features based on Haralick's texture analysis concepts, where 26 texture features are extracted. The features are: Angular second moment, contrast, correlation, variance, inverse difference moment, sum average, sum variance, sum entropy, entropy, difference variance, difference entropy, first information measure, second information measure, and invariance were achieved for each of these 13 features by averaging them over the four directional co-occurrence matrices [107].

Thus, for training the classifier, a collection of 29 features was extracted (two features from the first step, 26 features from the second step, and the average area). The statistics formulas for the haralick features are listed below:

**1. Angular Second Moment [asm]:** is a measure for image homogeneity.

$$f_1 = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \left( \frac{P(i,j)}{R} \right)^2 = \sum_i \sum_j p(i,j)^2 \quad (3.8)$$

Where,

$N_g$  is the number of gray levels,  $p$  is the normalized symmetric GLCM of dimension  $N_g \times N_g$ , and  $p(i,j)$  is the  $(i,j)$ th element of the normalized GLCM.

**2. Contrast [con]:** is a measurement of the local variances present in an image.

$$f_2 = \sum_{k=0}^{N_g-1} k^2 \left\{ \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \delta|i-j|, k P(i,j) \right\} = \sum_{k=0}^{N_g-1} k^2 p_{x-y}(k) \quad (3.9)$$

**3. Correlation [cor]:** The linear dependence of the gray levels of neighboring pixels is measured by correlation.

$$f_3 = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (ij)p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y} \quad (3.10)$$

Where,

$\mu_x, \mu_y, \sigma_x, \sigma_y$  are the means and standard deviations of  $px$  and  $py$ .

**4. Sum of Squares variance [var]:** The dispersion of the gray level distribution is measured by variance.

$$f_4 = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i - \mu)^2 p(i,j) \quad (3.11)$$

**5. Inverse Difference Moment [idm]:** measures the local homogeneity of an image.

$$f_5 = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \frac{1}{1+(i-j)^2} p(i,j) \quad (3.12)$$

Where,

The term  $(i-j)^2$ : weighting factor ( a squared term).

**6. Sum average [sav]:** is the image's average value for intensity.

$$f_6 = \sum_{i=2}^{2N_g} i p_{x+y}(i) \quad (3.13)$$

**7. Sum variance [sva]:** is a measure of the intensity variation around the mean.

$$f_7 = \sum_{i=2}^{2N_g} (i - f_8)^2 p_{x+y}(i) \quad (3.14)$$

**8. Sum entropy [sen]:**

$$f_8 = \sum_{i=2}^{2N_g} p_{x+y}(i) \log(p_{x+y}(i)) \quad (3.15)$$

**9. Entropy [ent]:** measure the randomness of intensity value.

$$f_9 = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) \log(p(j,j)) \quad (3.16)$$

**10. Difference variance [dva]:** calculate the difference in variances between the two response variables.

$$f_{10} = \text{variance of } p_{x-y} \quad (3.17)$$

**11. Difference entropy [den]:**

$$f_{11} = \sum_{i=0}^{N_g-1} p_{x-y}(i) \log(p_{x-y}(i)) \quad (3.18)$$

**12. First information measure of correlation:**

$$f_{12} = \frac{f_9 - HXY1}{\max(HX, HY)} \quad (3.19)$$

**13. Second information measure of correlation:**

$$f_{13} = [1 - \exp(-2(HXY2 - f_9))]^{1/2} \quad (3.20)$$

In this section, a sample is shown of the obtained results. It displays a comparison of some extracted features of the segmented images. Figure 3.8 (a, b, c, and d) shows the distribution of the angular second moment, variance, first information measure, and cross-correlation coefficient values for a group of 30 segmented samples for every two classes (malignant and benign), and illustrated the extracted features that can be used to distinguish and observed the values of the malignant tissue differed from the values of the benign tissue.

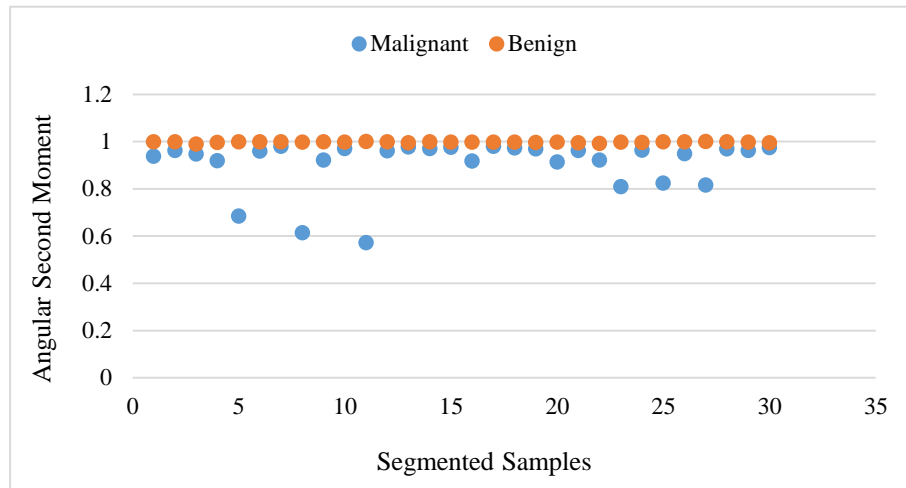


Figure 3.8 (a) Angular second moment values of malignant and benign samples

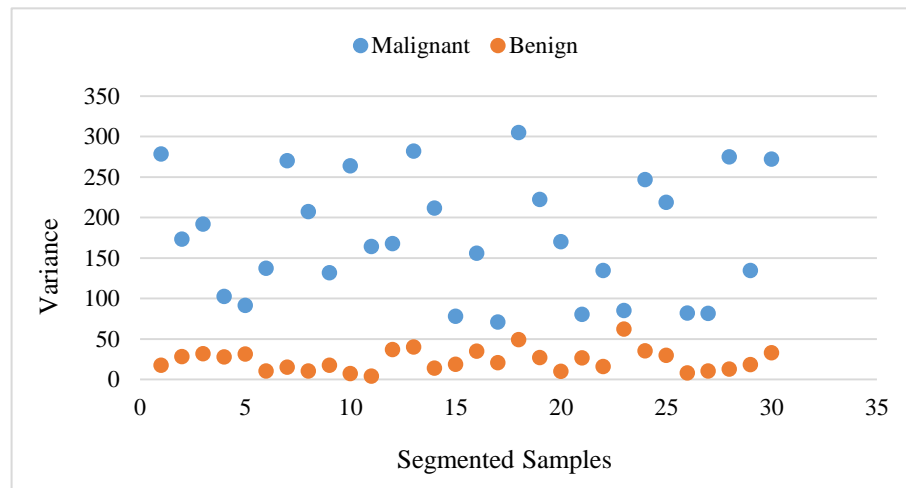


Figure 3.8 (b) Variance feature values of malignant and benign samples

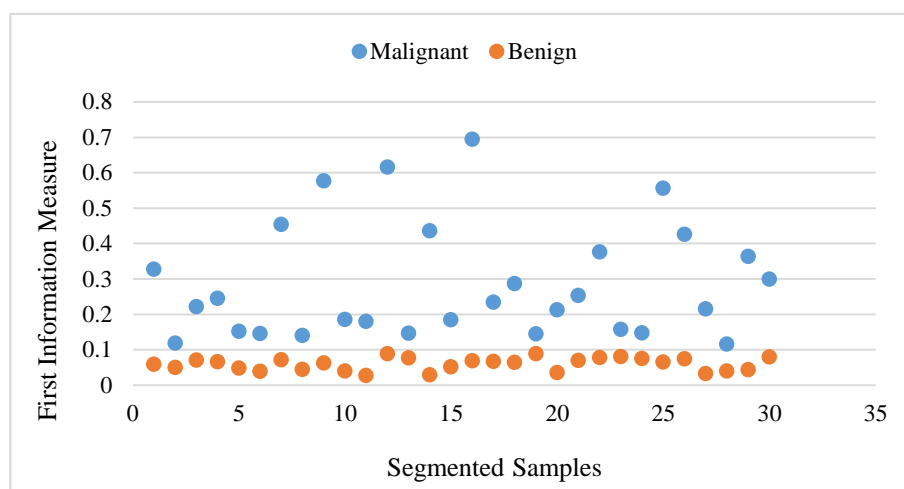
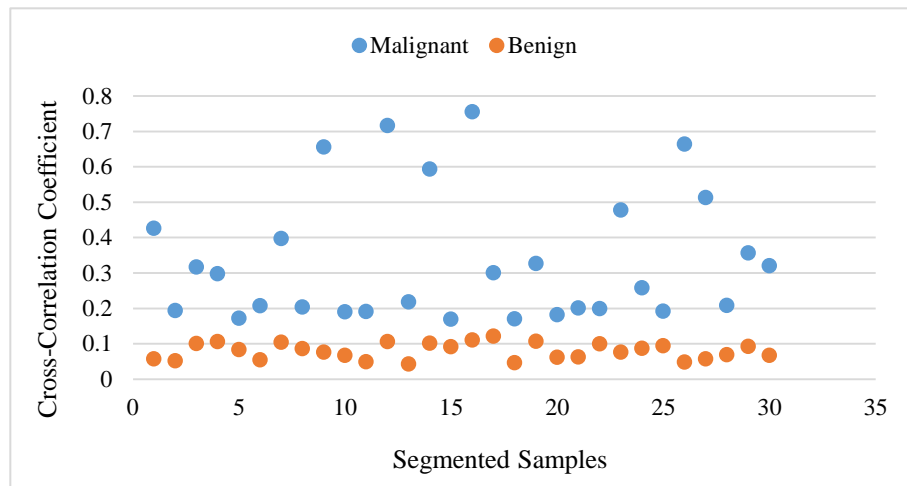


Figure 3.8 (c) First information measure feature values of malignant and benign samples



**Figure 3.8 (d)** Cross-Correlation Coefficient feature values of malignant and benign samples

### 3.3.6 Support Vector Machine Classifier

SVM is an effective statistical learning method for classification, which seeks to minimize the upper bound of the generalization error based on structural risk minimization. The working principle of the support vector machine is based on marginal calculations [141].

In this method, the Gaussian kernel function is used for transformation. A kernel function is a method for taking input data and transforming it into the needed form of processing data. In general, the kernel function transforms the training set of data so that a non-linear decision surface can transform to a linear equation in a higher number of dimension spaces. Basically, it returns the interior result between two points in a standard feature dimension. In the following the support vector machine classifier pseudocode is presented:

#### **Algorithm 3.2** Support Vector Machine Classifier

---

**Input:** Dataset **D**

**Output:** Confusion Matrix, Validation

**Step 1** (Divide the dataset into training and testing)

**Training**  $\leftarrow$  Split (**D**, size = 0.8)

**Testing**  $\leftarrow$  Split (**D**, size = 0.2)

**Step 2** (Algorithm training)

SupportVectorMachine<Gaussian>.Learn (**Training, Training\_Output**)

**Step 3** (Breast cancer prediction for the testing dataset)

**Prediction** = SupportVectorMachine<Gaussian>.Decide(**Testing**)

**Step 4** (Extract classification results)

Calculate **Scores(Prediction, Testing\_output)**

Compute **Confusion Matrix(Prediction, Testing\_output)**

**Step 5** (Evaluation of results)

**Validate Model**

---

### 3.4 Proposed Method 2: Pectoral Muscle Removal and Solving Data Imbalance Problem

This section proposes another approach in mammogram CAD systems for addressing two problems: pectoral muscle removal and data imbalance problems.

The segmentation of the pectoral muscle is useful in mammography image processing because this muscle typically appears as a dense region of incidences in mediolateral oblique (MLO) views of the mammograms. This may affect the performance of methods for the automatic detection of lesions.

Various image-processing techniques and segmentation algorithms have been used by researchers to analyze samples and enhance visual accuracy to discover and interpret regions of interest. In this model, the focus will be on removing the pectoral muscle to improve the segmentation process and solve the problem of data imbalance as in the Mini-MIAS dataset to avoid its impact on the accuracy of classification results.



The proposed method demonstrated in Figure 3.9, involves acquiring images, noise reduction, breast region differentiation, filtering (invert, subtract), pectoral muscle removal, segmentation, feature extraction, balancing the data in the feature space as well as performing the classification. In regards to the segmentation process, the proposed system performs many operations to improve the segmentation process, the most important of which is solving the pectoral muscle problem and ensuring that it does not appear which affects the efficiency of extracting characteristics and thus improves the performance of the classification process. Then, the proposed system solves the imbalance problem in the Mini-MIAS dataset utilizing SMOTE technique to raise the classification efficiency.

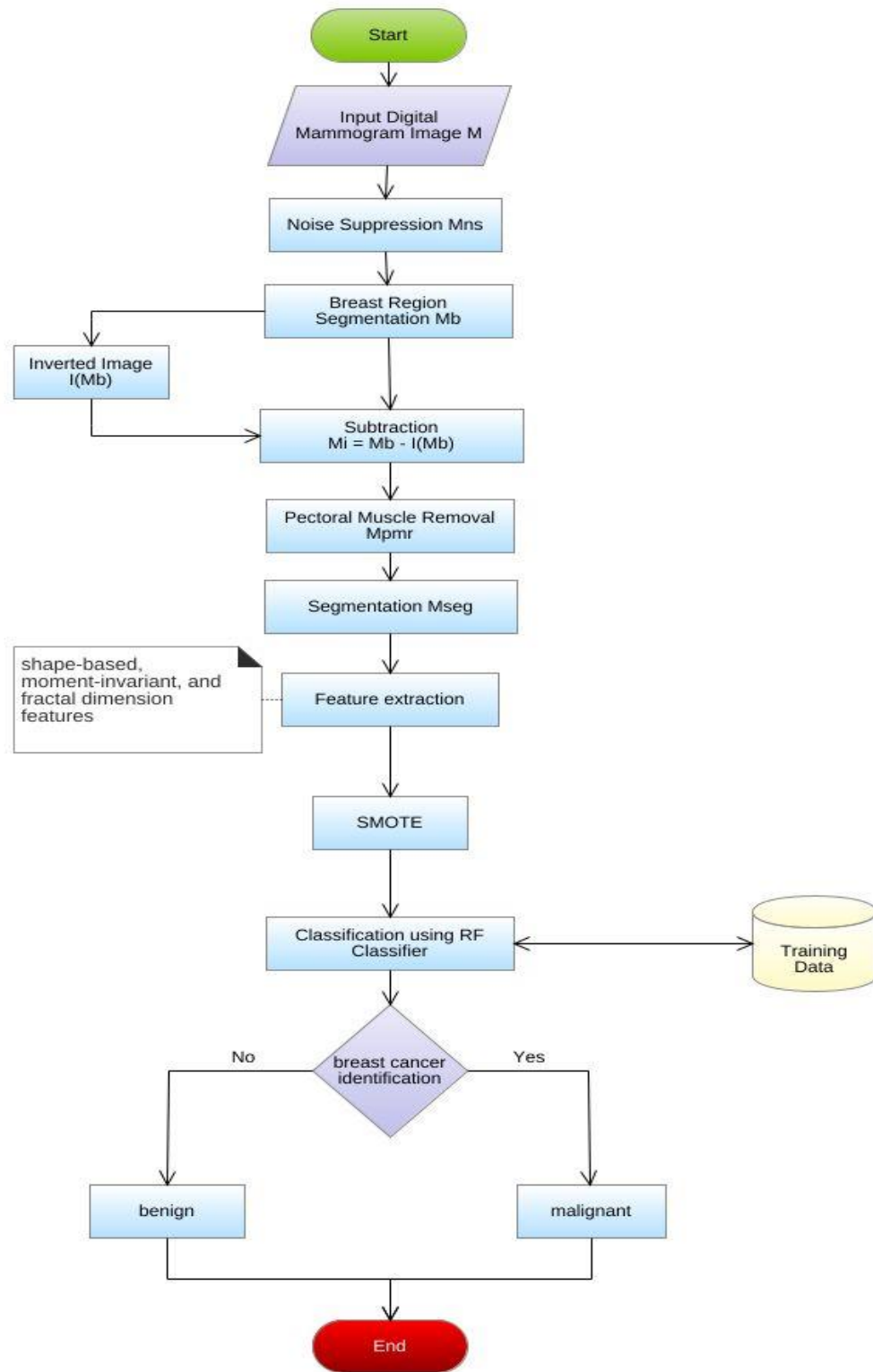


Figure 3. 9 Flowchart of the proposed method

### 3.4.1 Data Acquisition

The mammographic image analysis society (Mini-MIAS) is a freely accessible dataset for scientific purposes. The proposed model makes use of 322 mammography images, 280 of which are benign, while the remaining 42 are malignant in MLO views of both the left and right breasts. The database was compressed to a 200 micron pixel border and padded so that all images were  $1024 \times 1024$  pixels in size. The images are saved in grayscale in the JPG format.

### 3.4.2 Pre-processing

Mammogram preprocessing is one of the primary steps in a CAD system. In the preprocessing step, the unwanted objects are removed from the mammograms, which include annotations, labels, and background. The preprocessing helps the localization of region for abnormality search. In mammogram preprocessing, one of the major challenges is accurately defining the pectoral muscle (PM) boundary from the rest of the breast region. The PMs are mostly present in the MLO views of the mammograms.

Original mammogram images are binaries by a special threshold technique (Threshold = 20) that blacks the background only in order to separate it from the mammogram image content. Then, a morphological erosion operation is performed to remove small regions. Two sets of data are fed into the erosion operator. The first step is to erode the image. The second is a set of coordinate points known as a structuring element (also known as a kernel ( $3 \times 3$ )), which specifies the precise effect of the erosion on the input image. The filter assigns the minimum value of the surrounding pixels to each pixel of the resultant image. Surrounding pixels, which should be processed, are specified by the structuring element: (1) to process the neighbor, (-1) to skip it.

### 3.4.2.1 ROI Segmentation

In the cropping steps, the process of determining the breast region (region of interest) is carried out, and it depends on the direction of the breast and the intensity of the pixels, which is in two steps: The first is to determine the beginning and end of the breast region (left and right), where the maximum width of the breast is adopted. The second step includes determining the (top and bottom) of the breast region, at the top, the highest point belonging to the breast region is approved. As for the bottom, the end of the breast is determined and approved as a lower point for the region of interest. Thus, the area of the breast that is confined between the four points is adopted, and the cropping is performed.

### 3.4.2.2 Pectoral Muscle Removal

To produce a subtracted image, the obtained breast image was inverted. Then the segmented breast region is subtracted from the inverted image. Depending on the direction of the breast, the pectoral muscle appears in the left-top corner or right-top corner of MLO images, and in comparison to the main breast tissue, they have higher density values. In order to improve the segmentation process and the classification performance, an automated method for pectoral muscle removal is presented. A mammogram is clustered with the k-means clustering algorithm ( $k=7$ ) before applying the steps of the procedure for removing the pectoral muscle that has been proposed.

The method is divided into three phases: The alignment of the breast is determined in the first phase; mammograms are either right or left-aligned. The orientation is determined by examining the pixels intensities at the top of the image (left and right), the intensity values are compared on both sides to determine the orientation of the breast. Where the direction of the breast is opposite to the direction that contains the pixels with higher intensity. In the

second phase, a triangle is placed over the mammogram image to isolate the main breast region from the pectoral muscle.

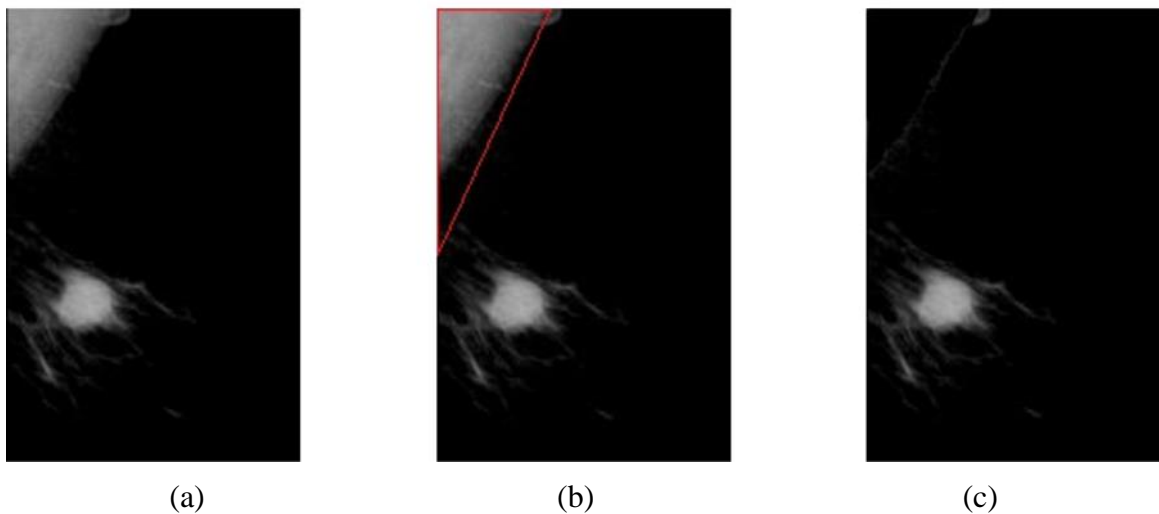
Determining the dimensions of the triangle and its three points depends on the direction of the breast and perform as follows:

- The first point (beginning): If the direction of the breast is to the right, the pectoral muscle will be located at the top left of the image, then we will take the first point from the top left corner of the image, and vice versa.
- The second point (muscle width): to determine the second point of the triangle, the intensity of the pixels at the top of the image (horizontally) was checked according to the direction of the breast. In the case where the direction of the breast is to the right, the examination of the pixel intensity is carried out from the right. If the intensity of the pixel is increased from lower to higher, this will be the beginning of the pectoral muscle area and the second point of the triangle.
- The third point (muscle length): to determine the third point of the triangle, the intensity of the pixels along the length of the image (vertically) was checked according to the direction of the breast, where the pectoral muscle will be located on the opposite side of the breast direction, and its value depends on the value of the second point. In the case where the direction of the breast is to the right, the examination of the pixel intensity is carried out from the top left corner toward the bottom. If the intensity of the pixel is decreased from higher to lower, this will be the end of the pectoral muscle area and the third point of the triangle. Since the length of the triangle = the width of the triangle / 100 \* the length of the image.

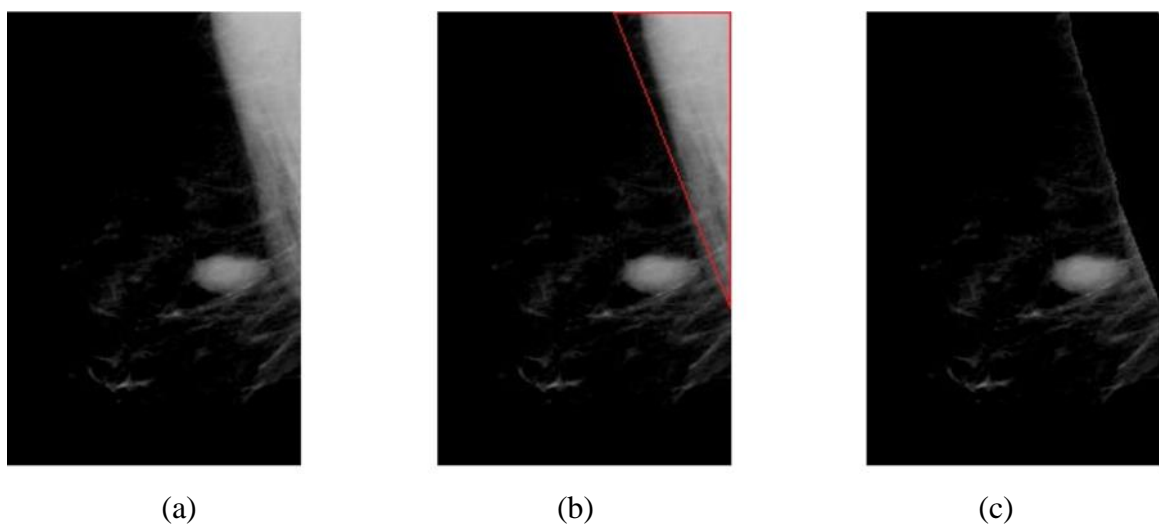
Then to suppress the pectoral muscle, a seeded region growing algorithm is being used in the third phase, which will be applied only to the triangular defined in the second step that is used in this procedure. The similarity criteria

will be the higher intensity of pixels that belonged to the defined triangular minus 120.

Figures 3.10 and 3.11 shows the process of removing the pectoral muscle in two cases (left and right), where (a) shows the subtracted image, then (b) defines the triangle resection, and (c) the image of the breast after removing the pectoral muscle. To eliminate the pectoral muscle we have applied a region growing method within the red triangle as seen in Figure 3.10.b, 3.11.b.



**Figure 3. 10** Left pectoral muscle removal process: (a) Subtracted image (b) Defines the truncating triangle (c) Pectoral muscle removal



**Figure 3. 11** Right pectoral muscle removal process: (a) Subtracted image (b) Defines the truncating triangle (c) Pectoral muscle removal

### 3.4.3 Segmentation

The primary goal of segmentation is to simplify the image so that it can be easily analyzed. The ROI was extracted from the original mammography image by suppressing the entire breast while excluding the pectoral muscle and any other artifacts. The final segmented unlabeled image is generated by superimposing the remaining breast area on the subtracted image. Segmentation is carried out to divide the images into uniform areas and extract the ROI. The segmentation process consists of two steps: in the first step, k-means clustering is applied to the obtained image after pectoral muscle removal. Then the thresholding algorithm is applied with an intensity threshold equal to 216. As a result of the threshold, each region with an intensity value above the threshold becomes white, and each region with intensity below the threshold becomes black. Finally, the suspicious mass area in the image is obtained, which represents the white area.

### 3.4.4 Features Extraction

In this process, the region of interest (ROI) is extracted for analyzing the image. It includes modifying the image from the lower level of pixel data into higher level representations. From these higher level representations we can gather useful information. Several types of image features have been examined and evaluated for classification applications. Fractal dimensions using pixel range calculation methodology, with three shape-based features i.e., eccentricity, solidity, and extent of the segmented region are collected. Then from the segmented suspicious mass region, seven-moment invariants are obtained. A feature vector is generated using a collection of shape-based, Hu moment-invariant, and fractal dimension features derived from each segmented suspicious area.

Hu moments (or rather Hu moment invariants) are a set of seven numbers calculated using central moments that are invariant to image transformations.

The first 6 moments have been proven to be invariant to translation, scale, rotation, and reflection, while the 7th moment's sign changes for image reflection.

In fractal dimension, the image of size  $M \times M$  is divided into different box lengths  $L$ . The size  $L \times L$  boxes with gray level  $L'$  are determined.

$N_r$  is a 2D matrix that stores the number of boxes and their gray levels.

The image's fractal dimension is precisely dependent on the size of the box. The reduction factor and gray level can be computed based on the box size as follows:

$$r = L / M \quad (3.21)$$

$L/M$  represents the ratio of the box length to the image length. As a result, the reduction factor will be:

$$1/r = M / L \quad (3.22)$$

Any image is represented by the cell counting method by the number of boxes with their gray levels. Here,  $L'$  is the gray level that can be possible with  $L \times L$  box size, which can be calculated as follows:

$$\text{Gray level } L' = L \times G / M \quad (3.23)$$

Where  $G$  has a value of 256

$$R_i = P_i(\max) - P_i(\min) + 1 \quad (3.24)$$

Where  $R_i$  is the range of intensities and  $P_i$  is the pixel intensity.

The gray levels covered by the intensity range can be determined as follows:

$$N_r = R_i / L' \quad (3.25)$$

$$FD = \log(N_r) / \log(1/r) \quad (3.26)$$



Where  $\log(Nr)$  is the  $\log$  of whole box covered by an image,  $\log(1/r)$  is the reduction factor based on box length, and  $FD$  is the image's fractal dimension of images.

---

**Algorithm 3.3** Pixel Range Calculation (PRC)

---

**Input:** Image

**Output:** Estimated Fractal dimension

**Begin**

- Step 1**      Determining the length of the box.
- Step 2**      Compute the reduction factor.
- Step 3**      Separate each image into boxes.
- Step 4**      Compute the gray level.
- Step 5**      Determin the gray intensity range for each box.
- Step 6**      Divide the gray intensity range for each box using the gray level.
- Step 7**      Determine the total box count.
- Step 8**      Divide the total box count by the reduction.

**End**

---

### 3.4.5 Balancing the Data

In the Mini-MIAS dataset, the number of observations coming from the malignant class is substantially less than those belonging to the benign class. The predictive model created utilizing typical machine learning methods may be biased and inaccurate in this situation. This happens because machine learning algorithms are normally designed to increase accuracy by reducing the number of errors. As a consequence, they don't consider the distribution of classes, their proportions, or the balance of classes. Because of that, employ synthetic minority over-sampling technique (SMOTE) to cover the impact of the

imbalance on the classifier's accuracy. A subset of data from the minority class is selected, and then new synthetic comparable instances are created. After that, the initial dataset is supplemented with these synthetic instances. The SMOTE algorithm is detailed below [142]:

- For each sample, find the  $k$ - nearest neighbours, ( $k=3$ ).
- Choose samples from a  $k$ -nearest neighbour randomly.
- Find new samples by multiplying the initial samples by the difference and by the gap of  $(0, 1)$ .
- Add the additional samples to the minority. Then, a new set is generated.

### Algorithm 3.4 SMOTE

---

**Input:** Number of minority class samples  $z$

Amount of SMOTE  $N\%$

Number of nearest neighbors  $K$

**Output:**  $(N/100) * z$

#### Local Variables

$N = (N/100)$  (*The amount of SMOTE is assumed to be in integral multiples of 100*)

$K =$  Number of nearest neighbors

$attrnum =$  Number of attributes

$z[ ][ ]:$  array for original minority class samples

$index:$  keeps a count of number of synthetic samples generated

$NewSynthetic[ ][ ]:$  array for synthetic samples

#### Step 1

If  $N < 100$

then Randomize the  $z$  minority class samples

$z = (N/100) * z$

```

        N = 100
    Endif

Step 2
    for i ← 1 to z
        Compute K nearest neighbors for i, and save the indices in the
        temparray
        Populate(N, i, temparray)
    Endfor

Step 3
    Populate(N, i, temparray) (Function to generate the synthetic samples)
    while N != 0
        Choose a random number between 1 and K, call it  $K_n$ . This step
        chooses one of
        the K nearest neighbors of i.
        for j ← 1 to attrnum
            Compute:  $dif = z[temparray[K_n]][j] - z[i][j]$ 
            Compute: gap = random number between 0 and 1
            newSynthetic[index][j] =  $z[i][j] + gap * dif$ 
        endfor
        index++
        N = N - 1
    Endwhile

Step 4
    Return newSynthetic

```

---

### 3.4.6 Random Forest (RF) Classifier

A random forest consists of a large number of individual decision trees that operate as an ensemble, and each tree in the RF generates a class prediction. RF consists of a set of tree-structured classifiers. The number of trees needed in the RF depends on the number of rows in the dataset. The rows in our dataset are 322 and after applying the SMOTE, the dataset became 552. The number of trees considered in our model is 10. Each tree performs a unit vote for the most

common class, and the class with the most votes becomes our model's prediction. The steps involved in the RF classifier are as follows:

**Step 1**  $n$  number of random records are taken from the dataset having  $k$  number of records.

**Step 2** Individual decision trees are constructed for each sample.

**Step 3** Each decision tree will generate an output.

**Step 4** Final output is considered based on majority voting or averaging for classification and regression respectively.

Since the classifier is conditioned by a given dataset, high classification accuracy can only be achieved for the training set, not yet for other independent datasets. We'd like to incorporate cross-validation into our system to eliminate the overfitting. Cross-validation will not improve overall classification precision, but it will make the classifier more accurate and allow it to be expanded to additional separate datasets.

To train the classification models, the newly generated dataset from the SMOTE is used for this purpose. The 5-fold cross-validation method trains the random forest classifier for each fold using 80% of data for training and 20% for testing. Each partition used for training and testing in each iteration is always different from the others. The random forest classifier identifies the segmented region as benign or malignant. The pseudocode of the random forest classifier is as follows:

### **Algorithm 3.5** Random Forest Classifier

---

**Input:** Dataset  $\mathbf{D}$

**Output:** Confusion Matrix, Validation

**Step 1** (Divide the dataset into training and testing)

**Training**  $\leftarrow$  Split ( $\mathbf{D}$ , size = 0.8)

**Testing**  $\leftarrow$  Split ( $\mathbf{D}$ , size = 0.2)

**Step 2** (Algorithm training)

**Teacher** = new RandomForestLearning()

{

    NumberOfTrees = 10, // use 10 trees in the forest

};

**Step 3** (Breast cancer prediction for the testing dataset)

**Forest** = **Teacher.Learn(Training, Training\_Output)**

**Prediction** = **Forest.Decide(Testing)**

**Step 4** (Extract classification results)

Calculate **Scores(Prediction, Testing\_output)**

Compute **Confusion Matrix(Prediction, Testing\_output)**

**Step 5** (Evaluation of results)

**Validate Model**

---

# *Chapter Four*

## *Results and Discussion*

## 4.1 Introduction

The experiments that have been designed to evaluate the suggested algorithms are illustrated in this chapter, and the initial results are then presented. The techniques and methods covered in the previous chapter have been tested on two datasets with various abnormalities, and the results have been evaluated and compared using several metrics. The first proposed model is a multi-points (seeds) region growing method for ROI segmentation are developed to classify the benign and malignant of mammogram images for breast cancer. The second model is pectoral muscle removal and solving the problem of data imbalance.

All practical results of the application of the two proposed models are presented and discussed in this chapter. Different experiments were carried out under different parameters and image processing characteristics. The results are discussed and evaluated in this chapter. All the experiments are performed on a computer with these attributes: Dell G5 15, Windows 11 Home 64-bit, Intel(R) Core i7-8750H CPU @ 2.21GHz, 2.40GHz, 8GB RAM. The proposed models were implemented in Visual Studio.Net framework 2019 uses the C# language, and the detailed figures of the proposed models are located in the appendices A and B.

## 4.2 Evaluation Results of the Proposed Models

Micro-calcification and mass are two common early signs of malignancy that help to identify a breast cancer. However, because of a variety of factors, including low contrast and the presence of pectoral muscle in the breast region, these automated methods detected a large proportion of false positives.

This dissertation presents two models of mammogram segmentation and breast cancer detection. The first model is based on the proposed multi-points (seeds) region growing method and SVM classifier. The second model involves a pectoral muscle removal approach with RF classifier and SMOTE technique.

### 4.3 Proposed Model 1: Evaluation Results of the Segmentation using K-means Clustering and Optimized Region-growing Technique

The proposed system initially, using a bilinear interpolation approach, reduces all instance images to  $256 \times 256$  pixels. Then the Gaussian filtering is utilized to preprocess the mammogram images, followed by the denoising step that consists of isolating the image background from its components using k-means clustering, eliminating noise objects, and breast area recovery. Then the optimized region growing used to segment ROI includes the MCs. The work deals with the extraction of features from segmented areas to detect and classify mammogram images as benign and malignant with the SVM classifier.

#### 4.3.1 Dataset

The proposed system processes the images of the left and right breast in craniocaudal (CC) views obtained from the digital database for screening mammography (DDSM) dataset. In this work, 440 images were used, including 329 benign cases and 111 malignant cases in the CC view. The images were randomly selected from the CBIS-DDSM dataset and were divided in advance into training and testing. For training, 355 images were used, and 85 images were utilized to assess the proposed method. Table 4.1 shows the selection and distribution of samples from the CBIS-DDSM dataset.

**Table 4. 1** The selection and distribution of samples from the CBIS-DDSM dataset

No. of selected images	Training		Testing	
440 Images selected randomly	355		85	
	Benign	Malignant	Benign	Malignant
	264	91	65	20



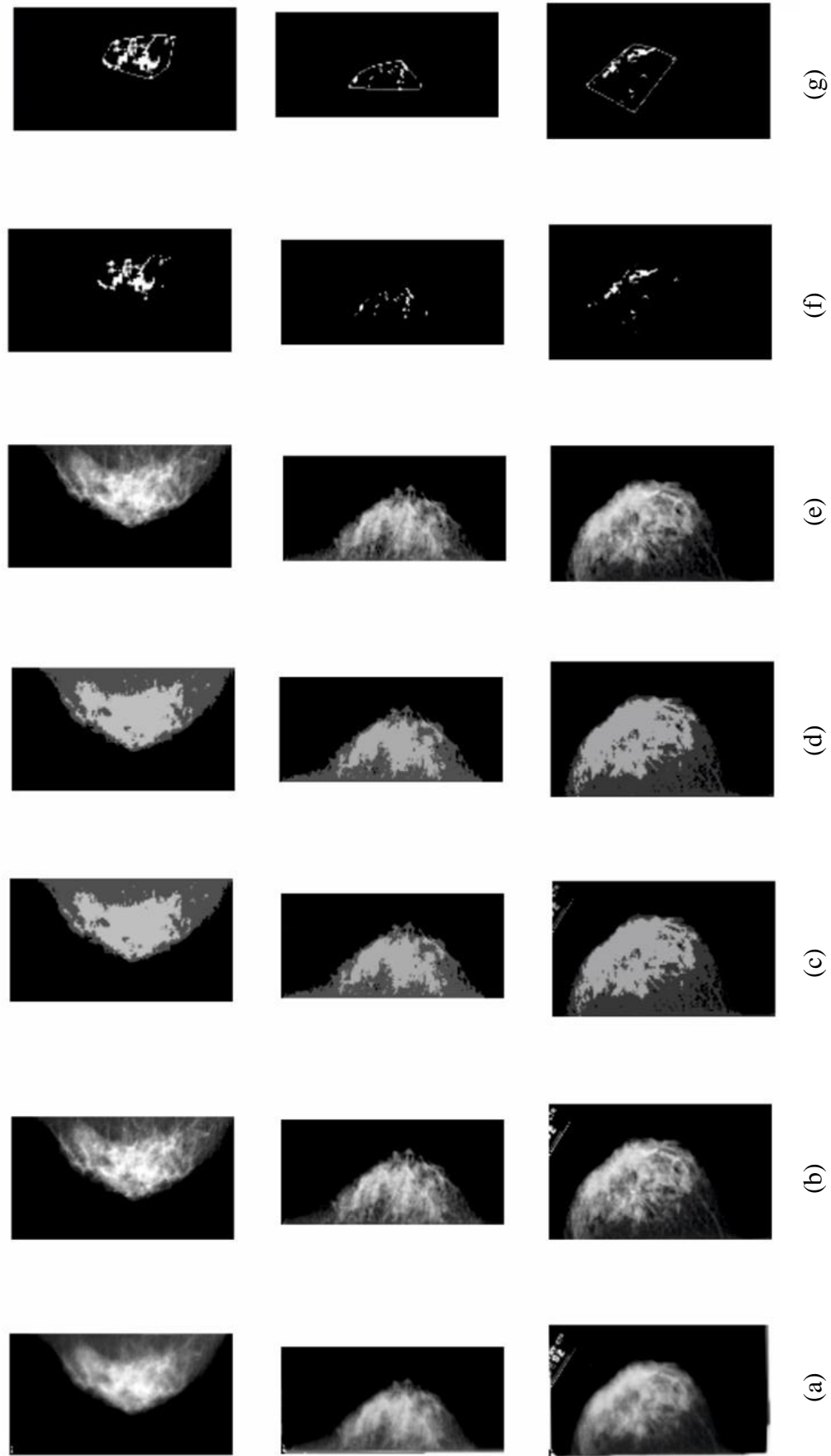
### 4.3.2 Preprocessing Results

The proposed method preprocesses digital mammograms using a Gaussian filter, which removes noise and softens the images. Figure 4.1.a is the original image, and Figure 4.1.b shows the result of applying a Gaussian filter to the image.

Medical images may include symbols, phrases, or characters that indicate the image's type or part of its medical-physical characteristics. This is commonly referred to as "image noise," and it can have an effect on classification accuracy. Denoising was used to solve these problems. Figure 4.1 (c, d, and e) shows the three stages of the denoising process (k-means clustering, noise object elimination, and recovering the important area).

### 4.3.3 Segmentation Results

The proposed segmentation algorithm presents an efficient pixel-based approach for growing regions that produces optimal seeds and thresholds. Multi-points (seeds) have been used to detect calcifications if they are in more than one area of the breast, which cannot be determined using the traditional region-growing algorithm due to the nature of the calcifications, which may be sporadic. Figure 4.1(f and g) shows the micro-calcifications (MCs) segmented image.

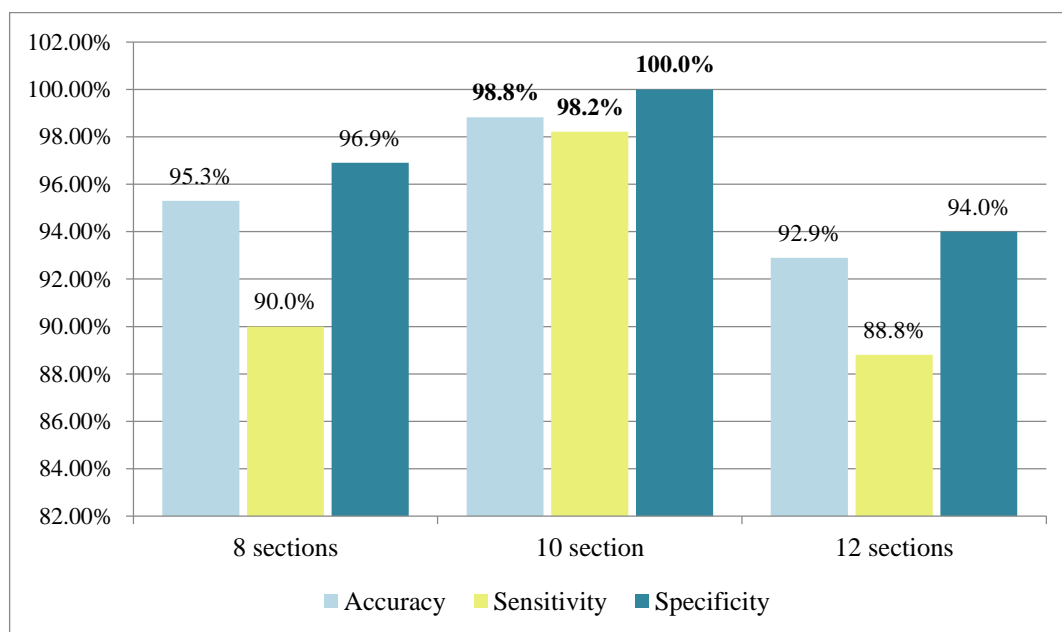


**Figure 4. 1** The pre-processing and segmentation Process: a) Original image, b) Gaussian filter, c) Applying k-means, d) Erosion filter, e) Breast area retrieval f) Segmented area, g) Micro-calcifications (ROI)

In the segmentation step, we proposed an optimized region-growing algorithm, and the sorted mammogram value was divided into 10 sections to prove that these sections have obtained better results compared with the two other examined sections (8 sections and 12 sections). As shown in the Table 4.2, and Figure 4.2, the result for the 10 sections exceeds the other two examined sections.

**Table 4. 2** Comparison of experimental results for the proposed method

Optimized Region Growing	10 Sections	12 Sections	8 Sections
Accuracy	98.82%	92.90%	95.30%
Sensitivity	98.2%	88.80%	90.00%
Specificity	100.00%	94.00%	96.90%



**Figure 4. 2** Comparison of experimental results for the proposed method

#### 4.3.4 Classification Results

To evaluate the proposed models, evaluation metrics such as accuracy, sensitivity, and specificity are required. For the calculation of these metrics in the classification process, the confusion matrix distinguishes the terms TP, TN, FP, and FN from the predicted and ground truth result. The confusion matrix indicates how accurate our models are at estimating and how often they predict erroneously. As shown in Table 4.3 false positives and false negatives were attributed to values that were incorrectly predicted, whereas true positives and true negatives were assigned to values that were correctly predicted.

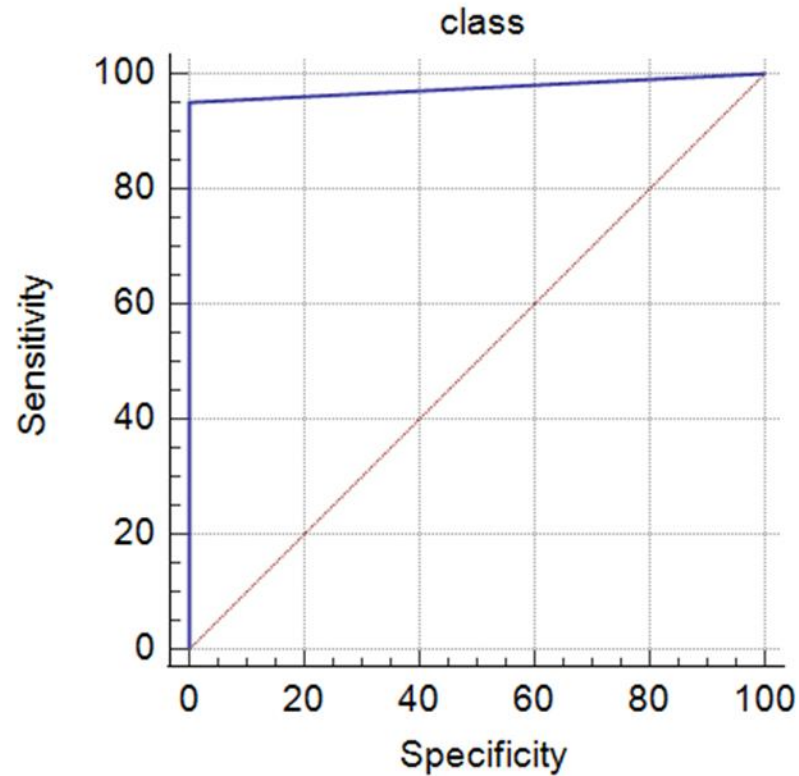
**Table 4. 3** Confusion matrix for the proposed method

Predicted	Actual class		
	Positive	TP = 55	FP = 0
	Negative	FN = 1	TN = 29

To measure the efficiency of the proposed method outlined in the model, accuracy, sensitivity, and specificity are calculated utilizing Eqs. (2.9, 2.12, and 2.13) as described in Chapter 2. The proposed method was validated using SVM for the desired results. Furthermore, the proposed method achieved significant accuracy in classifying the CBIS-DDSM dataset. It achieved good results with the proposed multi-points (seeds) region growing method and SVM, showing 98.2% sensitivity, 100% specificity, and 98.82% accuracy in identifying both benign and malignant samples.

A ROC curve is created by combining the sensitivities and specificities for various values of a continuous test measure. This gives a list of different test values, as well as the sensitivity and specificity of the test at those levels. For each of the tabulated data, the ROC curve is generated by plotting sensitivity on

the y-axis versus  $1 - \text{specificity}$  on the x-axis. As a result, a diagnostic test with adequate accuracy should have a ROC curve in the top left triangle. The ROC curve area of the proposed method was validated, as shown in Figure 4.3.



**Figure 4. 3** ROC curve of the classification results

**Table 4. 4** Comparison of the existing techniques with the proposed system

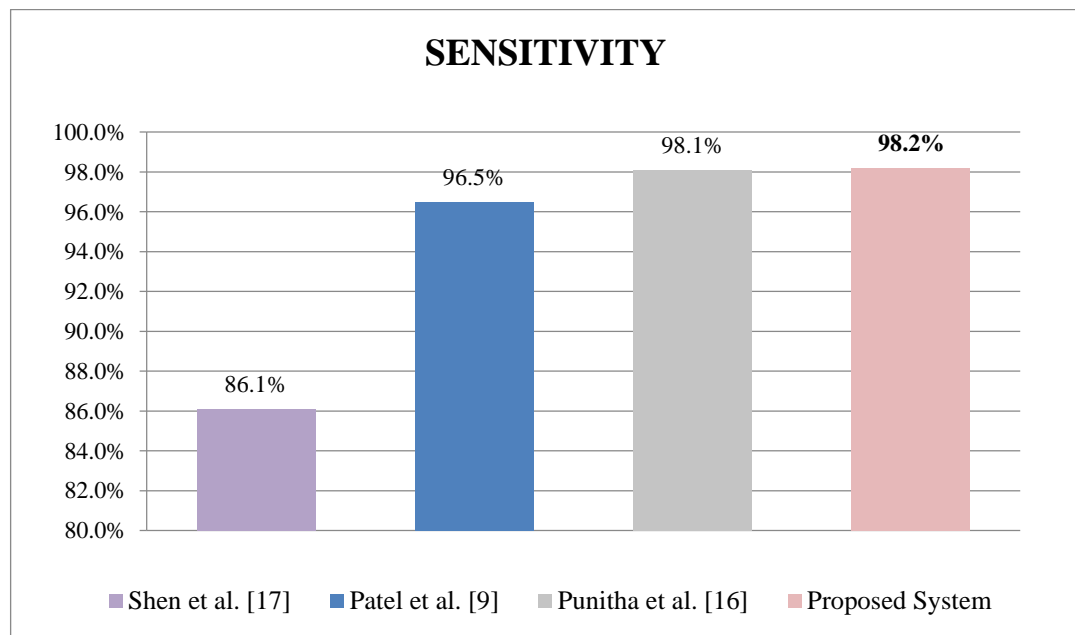
Reference	Segmentation	Features	Classifier	Dataset	Sensitivity	Specificity	Accuracy
Rouhi et al. [13]	Region Growing optimized using GA Adaptive Threshold method	GLCM , contour-related, and Morphological features	Random Forest, NB, SVM, and KNN	MIAS and DDSM	96.87%	95.94%	96.47%
Patel et al. [9]	Region Growing	Shape and texture	Multilayer Perceptron Neural Network	DDSM	96.5%	89%	95.6%
Shen et al. [17]	-	Pixel-level annotations	CNN	DDSM	86.1%	80.1%	-
Xie et. al. [14]	level set model	Gray-level features and textural features	ELM and SVM	Mini-MIAS+DDSM	-	-	96.02%
Punitha et al. [16]	Dragonfly Region Growing Optimization	GLCM and GLRLM Texture features	FFNN using Backpropagation	DDSM	98.1%	97.8%	98%
Proposed System	Multi-points (seeds) Optimized Region Growing	Cross-correlation coefficient, Pearson Correlation, the average area of segmented spots and texture features based on Haralick definitions	<b>SVM</b>	<b>DDSM</b>	<b>98.2 %</b>	<b>100 %</b>	<b>98.82 %</b>

The experimental results of the proposed algorithm are compared with the results of prior research and similar models to classify the breast masses. Our proposed CAD system and the proposed segmentation method outperformed the previous models applied, as shown in Table 4.4.

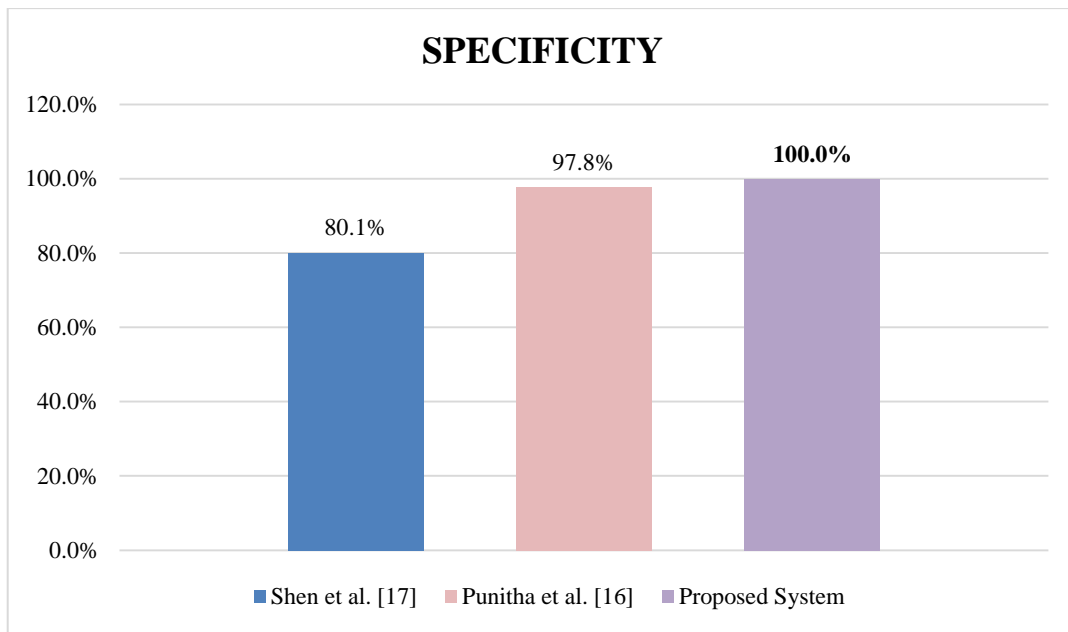
Compared with other techniques, this work exceeded the performance of the work by Rouhi et al. [13], which only achieved an accuracy of 96.47% using optimized region growing and the GA adaptive threshold method on the MIAS

and DDSM datasets. On the other hand, this work also outperformed the work of Patel et al. [9], which reported an accuracy of 95.6%. Moreover, recent works on the DDSM dataset were all surpassed, where the highest accuracy of 98% was reported by Punitha et al. [16] using dragonfly region growing optimization.

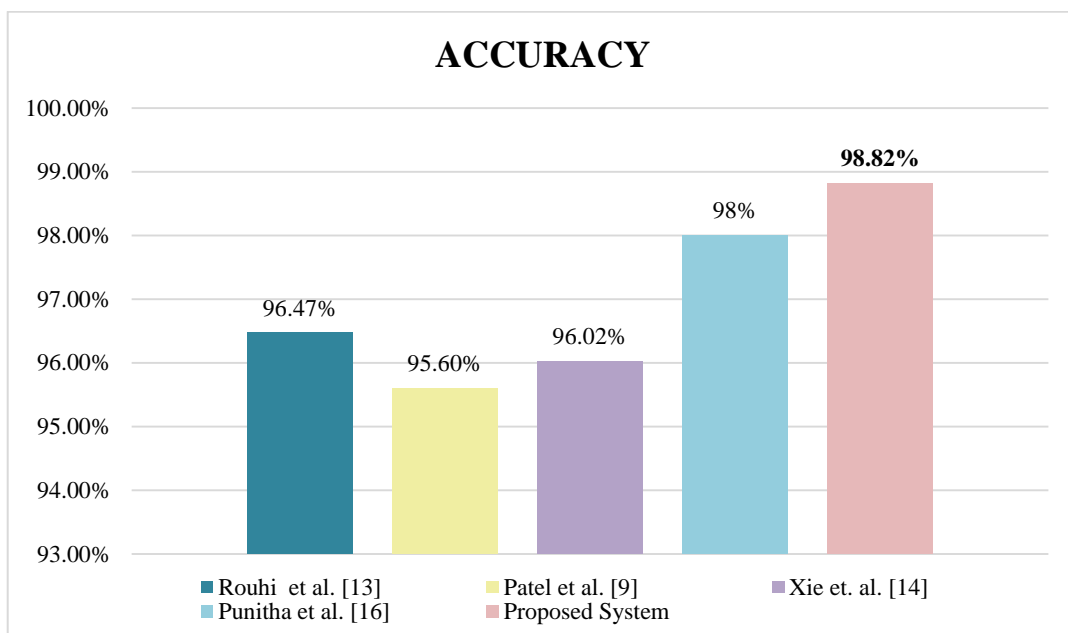
In Figure 4.4 (a, b, and c), the obtained results have shown that the proposed method outperforms other reported literature approaches.



**Figure 4. 4.a** Sensitivity comparison of the existing techniques with the proposed system



**Figure 4.4.b** Specificity comparison of the existing techniques with the proposed system



**Figure 4.4.c** Accuracy comparison of the existing techniques with the proposed system



## 4.4 Proposed Model 2: Evaluation Results of the Pectoral Muscle Removal and Solving the Data Imbalance Problem

This model presents a diagnosis method to detect an abnormality in mammograms automatically. Before abnormality identification, image-processing techniques used to correctly segment the suspicious region-of-interest ROI. The background of the mammograms has darkened to distinguish the breast area from any blemishes or writing that will be removed. Then the breast area is extracted after ignoring the empty regions around the breast in mammogram images. After that, the mammogram image is inverted, and the inverted image is then subtracted from the segmented breast region. For pectoral muscle removal, a region-growing method with the k-means clustering is used. Afterward, the segmented suspicious ROI is extracted utilizing the k-means with thresholding technique. Shape-based features, moment invariants, and fractal dimensions are extracted from the segmented ROI. Due to the imbalance of the Mini-MIAS dataset, the SMOTE algorithm is used to accomplish far better classifier efficiency, which presents new samples from the minority classes to get a balanced dataset. A random forest classifier is utilized to classify the segmented region as benign or malignant.

### 4.4.1 Dataset

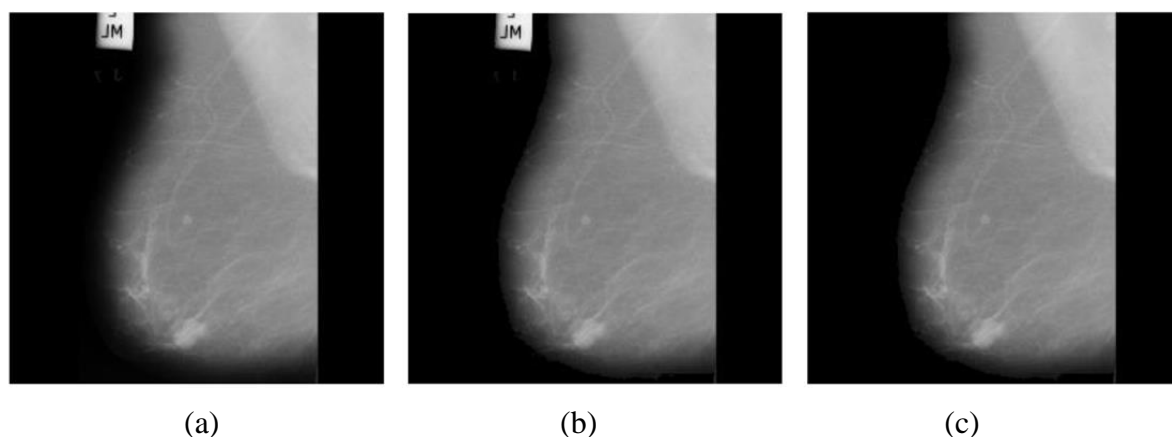
The suggested algorithm is evaluated using 322 mammogram images and processes the images of the left and right breast in MLO views from the mammographic image analysis society (Mini-MIAS) dataset, which is publicly available. All of the images are 1024×1024 pixels in size. Table 4.5 illustrates the distribution of samples in the Mini-MIAS dataset.

**Table 4. 5** Mini-MIAS dataset description

Mini-MIAS dataset number	Character of background tissue	Fatty	Fatty-glandular	Dense-glandular
322		106	104	112
	Severity of abnormality		Benign 280	Malignant 42

#### 4.4.2 Pre-processing Results

Initially, using a bilinear interpolation approach, all instance images are reduced to  $256 \times 256$  pixels. After that, tape artifacts and labels are removed. The intermediate findings of breast area segmentation are seen in Figure 4.5. The initial mammogram, darkening background mammogram after performing special threshold operation, and mammogram after performing morphological erosion operation (enhanced image) are shown in Figure 4.5 (a, b, and c) respectively.

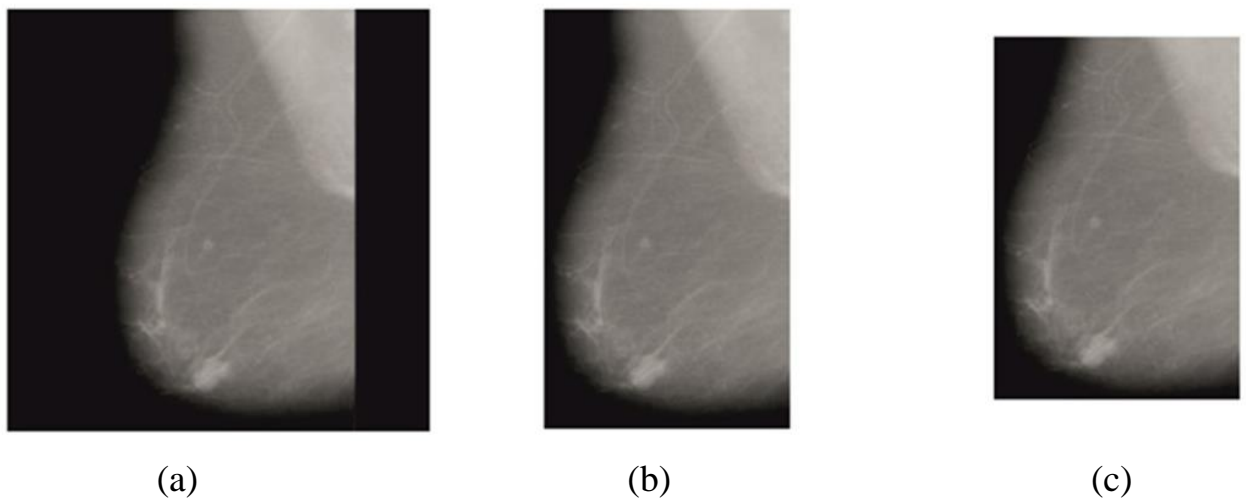


**Figure 4. 5** Breast region identification results: a) Original mammogram, b) Darkening the background, c) Erosion operation

#### 4.4.3 ROIs Segmentation Results

In fact, the primary aim of extracting the ROIs is to investigate their content and composition further. The other goal of this work is to separate breast tissue from the surrounding organs. Because breast cancer most typically arises in cells

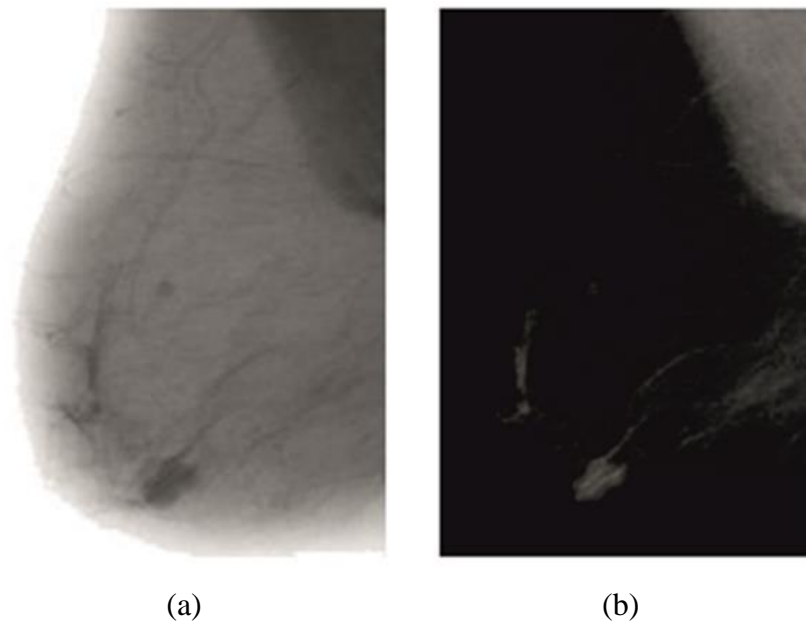
from the lining of milk ducts, based on the entire mammograms in the dataset, the breast tissue is considered ROI. Therefore, it is possible to limit the search area and make object analysis easier to detect suspicious objects or masses, which is the final goal of this work. Figure 4.6 shows the breast region segmentation process. The original revised image, the first step of the breast region segmentation, and the second step of the breast region segmentation, are shown in Figure 4.6 (a, b, and c) respectively.



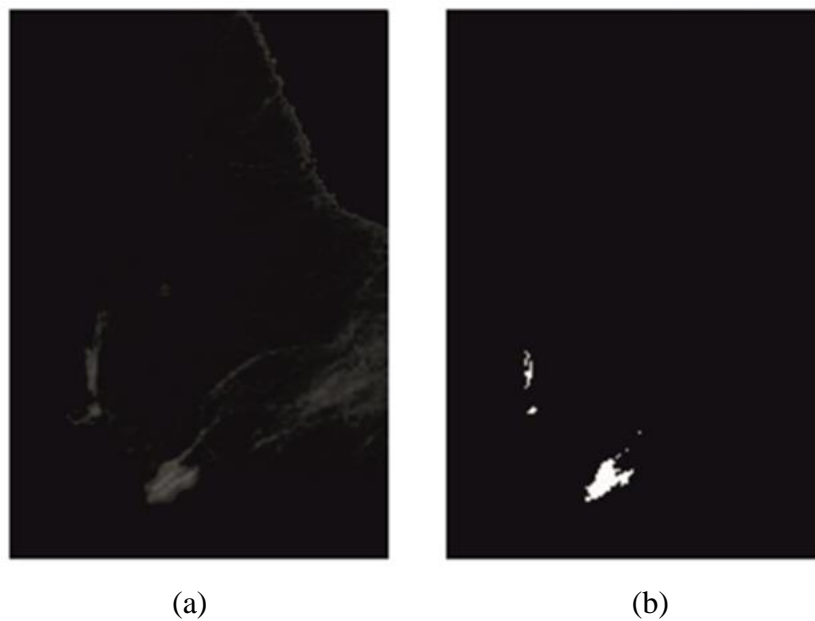
**Figure 4. 6** Breast region segmentation process: a) Original revised image, b) (step 1) breast segmentation, c) (step 2) breast segmentation

#### 4.4.4 Segmentation Results

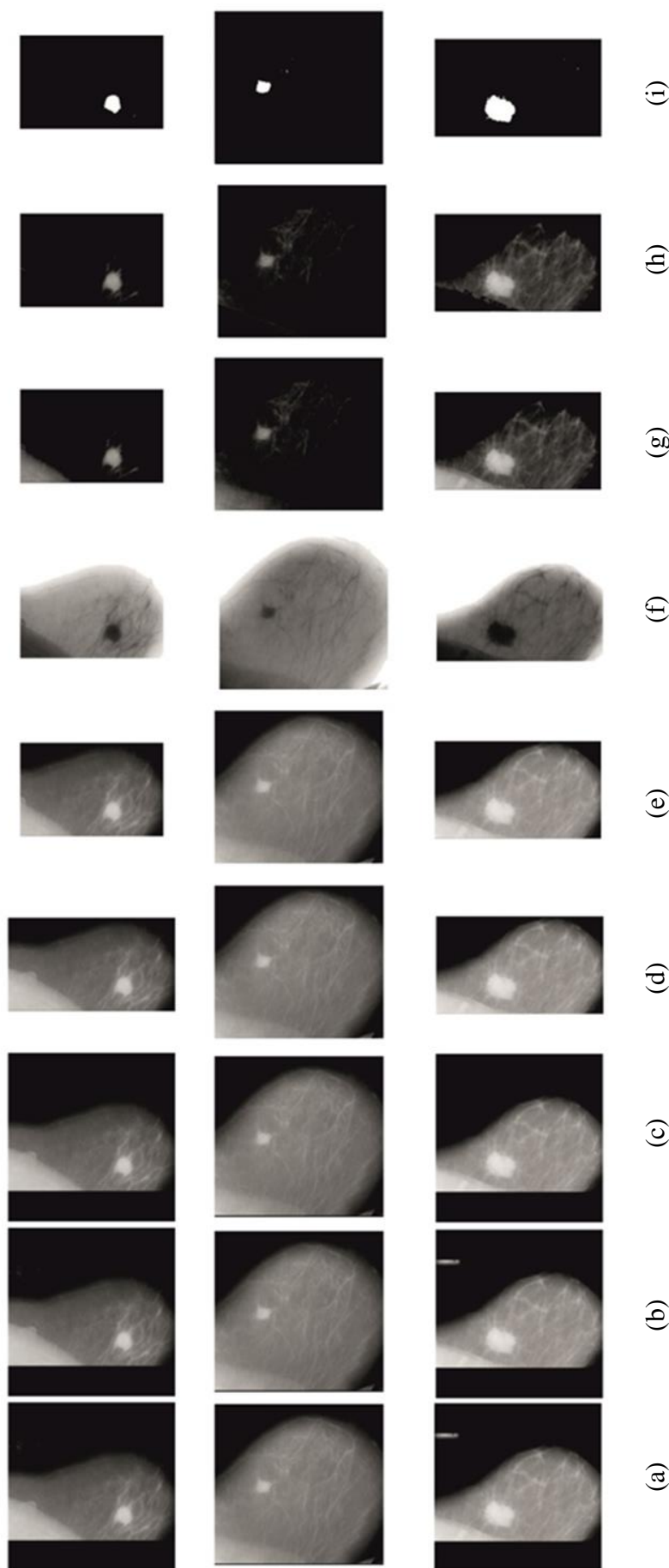
Inverted and subtracted images are shown in Figure 4.7. In the subtracted images, the suspicious mass area is clearly improved. Figure 4.8 illustrates the elimination of the pectoral muscle and the segmentation of suspicious mass regions. Figure 4.9 shows the sample segmented images after applying all of the proposed model's operations.



**Figure 4. 7** Invert and subtract process: (a) Inverted image (b) Subtracted image



**Figure 4. 8** Segmentation process: (a) Pectoral muscle removal process (b) Segmented image

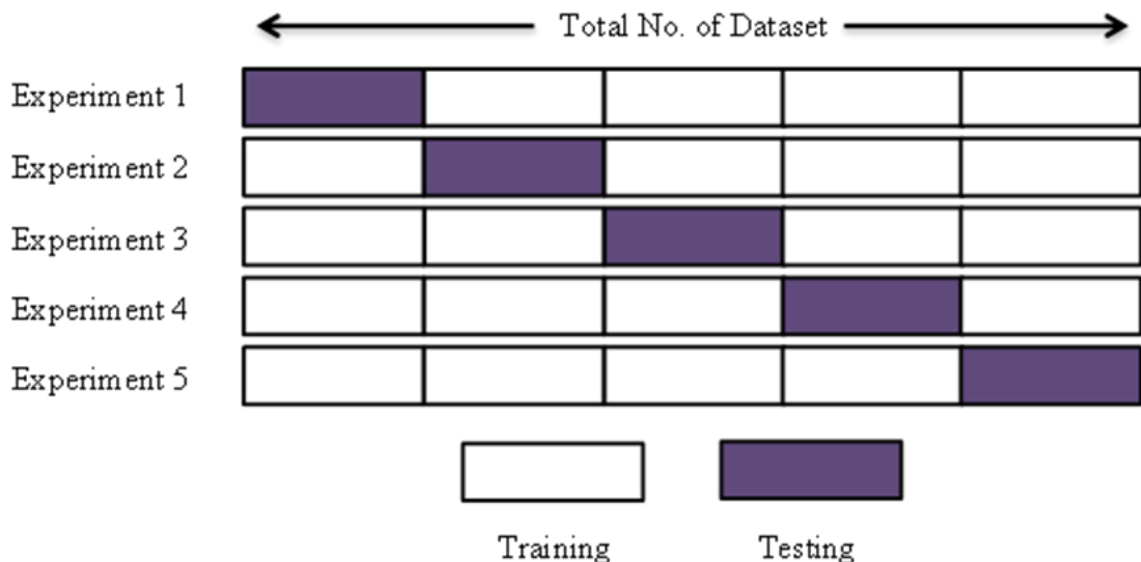


**Figure 4. 9** Segmentation of suspicious regions outputs: a) Original image , b) Binarized image , c) Morphological erosion, d) First step of the breast region segmentation, e) Second step of the breast region segmentation, f) Inverted image, g) Subtracted image, h) Pectoral muscle removal process, i) Segmented image

#### 4.4.5 K-Fold Cross-Validation

In this section, 5-fold cross-validation is applied, and all data is used for training and validation. The method involves partitioning the entire dataset into 5-folds, repeating 5 times for training using 4 folds and a left fold for validation, and then averaging the error rates of the 5 experiments, as illustrated in Figure 4.10.

5-folds are mainly partitioned at random, but some folds may have a slightly different distribution than others. Stratified 5-fold cross-validation was also employed, with each fold having an equal class distribution.



**Figure 4. 10** A 5-fold cross-validation

#### 4.4.6 Classification Results

Three types of features are derived from the segmented suspicious mass region: seven invariant moments, a fractal dimension using pixel range calculation, and the three shape-based features (eccentricity, solidity, and extent). The feature extraction process ends after the completion of the

extraction of the above features for all the mammograms in the Mini-MIAS dataset.

As we mentioned earlier, the number of mammograms coming from the malignant class is much lower than those from the benign class. To fix this weakness, the SMOTE algorithm was used to produce a balanced dataset. Table 4.6 shows the selection and distribution of samples before and after applying the SMOTE algorithm.

**Table 4. 6** Mini-MIAS dataset samples distribution before and after applying the SMOTE

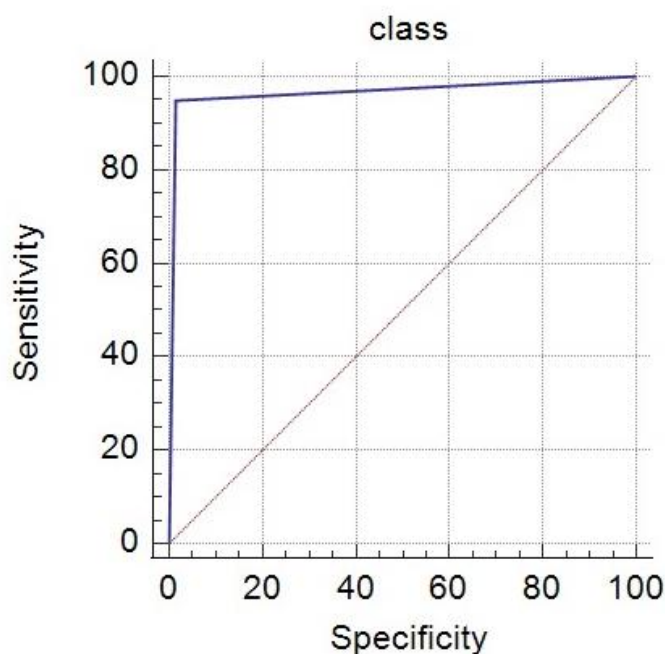
No. of Mini-MIAS samples	No. of majority samples (Benign)	No. of minority samples (Malignant)	No. of minority samples after SMOTE	No. of balanced dataset samples	5-fold cross validation	
					Training	Testing
322	280	42	272	552	442	110

A random forest classifier was trained with 5-fold cross-validation for the balanced dataset. After training the classifier, the testing set was used to test the performance of the classifier and its ability to accurately classify the mammogram images as either benign or malignant. To evaluate our classifier, a confusion matrix was employed as presented in Table 4.7. Confusion matrices are typically used with class output models. According to the confusion matrix, the correct or positive values are more than the incorrect or negative values making the model more accurate.

**Table 4. 7** Confusion matrix for the balanced dataset of the proposed method

Predicted	Actual		
	Positive	TP = 276	FP = 4
	Negative	FN = 12	TN = 260

In addition, sensitivity, specificity, accuracy, and the ROC curve are all used to assess the classification's efficiency as shown in Figure 4.11. The percentage of actual positives that are correctly identified as malignant is measured by sensitivity, the proportion of actual negatives correctly identified as benign is measured by specificity.



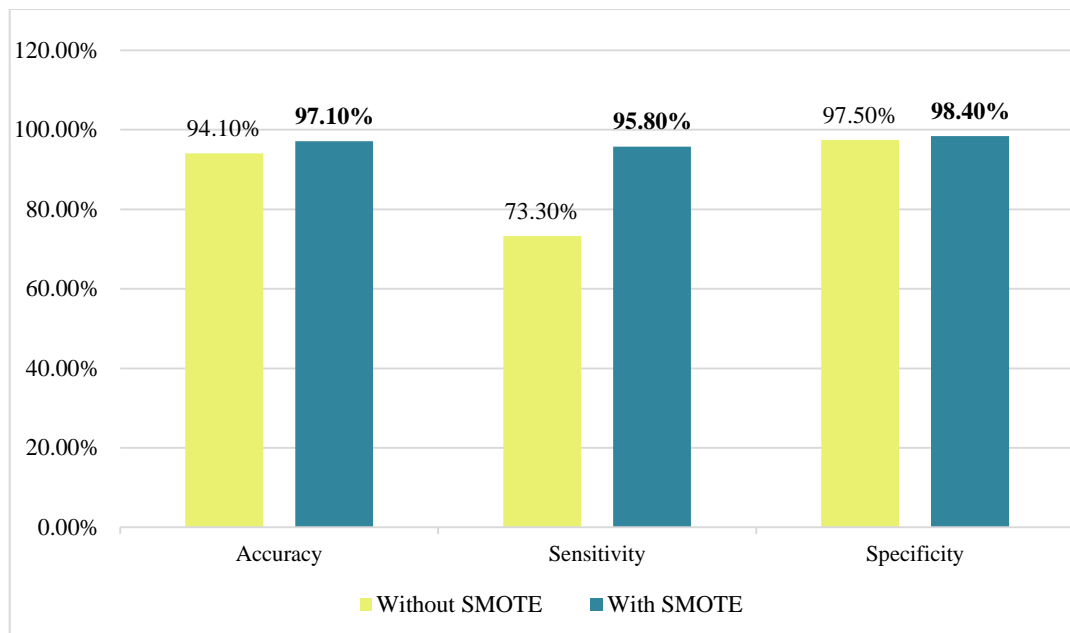
**Figure 4. 11** ROC curve of the classification result

To present the effect of using SMOTE technique on an imbalance dataset, the dataset was examined without using SMOTE technique to evaluate the accuracy, sensitivity, and specificity. As seen in Table 4.8 and Figure 4.12, the results show that the evaluation metrics with a balanced dataset outperformed compared with an imbalanced dataset.

**Table 4. 8** Comparison of evaluation metrics with and without SMOTE

Model 2	Without SMOTE	With SMOTE
Accuracy	94.1%	97.1%
Sensitivity	73.3%	95.8%
Specificity	97.5%	98.4%





**Figure 4. 12** Comparison of evaluation metrics with and without SMOTE

Table 4.9 shows the performance dimension of the proposed algorithm. The results are promising, with a 97.1% accuracy. It performed well in detecting both benign and malignant, with sensitivity and specificity values of 95.8% and 98.4%, respectively. Table 4.10 compares the proposed algorithm's efficiency with the various other existing approaches in this domain.

**Table 4. 9** The performance measures of the proposed algorithm for the Mini-MIAS dataset

Features	Sensitivity	Specificity	Accuracy
Moment invariant, fractal dimension, and Region-based features	95.8%	98.4%	97.1%

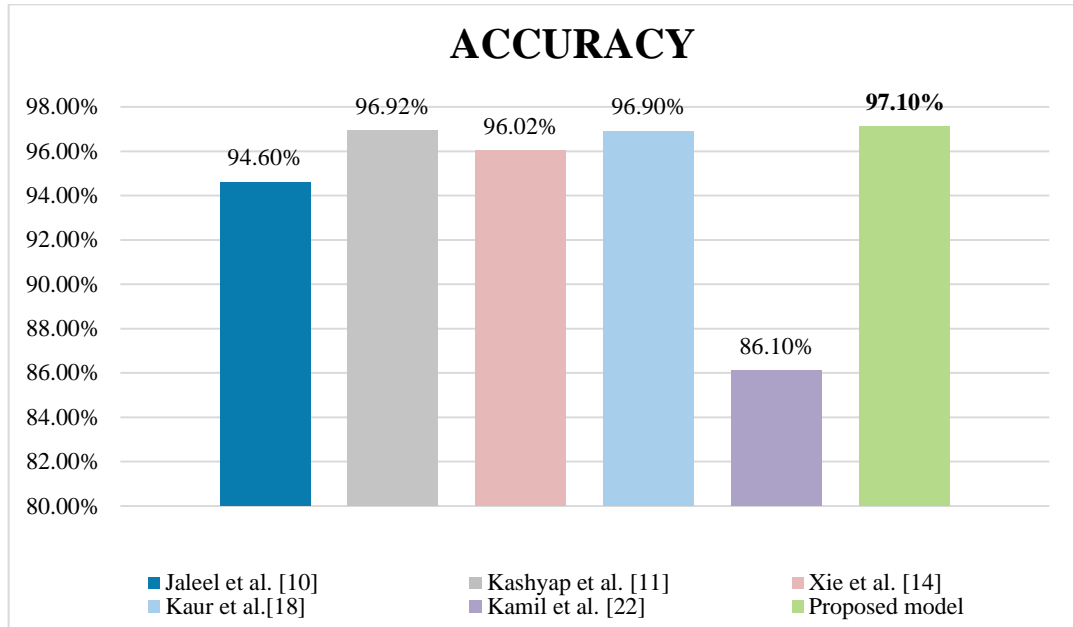
**Table 4. 10** Comparison of the proposed method with the existing techniques

Methods	Dataset	Features	Classifier	Accuracy
Jaleel et al. [10]	Mini-MIAS	Discrete Wavelet Transform (DWT) and GLCM	ANN	93.7% with GLCM and 94.6% with DWT
Kashyap et al. [11]	Mini-MIAS	Moment invariant, Fractal dimension	SVM	96.92%
Xie et al. [14]	Mini-MIAS+DDSM	Gray level features and textural features	ELM and SVM	96.02%
Kaur et al.[18]	Mini-MIAS	SURF	SVM, KNN, LDA and DT	96.9%, 93.8%, 89.7% and 88.7%,
Kamil et al. [22]	Mini-MIAS	Gray Level Co-occurrence Matrix	KNN	86.1%
Proposed model	Mini-MIAS	Moment invariant, Fractal Dimension and Region based	<b>RF</b>	<b>97.1%</b>

Compared with other techniques, we exceeded the performance of the work by Jaleel et al. [10] that only achieved an accuracy of 93.7% with GLCM and 94.6% with DWT on the Mini-MIAS dataset. We also outperformed the work of Xie et al. [14] that reported an accuracy of 96.02%. Recent works on the Mini-MIAS dataset were all surpassed where the highest accuracy of 96.92% was reported by Kashyap et al. [11] using moment invariant, fractal dimension with SVM classifier.

Figure 4.13 demonstrates that the proposed method which includes (noise reductions, breast region differentiation, filtering, pectoral muscle removal, segmentation, feature extraction, balancing data) obtained an accuracy of 97.1%,

the sensitivity achieved is 95.8%, and the specificity achieved is 98.4%, over other popular methods in recent literature.



**Figure 4. 13** Comparison of the existing techniques with the proposed model

# *Chapter Five*

*Conclusions and Future Works*

## 5.1 Conclusions

Computer-aided detection/diagnosis (CAD) system is applied to digital mammography to help radiologists make a fast and accurate diagnosis of breast cancer. The primary purpose of this dissertation was to increase the performance of the CAD system and to accurately identify and detect the abnormality of the breast tissue and determine it as a region of interest. The dissertation proposed two models for detection and classification of breast cancer into benign and malignant using a combination of image processing and machine learning approaches. The findings of the models have led to the following conclusions:

- In the first proposed model, the optimized region growing method was presented, where multi-points (seeds) have been used to detect and segment micro-calcifications (MCs) of mammographic images accurately. The preprocessing phase used a Gaussian filter to remove noise and soften the images in order to obtain a clearer image. Then, in pre-segmentation, the breast area has been isolated from the image using k-means clustering. For the segmentation, an optimized region growing (ORG) approach has been used, where multi-seed points and thresholds are generated optimally depending on the color values of the image pixels. Then, twenty-six texture features based on Haralick's texture analysis and the average area of segmented spots were extracted. Furthermore, two statistical textural analysis features, including the cross-correlation coefficient and Pearson correlation information, were extracted from the comparison of the de-noising image with the segmented image intensities. 440 images from the DDSM dataset of breast micro-calcification in craniocaudal (CC) views were used. This collection comprises 329 benign cases and 111 malignancies in both the left and right breasts. To evaluate the efficiency of the system, a support vector machine (SVM) classifier was utilized to distinguish between benign and malignant tissue. The proposed system's sensitivity reached up to 98.2%, the specificity

obtained was 100%, and the accuracy was 98.82%. The results also show that computer-aided detection/diagnosis is a promising area for reducing mortality through early detection of breast cancer. We can conclude from the first proposed model that finding the various MCs in a mammographic image has an important role in the performance of breast cancer detection systems, and multi-seed segmentation is a possible solution for the purpose of calcification detection.

- The second proposed model focuses on pectoral muscle removal and solving the data imbalance problem. The ROI (breast area) was determined from the image. After that, the obtained breast area was inverted, and the inverted image was then subtracted from the segmented breast area. In the segmentation process, a region-growing method and k-means clustering were performed to isolate the pectoral muscle and ensure that it does not appear, which affects the efficiency of extracting characteristics. Then, a k-means clustering and thresholding technique were utilized to segment the suspicious ROI. A set of eleven texture features were extracted from the ROI, including three shape-based features (i.e., eccentricity, solidity, and extent), seven Hu moment invariants, and fractal dimensions. Afterward, the proposed system solves the imbalance problem in the Mini-MIAS dataset using the SMOTE technique to provide new samples from the minority classes and achieve better classification efficiency. In this model, 322 images were used, of which 280 are benign and the remaining 42 are malignant in MLO views of both the left and right breasts. After applying the SMOTE, the newly generated dataset grew to 552 and was used to train the classification efficiency. The 5-fold cross-validation method trains the RF classifier for each fold using 80% of the data for training and 20% for testing. The experimental results achieved an accuracy of 97.1%, a sensitivity of 95.8%, and a specificity of 98.4%. The proposed method outperformed other widely used studies in recent literature for detecting benign and malignant samples.

From this model, we found that reducing the false positive rates is possible by removing the pectoral muscle, and the proposed region-growing technique is an effective way in this regard. In addition, handling the imbalance data problem results in a higher detection rate, especially for the class with the lower number of samples, and the proposed SMOTE algorithm is an appropriate technique for that.

## 5.2 Future Works

In the future, several suggestions can be taken into account to improve the proposed algorithms. Below are the details of the suggestions:

- In the future, the comparison of the segmentation methods can be implemented based on mass and micro-calcification in mammogram images, as each of these lesions has different methods.
- To improve the classifier performance for detecting and diagnosing masses and micro-calcifications, different screening modalities with machine learning algorithms can be utilized.
- In the future different mammography images view such as MLO and CC can be utilized.
- Implementing methods based on deep learning may enhance classifier performance for the detection of small obscured masses.
- Deep learning methods combined with contrast-enhanced digital mammography can also increase the efficacy of current diagnostic techniques.

Overall, it is crucial to continue working on the suggested system to enhance and aid in the research of breast cancer, developing algorithms that may support experts and reduce their examination time and subjectivity. Thus, the approach displayed promising results, and it is suitable to continue working on it for further improvement in the classification and identification of lesions.



## Publications

### (Scopus)

1. Srwa H. A., Ali M. S. and, Hadi V. “Breast cancer segmentation using K-means clustering and optimized region-growing technique.” , *Bulletin of Electrical Engineering and Informatics*, Vol. 11, No. 1, pp. 158-167, February 2022.

### (Scopus)

2. Srwa H. A., Ali M. S. and, Hadi V. “Improving Breast Cancer Classification Using (SMOTE) Technique and Pectoral Muscle Removal in Mammographic Images.”, *Mendel*, ISSN: 1803-3814 (Printed), 2571-3701 (Online), <https://doi.org/10.13164/mendel.2021.2.036>.

### (Scopus)

3. Srwa H. A., Ali M. S. and, Hadi V.” Breast Cancer Classification Using Machine Learning Techniques: A Review.”, *Turkish Journal of Computer and Mathematics Education*, Vol.12 No.14, pp. 1970- 1979, 2021.

## References

- [1] Jiménez Gaona Y, Rodríguez-Álvarez MJ, Freire J, Castillo D, Lakshminarayanan V, editors. Preprocessing fast filters and mass segmentation for mammography images. *Optical Engineering + Applications*; 2021.
- [2] Tang J, Rangayyan RM, Xu J, El Naqa I, Yang Y. Computer-aided detection and diagnosis of breast cancer with mammography: recent advances. *IEEE Trans Inf Technol Biomed.* 2009;13(2):236-51.
- [3] Laal M. Innovation process in medical imaging. *Procedia-Social and Behavioral Sciences.* 2013 Jun 28;81:60-4.
- [4] Ali MT, Lahmood FM, Chisab RF. Mammogram image segmentation for improving the diagnosis of dense breast issues. *International Journal of Engineering and Technology.* 2019;8(1):44-52.
- [5] Paruchuri S, Sim S, Ganesh P, Thamby SA, Ping NY. Knowledge Assessment on Breast Cancer and Breast Self-Examination Practice among Female University Students in Kedah, Malaysia. *Systematic Reviews in Pharmacy.* 2021;12(9):3044-52.
- [6] Fatima M, Pasha M. Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Learning Systems and Applications.* 2017;9(01):1-16.
- [7] Taheri M, Hamer G, Son SH, Shin SY. Enhanced breast cancer classification with automatic thresholding using SVM and Harris corner detection. In *Proceedings of the International Conference on Research in Adaptive and Convergent Systems 2016 Oct 11* (pp. 56-60).
- [8] Varela C, Tahoces PG, Mendez AJ, Souto M, Vidal JJ. Computerized detection of breast masses in digitized mammograms. *Computers in Biology and Medicine.* 2007 Feb 1;37(2):214-26.
- [9] Patel BC, Sinha GR. Mammography feature analysis and mass detection in breast cancer images. In *2014 International Conference on Electronic Systems, Signal Processing and Computing Technologies 2014 Jan 9* (pp. 474-478). IEEE.
- [10] Jaleel JA, Salim S, Archana S. Textural features based computer aided diagnostic system for mammogram mass classification. In *2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT) 2014 Jul 10* (pp. 806-811). IEEE.

- [11] Lata KK, Kumar BM, Pritee K. Breast cancer detection in digital mammograms. InIEEE international conference on imaging systems and techniques (IST) 2015 (p. 16e8).
- [12] Setiawan AS, Wesley J, Purnama Y. Mammogram classification using law's texture energy measure and neural networks. *Procedia Computer Science*. 2015 Jan 1;59:92-7.
- [13] Rouhi R, Jafari M, Kasaei S, Keshavarzian P. Benign and malignant breast tumors classification based on region growing and CNN segmentation. *Expert Systems with Applications*. 2015 Feb 15;42(3):990-1002.
- [14] Xie W, Li Y, Ma Y. Breast mass classification in digital mammography based on extreme learning machine. *Neurocomputing*. 2016 Jan 15;173:930-41.
- [15] Last F, Douzas G, Bacao F. Oversampling for imbalanced learning based on k-means and smote. *arXiv preprint arXiv:1711.00837*. 2017 Nov 2.
- [16] Punitha S, Amuthan A, Joseph KS. Benign and malignant breast cancer segmentation using optimized region growing technique. *Future Computing and Informatics Journal*. 2018 Dec 1;3(2):348-58.
- [17] Shen L, Margolies LR, Rothstein JH, Fluder E, McBride R, Sieh W. Deep learning to improve breast cancer detection on screening mammography. *Scientific reports*. 2019 Aug 29;9(1):1-12.
- [18] Kaur P, Singh G, Kaur P. Intellectual detection and validation of automated mammogram breast cancer images by multi-class SVM using deep learning classification. *Informatics in Medicine Unlocked*. 2019 Jan 1;16:100151.
- [19] Mabrouk MS, Afify HM, Marzouk SY. Fully automated computer-aided diagnosis system for micro calcifications cancer based on improved mammographic image techniques. *Ain Shams Engineering Journal*. 2019 Sep 1;10(3):517-27.
- [20] Viswanath VH, Guachi-Guachi L, Thirumuruganandham SP. Breast cancer detection using image processing techniques and classification algorithms.2019.
- [21] Azary H, Abdoos M. A semi-supervised method for tumor segmentation in mammogram images. *Journal of medical signals and sensors*. 2020 Jan;10(1):12.
- [22] Kamil MY, Jassam AL. Analysis of tissue abnormality in mammography images using gray level co-occurrence matrix method. In*Journal of Physics: Conference Series* 2020 May 1 (Vol. 1530, No. 1, p. 012101). IOP Publishing.
- [23] Khoulqi I, Idrissi N, Sarfraz M. Segmentation of pectoral muscle in mammogram images using k-means and region growing. *Information Sciences Letters*. 2021;10(1):7.

- [24] Elsadig MA. A machine learning approach for breast cancer early detection. *J Theor Appl Inf Technol*. 2021 Mar 15;99(5).
- [25] Almalki YE, Soomro TA, Irfan M, Alduraibi SK, Ali A. Impact of Image Enhancement Module for Analysis of Mammogram Images for Diagnostics of Breast Cancer. *Sensors*. 2022 Feb 26;22(5): p.1868.
- [26] Hikmah NF, Sardjono TA, Mertiana WD, Firdi NP, Purwitasari D. An Image Processing Framework for Breast Cancer Detection Using Multi-View Mammographic Images. *EMITTER International Journal of Engineering Technology*. 2022 Jun 20:136-52.
- [27] Singh L, Alam A. An efficient hybrid methodology for an early detection of breast cancer in digital mammograms. *Journal of Ambient Intelligence and Humanized Computing*. 2022 May 28:1-24.
- [28] Sakib S, Yasmin N, Tanzeem AK, Shorna F, Alam SB. Breast Cancer Detection and Classification: A Comparative Analysis Using Machine Learning Algorithms. In *Proceedings of Third International Conference on Communication, Computing and Electronics Systems 2022* (pp. 703-717). Springer, Singapore.
- [29] Dafni Rose J, VijayaKumar K, Singh L, Sharma SK. Computer-aided diagnosis for breast cancer detection and classification using optimal region growing segmentation with MobileNet model. *Concurrent Engineering*. 2022 Apr 14:1063293X221080518.
- [30] Sarvestani ZM, Jamali J, Taghizadeh M, Dindarloo MH. A novel machine learning approach on texture analysis for automatic breast microcalcification diagnosis classification of mammogram images. 2022.
- [31] Almalki YE, Soomro TA, Irfan M, Alduraibi SK, Ali A. Computerized Analysis of Mammogram Images for Early Detection of Breast Cancer. In *Healthcare 2022* Apr 25 (Vol. 10, No. 5, p. 801). MDPI.
- [32] Malek AA, Rahman WE, Haris MH, Jalil UM. Segmenting masses in ultrasound images by using seed based region growing and mathematical morphology. *Advanced Science Letters*. 2017 Nov 1;23(11):11512-6.
- [33] Embong R, Aziz NN, Abd Karim AH, Ibrahim MR. Colour application on mammography image segmentation. In *Journal of Physics: Conference Series* 2017 Sep 1 (Vol. 890, No. 1, p. 012066). IOP Publishing.
- [34] Michael E, Ma H, Li H, Kulwa F, Li J. Breast cancer segmentation methods: current status and future potentials. *BioMed Research International*. 2021 Jul 22.

- [35] Zhou K, Li W, Zhao D. Deep learning-based breast region extraction of mammographic images combining pre-processing methods and semantic segmentation supported by Deeplab v3+. *Technology and Health Care*. 2022 Jan 1(Preprint):1-8.
- [36] Zhang YN, Xia KR, Li CY, Wei BL, Zhang B. Review of breast cancer pathological image processing. *BioMed Research International*. 2021 Sep 21.
- [37] García E, Diez Y, Diaz O, Lladó X, Martí R, Martí J, Oliver A. A step-by-step review on patient-specific biomechanical finite element models for breast MRI to x-ray mammography registration. *Medical physics*. 2018 Jan;45(1):e6-31.
- [38] Soleimani H. *Information Fusion of Magnetic Resonance Images and Mammographic Scans for Improved Diagnostic Management of Breast Cancer*. 2017.
- [39] Gardezi SJ, Elazab A, Lei B, Wang T. Breast cancer detection and diagnosis using mammographic data: Systematic review. *Journal of medical Internet research*. 2019 Jul 26;21(7):e14464.
- [40] Adusei EV. *Circulating Cell-Free DNA as a Blood Biomarker in Monitoring Response to Chemotherapy in Breast Cancer (Doctoral dissertation, University Of Ghana)*.2020.
- [41] Houfani D, Slatnia S, Kazar O, Zerhouni N, Merizig A, Saouli H. Machine learning techniques for breast cancer diagnosis: literature review. In *International Conference on Advanced Intelligent Systems for Sustainable Development 2019 Jul 8 (pp. 247-254)*. Springer, Cham.
- [42] Alshanbari H. *Automatic Breast Cancer Classification Using Novel Feature Extraction for Magnetic Resonance Imaging and Image Processing Technique: Coventry University*; 2013.
- [43] Mughal B. *Early Detection and Classification of Breast Tumor From Mammography: COMSATS Institute of Information Technology, Islamabad*; 2019.
- [44] Ionescu GV. *Automated prediction and early detection of breast cancer in mammograms (Doctoral dissertation, University of Manchester)*. 2020.
- [45] Jiang L. *Association between use of a specialized diagnostic assessment unit and the diagnostic interval in Ontario breast cancer patients*. Queen's University (Canada); 2014.
- [46] *Study Of Breast Cancer Treatment Reveals Paradox Of Precision Medicine [Internet]. Breast Cancer | Breast Cancer Symptoms | Breast Cancer Stages | Breast Cancer Sign and Types. [cited 2022 May 10]. Available from: <https://breast-cancer-sign.blogspot.com/2016/08/study-of-breast-cancer-treatment.html>*

- [47] Petridis C. *Molecular Genetics of Lobular Breast Cancer Ductal Carcinoma in Situ*: King's College London; 2018.
- [48] Zhang Y. *Automated Analysis of Mammograms using Evolutionary Algorithms*: University of York; 2012.
- [49] Mwadulo MW. *ALOCAL DIRECTIONAL TERNARY PATTERN TEXTURE DESCRIPTOR FOR MAMMOGRAPHIC BREAST CANCER CLASSIFICATION*: MMUST; 2020.
- [50] Agarwal R. *Computer aided detection for breast lesion in ultrasound and mammography*. 2019.
- [51] ANGAYARKANNI N. *ANALYSIS AND DETECTION OF BREAST CANCER IN MAMMOGRAM BY IMAGE PROCESSING TECHNIQUES*. 2016.
- [52] Oza P, Sharma P, Patel S, Bruno A. A bottom-up review of image analysis methods for suspicious region detection in mammograms. *Journal of Imaging*. 2021 Sep 18;7(9):190.
- [53] Akram Z. *Detection and Classification of Mammographic Abnormalities*: Aberystwyth University (Computer Science); 2019.
- [54] Gaur S, Dialani V, Slanetz PJ, Eisenberg RL. Architectural distortion of the breast. *American Journal of Roentgenology*. 2013 Nov;201(5):p, 662-70.
- [55] Worku B. *Breast Cancer Classification Using Image Processing Technique and Support Vector Machine*: St. Mary's University; 2017.
- [56] Huang N, Chen L, He J, Nguyen QD. The Efficacy of Clinical Breast Exams and Breast Self-Exams in Detecting Malignancy or Positive Ultrasound Findings. *Cureus*. 2022 Feb 21;14(2).
- [57] Abrahamsson L. *Statistical models of breast cancer tumour progression for mammography screening data*: Karolinska Institutet (Sweden); 2018.
- [58] Puig Vives M. *Breast cancer epidemiology: mammographic screening and molecular subtypes*: Universitat de Girona; 2015.
- [59] Henrysson R. *Evaluation of Quantitative PET/CT Usage for Cancer Treatment*. 2016.
- [60] dos Santos Teixeira RF. *Automatic analysis of mammography images: classification of breast density*. 2013.
- [61] Wu H. *Automatic computer aided diagnosis of breast cancer in dynamic contrast enhanced magnetic resonance images*: University of Toronto (Canada); 2016.

- [62] Breast Cancer, Breast cancer care, Breast Care, Cosmetic Surgery [Internet]. Ganga Breast Care Centre. [cited 2022 Jul 20]. Available from: <https://gangabreastcare.com/mammography.php>
- [63] Ragab DA, Sharkas M, Attallah O. Breast cancer diagnosis using an efficient CAD system based on multiple classifiers. *Diagnostics*. 2019 Oct 26;9(4):165.
- [64] Pereira DC, Ramos RP, Do Nascimento MZ. Segmentation and detection of breast cancer in mammograms combining wavelet analysis and genetic algorithm. *Computer methods and programs in biomedicine*. 2014 Apr 1;114(1):88-101.
- [65] Mughal B, Muhammad N, Sharif M, Rehman A, Saba T. Removal of pectoral muscle based on topographic map and shape-shifting silhouette. *BMC cancer*. 2018;18(1):1-14.
- [66] Sreedevi S, Sherly E. A novel approach for removal of pectoral muscles in digital mammogram. *Procedia Computer Science*. 2015 Jan 1;46:1724-31.
- [67] Imaging of the Breast | Concise Medical Knowledge [Internet]. [www.lecturio.com](http://www.lecturio.com). [cited 2022 Feb 16]. Available from: <https://www.lecturio.com/concepts/breast-imaging/>
- [68] Iranmakani S, Mortezaazadeh T, Sajadian F, Ghaziani MF, Ghafari A, Khezerloo D, Musa AE. A review of various modalities in breast imaging: technical aspects and clinical outcomes. *Egyptian Journal of Radiology and Nuclear Medicine*. 2020 Dec;51(1):1-22.
- [69] Xing D, Lv Y, Sun B, Xie H, Dong J, Hao C, Chen Q, Chi X. Diagnostic value of contrast-enhanced spectral mammography in comparison to magnetic resonance imaging in breast lesions. *Journal of Computer Assisted Tomography*. 2019 Mar;43(2):245.
- [70] Anwar R, Farouk MA, Abdel Hamid WR, Abu El Maati AA, Eissa H. Breast cancer in dense breasts: comparative diagnostic merits of contrast-enhanced mammography and diffusion-weighted breast MRI. *Egyptian Journal of Radiology and Nuclear Medicine*. 2021 Dec;52(1):1-3.
- [71] Mahmoud MAM. Development of advanced computer methods for breast cancer image interpretation through texture and temporal evolution analysis: Universitat Rovira i Virgili; 2016.
- [72] Hossam A, Harb HM, Abd El Kader HM. Performance analysis of breast cancer imaging techniques. *Int. J. Comput. Sci. Inf. Secur*. 2017 May.
- [73] Shubbar S. Ultrasound medical imaging systems using telemedicine and blockchain for remote monitoring of responses to neoadjuvant chemotherapy in women's breast cancer: concept and implementation: Kent State University; 2017.

- [74] Bhushan A, Gonsalves A, Menon JU. Current state of breast cancer diagnosis, treatment, and theranostics. *Pharmaceutics*. 2021 May 14;13(5):723.
- [75] Yurttakal AH, Erbay H, İkizceli T, Karaçavuş S. Detection of breast cancer via deep convolution neural networks using MRI images. *Multimedia Tools and Applications*. 2020 Jun;79(21):15555-73.
- [76] García-Barquín P, Páramo M, Elizalde A, Pina L, Etxano J, Fernandez-Montero A, Caballeros M. The effect of the amount of peritumoral adipose tissue in the detection of additional tumors with digital breast tomosynthesis and ultrasound. *Acta Radiologica*. 2017 Jun;58(6):645-51.
- [77] Abbas S. *Digital Image Processing using Soft Computing Techniques and Spline Representations*: University of the Punjab, Lahore, Pakistan; 2017.
- [78] Gonzalez RC. *Digital image processing*. Third edit. London, 2008.
- [79] Umbaugh SE. *Digital image processing and analysis: human and computer vision applications with CVIPtools*. CRC press; 2010 Nov 19.
- [80] Abdulla AA. Efficient computer-aided diagnosis technique for leukaemia cancer detection. *IET Image Processing*. 2020 Dec;14(17):4435-40.
- [81] Ahmad MA. *Mining health data for breast cancer diagnosis using machine learning*. Canberra, Australia: University of Canberra; 2013 Dec.
- [82] Gonzalez RC. *Digital image processing*. Fourth edit. London, 2018.
- [83] Dewangan SK. Importance & Applications of Digital Image Processing. *International Journal of Computer Science & Engineering Technology (IJCSET)*. 2016 Jul;7(7):316-20.
- [84] Zanova MA. Image Processing & Neural Network Based Breast Cancer Detection. *Comput. Inf. Sci.*. 2019;12(2):146.
- [85] Wang X, Liang G, Zhang Y, Blanton H, Bessinger Z, Jacobs N. Inconsistent performance of deep learning models on mammogram classification. *Journal of the American College of Radiology*. 2020 Jun 1;17(6):796-803.
- [86] Khairi SS, Bakar MA, Alias MA, Bakar SA, Liong CY, Rosli N, Farid M. Deep learning on histopathology images for breast cancer classification: a bibliometric analysis. *InHealthcare* 2021 Dec 22 (Vol. 10, No. 1, p. 10). MDPI.
- [87] Oliveira JE, Gueld MO, Araújo AD, Ott B, Deserno TM. Toward a standard reference database for computer-aided mammography. In *Medical imaging 2008: Computer-aided diagnosis 2008* Mar 27 (Vol. 6915, pp. 606-614). SPIE.



- [88] Sadad T. Tumor Detection, Classification and Risk Assessment in Digital Mammograms: International Islamic University, Islamabad; 2019.
- [89] Zebari DA, Ibrahim DA, Zeebaree DQ, Haron H, Salih MS, Damaševičius R, Mohammed MA. Systematic review of computing approaches for breast cancer detection based computer aided diagnosis using mammogram images. *Applied Artificial Intelligence*. 2021 Dec 15;35(15):2157-203.
- [90] Vagssa P, Doudou NM, Jolivo T, Videme O, Kolyang DT. Pectoral muscle deletion on a mammogram to aid in the early diagnosis of breast cancer. *International Journal of Engineering, Science and Technology*. 2020 Sep 15;12(3):57-65.
- [91] Nandan D, Kanungo J, Mahajan A. An error-efficient Gaussian filter for image processing by using the expanded operand decomposition logarithm multiplication. *Journal of ambient intelligence and humanized computing*. 2018 Jul 14:1-8.
- [92] Boumaraf S, Liu X, Ferkous C, Ma X. A new computer-aided diagnosis system with modified genetic feature selection for bi-RADS classification of breast masses in mammograms. *BioMed Research International*. 2020 May 11.
- [93] Manohar KM, Patil AS. A review on techniques of image segmentation. *International Journal of Advanced Research in Computer and Communication Engineering*. 2016 Mar;5(3):201-9.
- [94] Quintanilla-Domínguez J, Ruiz-Pinales J, Barrón-Adame JM, Guzmán-Cabrera R. Microcalcifications detection using image processing. *Computación y Sistemas*. 2018 Mar;22(1):291-300.
- [95] Dougherty G. *Digital image processing for medical applications*. Cambridge University Press; 2009.
- [96] Senthilkumaran N, Vaithegi S. Image segmentation by using thresholding techniques for medical images. *Computer Science & Engineering: An International Journal*. 2016 Feb;6(1):1-3.
- [97] Tam C, Li S, Peters T. Machine Learning towards General Medical Image Segmentation. *Machine Learning*. 2020 Mar 9;10.
- [98] Ittannavar SS, Havaldar RH, Khot BA. A Research on Breast Cancer Detection using Mammography. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*. 2019. 9(1):3459-3463.

- [99] Muthukrishnan R, Radha M. Edge detection techniques for image segmentation. *International Journal of Computer Science & Information Technology*. 2011 Dec 1;3(6):259.
- [100] Jia W. Edge Detection Operators for X-ray Images Based on Hessian Matrices. 2020.
- [101] Thwe YM, Ogawa M, Dung PN. Applying Clustering Techniques for Refining Large Data Set: Case Study on Malware. In *2019 International Conference on Advanced Information Technologies (ICAIT) 2019 Nov 6* (pp. 238-243). IEEE.
- [102] Hossain M. AN IMPROVED CLUSTERING ALGORITHM FOR FINDING SPHERICAL AND NON SPHERICAL CLUSTERS ON DATA MINING 2019.
- [103] Abbas AH, Kareem AA, Kamil MY. Breast cancer image segmentation using morphological operations. *International journal of electronics and communication engineering and technology*. 2015 Apr;6(4):8-14.
- [104] Al-Khalidi FQ, Alkindy B, Abbas T. EXTRACT THE BREAST CANCER IN MAMMOGRAM IMAGES. *Technology*. 2019;10(02):96-105.
- [105] Shen S. MRI brain tumour classification using image processing and data mining. 2004.
- [106] Ippolito PP. Feature Extraction Techniques [Internet]. Medium. 2019. [cited 2022 Mar 5]. Available from: <https://towardsdatascience.com/feature-extraction-techniques-d619b56e31be>.
- [107] Haralick RM, Shanmugam K, Dinstein IH. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*. 1973 Nov(6):610-21.
- [108] Mateen M, Wen J, Song S, Huang Z. Fundus image classification using VGG-19 architecture with PCA and SVD. *Symmetry*. 2018 Dec 20;11(1):1.
- [109] Mitchell H. Image similarity measures. *Image Fusion: Springer*; 2010. p. 167-85.
- [110] Benesty J, Chen J, Huang Y, Cohen I. Pearson correlation coefficient. In *Noise reduction in speech processing 2009* (pp. 1-4). Springer, Berlin, Heidelberg
- [111] Liu T. Improved K-means clustering algorithms: a thesis presented in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Computer Science, Massey University, New Zealand: Massey University; 2020.
- [112] Giri P, Saravanakumar K. Breast cancer detection using image processing techniques. *Oriental journal of computer science and technology*. 2017 Jun 20;10(2):391-9.

- [113] Al-Waeli AM. An automated system for the classification and segmentation of brain tumours in MRI images based on the modified grey level co-occurrence matrix. University of Salford (United Kingdom); 2017.
- [114] Fatima M, Pasha M. Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Learning Systems and Applications*. 2017;9(01):1.
- [115] Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*. 2015 Jan 1;13:8-17.
- [116] Abdulqader DM, Abdulazeez AM, Zeebaree DQ. Machine learning supervised algorithms of gene selection: A review. *Machine Learning*. 2020 Apr;62(03):233-44.
- [117] Farooqui NA. A study on early prevention and detection of breast cancer using three-machine learning techniques. *Education*. 2017 Dec.
- [118] Gaikwad VJ. Detection of breast cancer in mammogram using support vector machine. *International Journal of Scientific Engineering and Research (IJSER)*. 2015 Feb;10(1):19-21.
- [119] Atrey K, Sharma Y, Bodhey NK, Singh BK. Breast cancer prediction using dominance-based feature filtering approach: A comparative investigation in machine learning archetype. *Brazilian Archives of Biology and Technology*. 2019 Nov 25;62.
- [120] Nguyen C, Wang Y, Nguyen HN. Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. *Journal of Biomedical Science and Engineering*. 2013 May 20;6(5):551-60.
- [121] Sirikulviriyaya N, Sinthupinyo S. Integration of rules from a random forest. In *International Conference on Information and Electronics Engineering* 2011 May 28 (Vol. 6, pp. 194-198).
- [122] Pavlov YL. Random forests. *Random Forests: De Gruyter*; 2019.
- [123] Chachalo Chachalo B, Solis E, Guachi R, Viswanath H, Guachi-Guachi L, Thirumuruganandham SP. BREAST CANCER DETECTION USING IMAGE PROCESSING TECHNIQUES AND CLASSIFICATION ALGORITHMS. *Journal of Natural Remedies*. 2020;21(2):37-47.
- [124] Maeda-Gutiérrez V, Galván-Tejada CE, Cruz M, Valladares-Salgado A, Galván-Tejada JI, Gamboa-Rosales H, García-Hernández A, Luna-García H, Gonzalez-Curiel I, Martínez-Acuña M. Distal symmetric polyneuropathy identification in type 2 diabetes

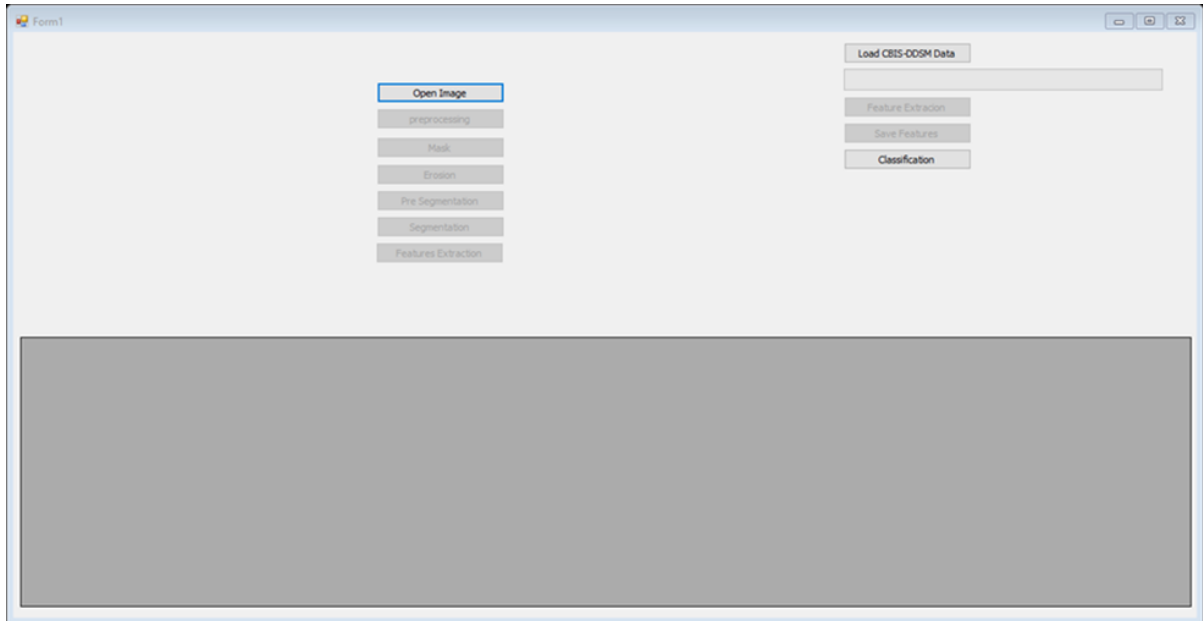
- subjects: A random forest approach. In *Healthcare* 2021 Feb 1 (Vol. 9, No. 2, p. 138). MDPI.
- [125] Random Forests Understanding [Internet]. ai-pool.com. [cited 2022 Jun 15]. Available from: <https://ai-pool.com/a/s/random-forests-understanding>
- [126] FOROOSHANI MK. Breast Tumor Detection by Mammogram Image Segmentation: POLITECNICO DI TORINO; 2021.
- [127] Ragab DA, Sharkas M, Marshall S, Ren J. Breast cancer detection using deep convolutional neural networks and support vector machines. *PeerJ*. 2019 Jan 28;7:e6201.
- [128] Xu J, Zhang Y, Miao D. Three-way confusion matrix for classification: A measure driven view. *Information sciences*. 2020 Jan 1;507:772-94.
- [129] Zhu Z. Detecting Coronavirus Disease 2019 Pneumonia in Chest X-Ray Images Using Deep Learning. University of California, Los Angeles; 2020.
- [130] Biratu ES, Schwenker F, Ayano YM, Debelee TG. A survey of brain tumor segmentation and classification algorithms. *Journal of Imaging*. 2021 Sep 6;7(9):179.
- [131] Ragab DA, Sharkas M, Attallah O. Breast cancer diagnosis using an efficient CAD system based on multiple classifiers. *Diagnostics*. 2019 Oct 26;9(4):165.
- [132] bin Alias MS, Ibrahim NB, Zin ZB. Improved sampling data Workflow using Smtmk to increase the classification accuracy of imbalanced dataset. *European Journal of Molecular & Clinical Medicine*. 2021 Jan 15;8(02).
- [133] Gu Q, Wang XM, Wu Z, Ning B, Xin CS. An improved SMOTE algorithm based on genetic algorithm for imbalanced data classification. *Journal of Digital Information Management*. 2016 Apr;14(2):92-103.
- [134] Picek S, Heuser A, Jovic A, Bhasin S, Regazzoni F. The curse of class imbalance and conflicting metrics with machine learning for side-channel evaluations. *IACR Transactions on Cryptographic Hardware and Embedded Systems*. 2019;2019(1):1-29.
- [135] Gao T. Hybrid classification approach for imbalanced datasets. *Graduate Theses and Dissertations*. 2015 Jan 1;14331.
- [136] Hussain L, Huang P, Nguyen T, Lone KJ, Ali A, Khan MS, Li H, Suh DY, Duong TQ. Machine learning classification of texture features of MRI breast tumor and peri-tumor of combined pre-and early treatment predicts pathologic complete response. *BioMedical Engineering OnLine*. 2021 Dec;20(1):1-23.

- [137] Tanha J, Abdi Y, Samadi N, Razzaghi N, Asadpour M. Boosting methods for multi-class imbalanced data classification: an experimental review. *Journal of Big Data*. 2020 Dec;7(1):1-47.
- [138] Maqsood S, Damaševičius R, Maskeliūnas R. TTCNN: A Breast Cancer Detection and Classification towards Computer-Aided Diagnosis Using Digital Mammography in Early Stages. *Applied Sciences*. 2022 Mar 23;12(7):3273.
- [139] CBIS-DDSM - The Cancer Imaging Archive (TCIA) Public Access - Cancer Imaging Archive Wiki [Internet]. [wiki.cancerimagingarchive.net](http://wiki.cancerimagingarchive.net). [cited 2021 May 6]. Available from: <https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM/>
- [140] Khaledyan D, Amirany A, Jafari K, Moayeri MH, Khuzani AZ, Mashhadi N, editors. Low-Cost Implementation of Bilinear and Bicubic Image Interpolation for Real-Time Image Super-Resolution. 2020 IEEE Global Humanitarian Technology Conference (GHTC); 2020: IEEE.
- [141] Yadav A, Jamir I, Jain RR, Sohani M. Comparative study of machine learning algorithms for breast cancer prediction-a review. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT)*, ISSN. 2019 Apr:2456-3307.
- [142] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*. 2002 Jun 1;16:321-57.

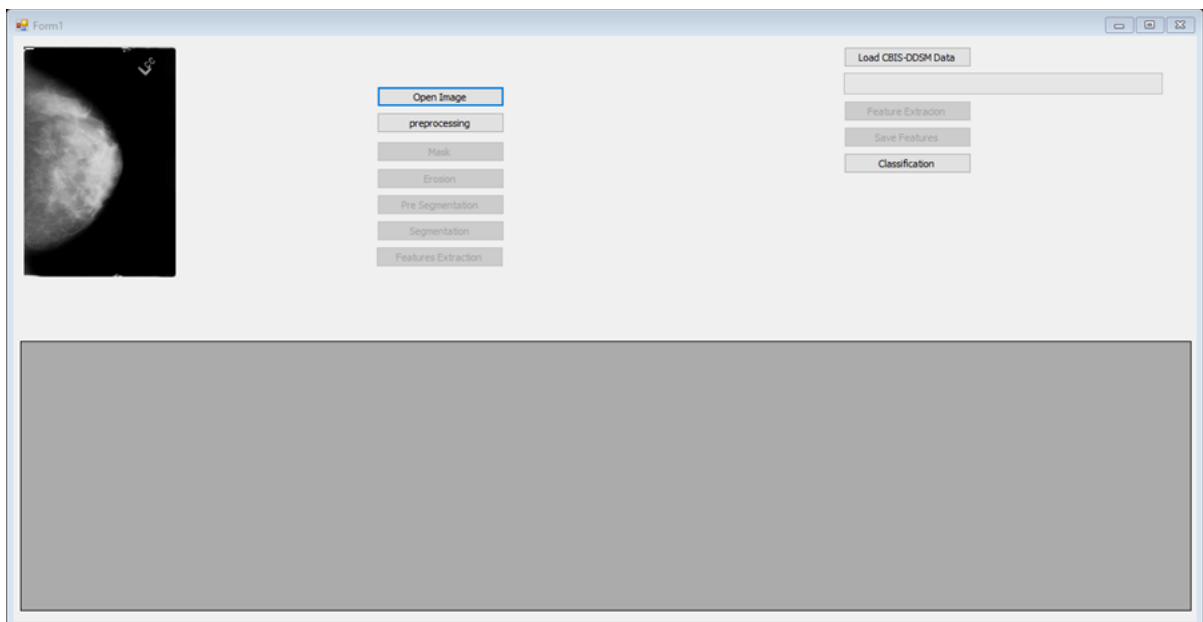
## Appendices

### Appendix A

In this appendix, the first model (Segmentation using k-means clustering and optimized region-growing technique) illustrated through the following figures:



**Figure A. 1** The main form of the first model



**Figure A. 2** Open a mammogram image

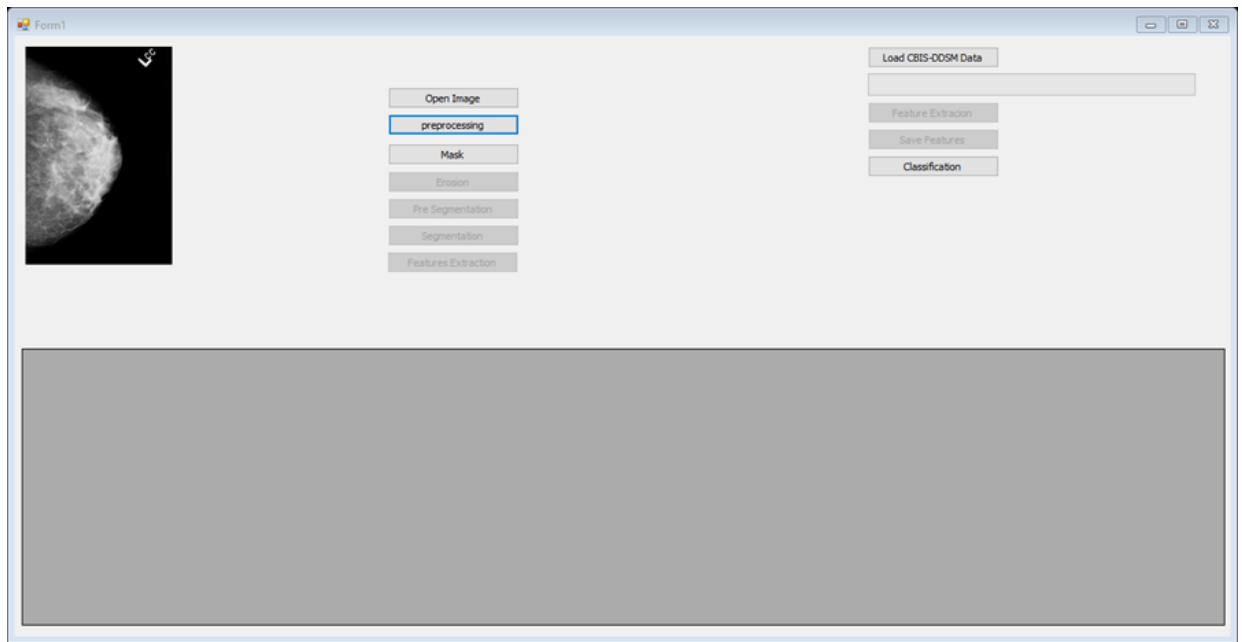


Figure A. 3 Pre-processing step

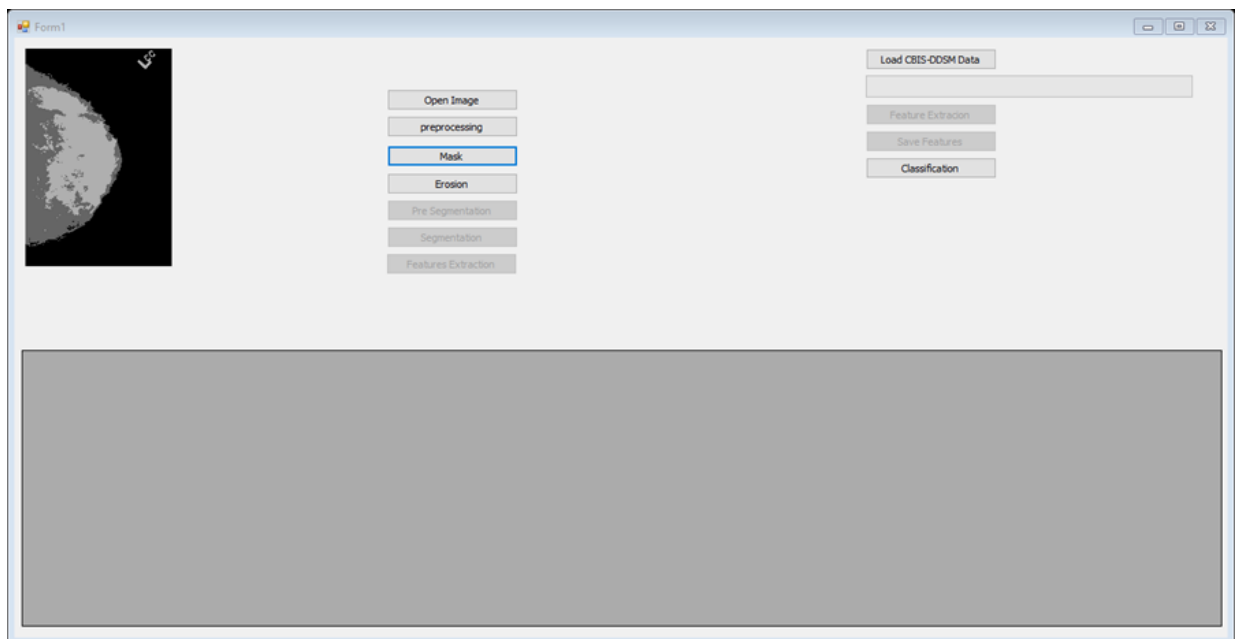
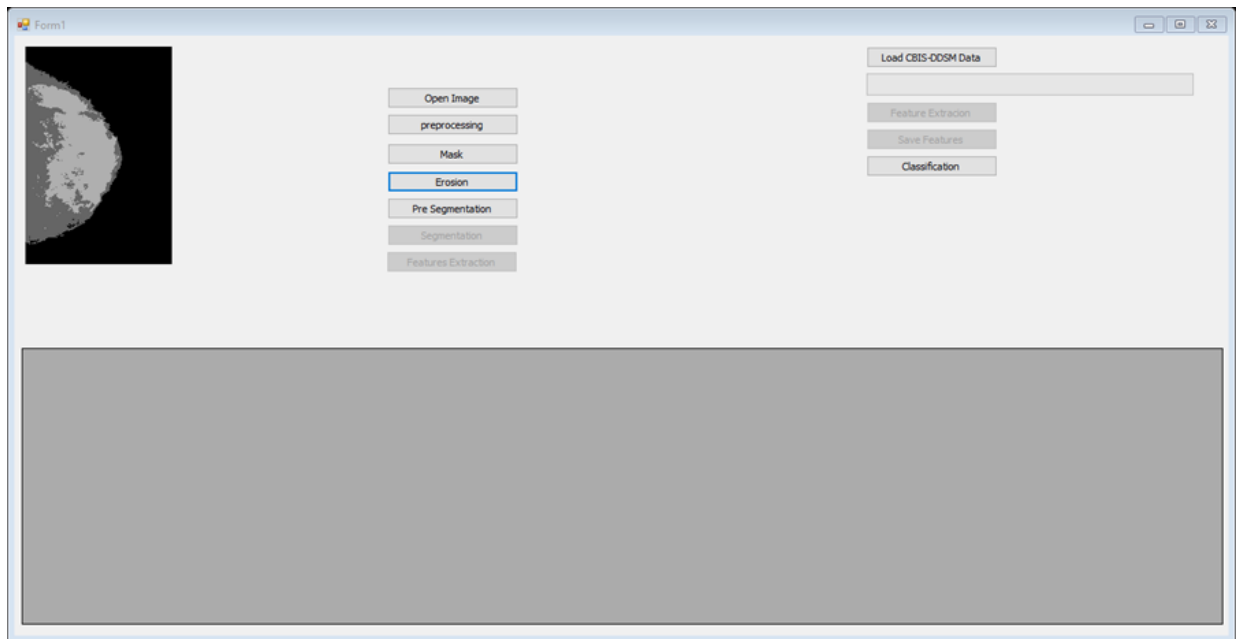
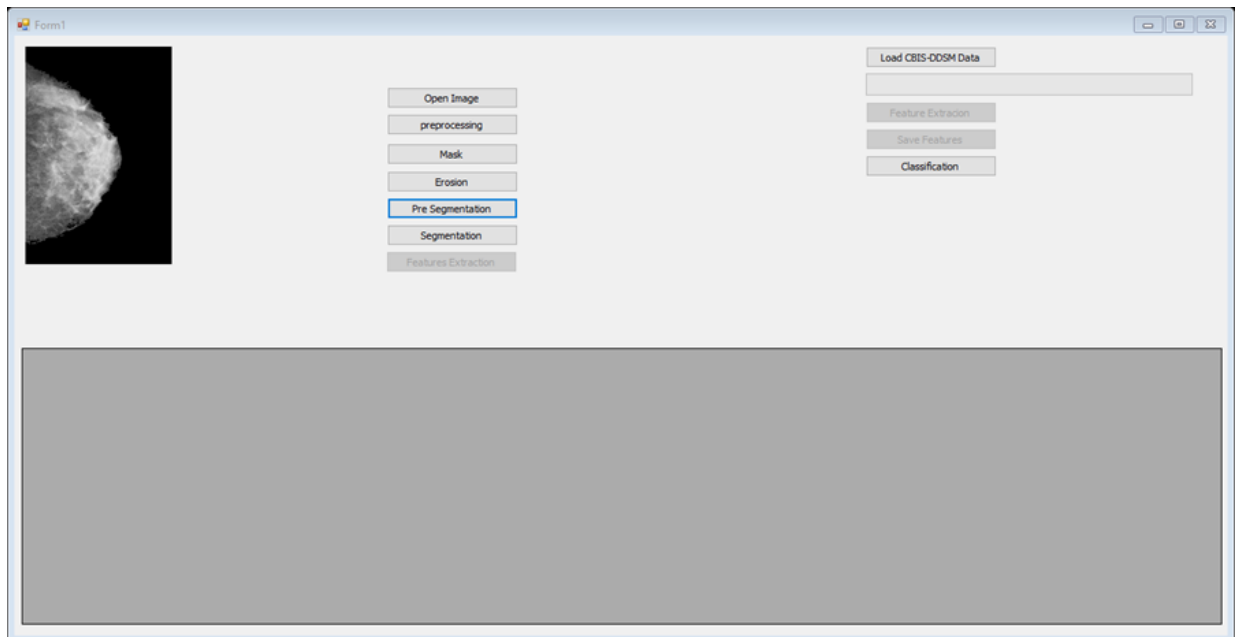


Figure A. 4 Create a mask for erosion step

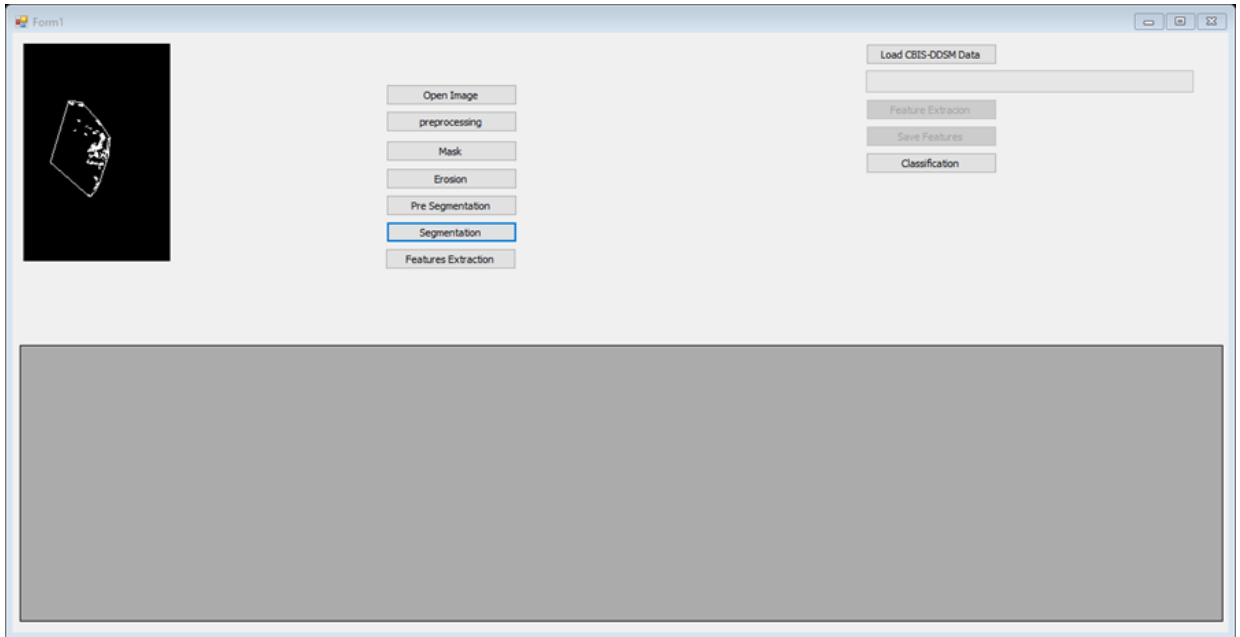


**Figure A. 5** Erosion step

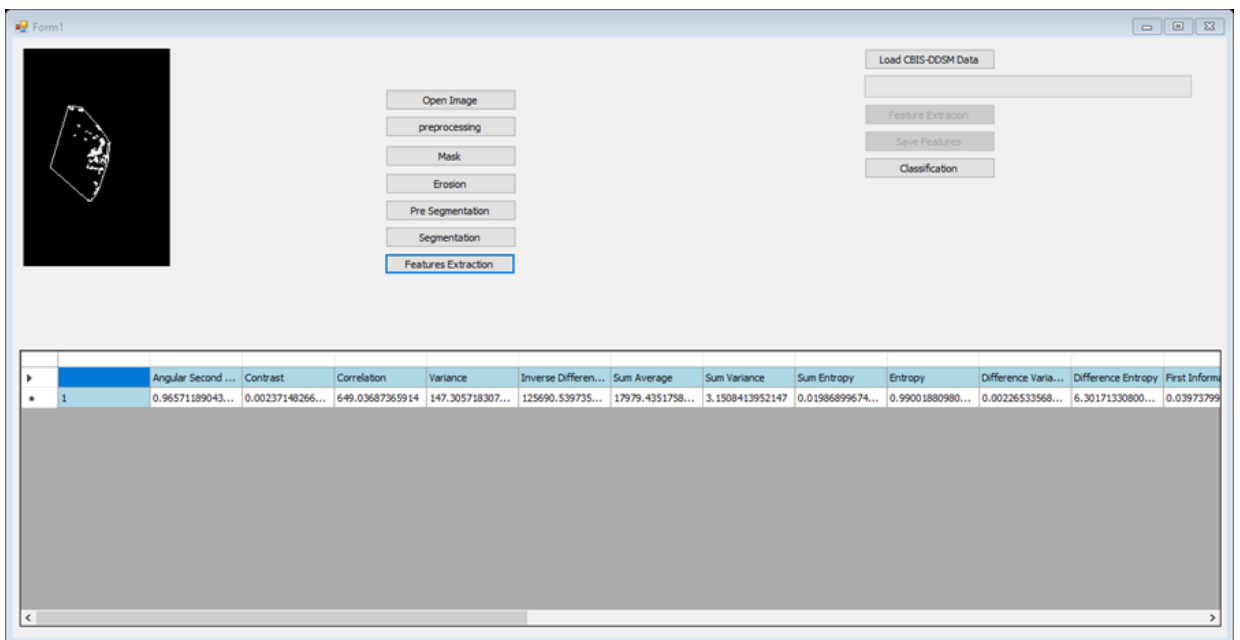


**Figure A. 6** Pre-segmentation step (breast area retrieval)

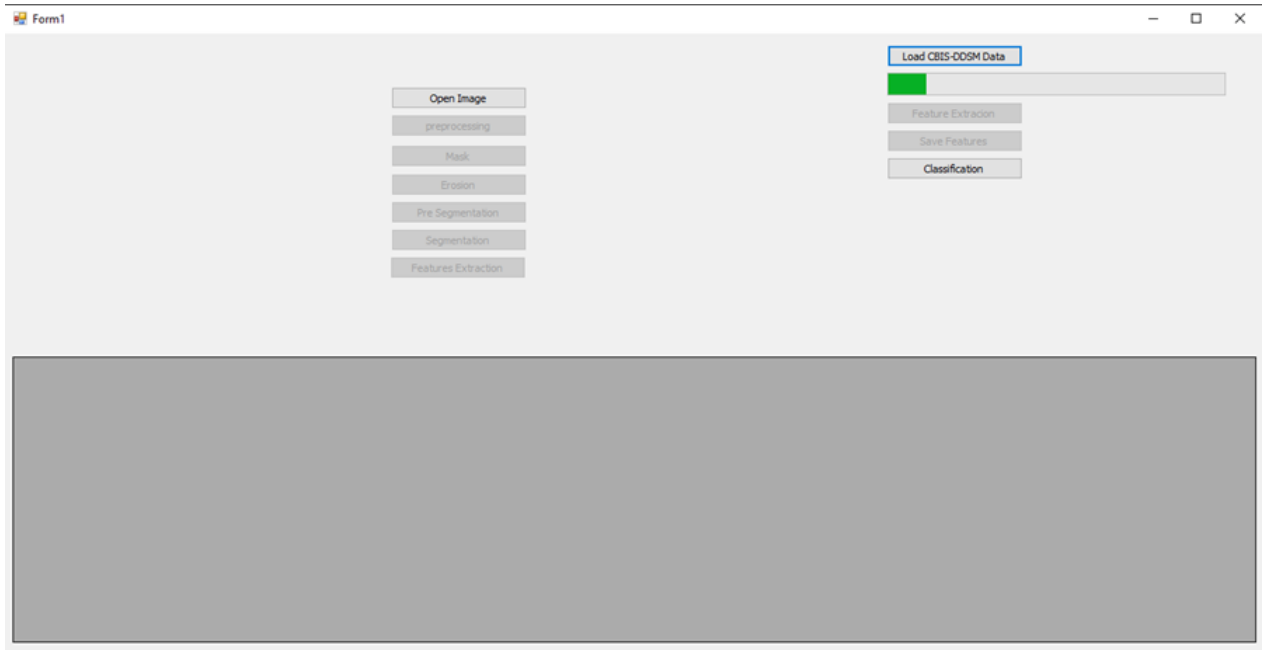




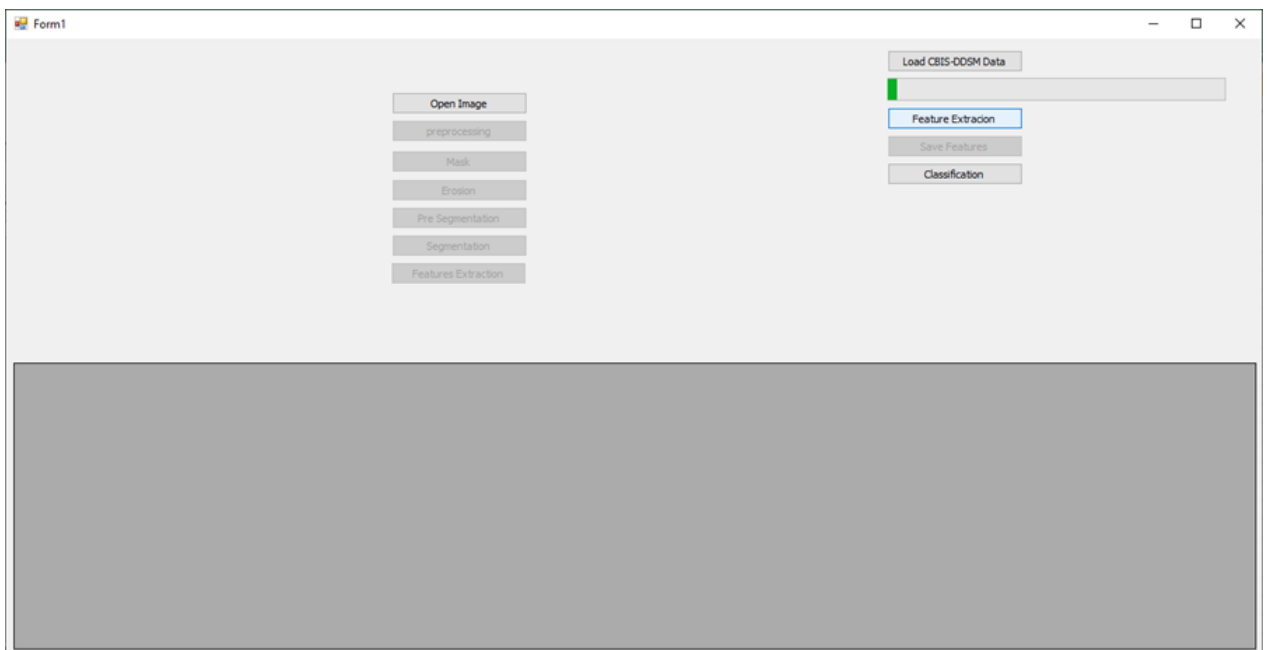
**Figure A. 7** Micro-calcification segmentation step



**Figure A. 8** Feature extraction step for one sample



**Figure A. 9** Dataset loading step



**Figure A. 10** Feature extraction step for the loaded dataset

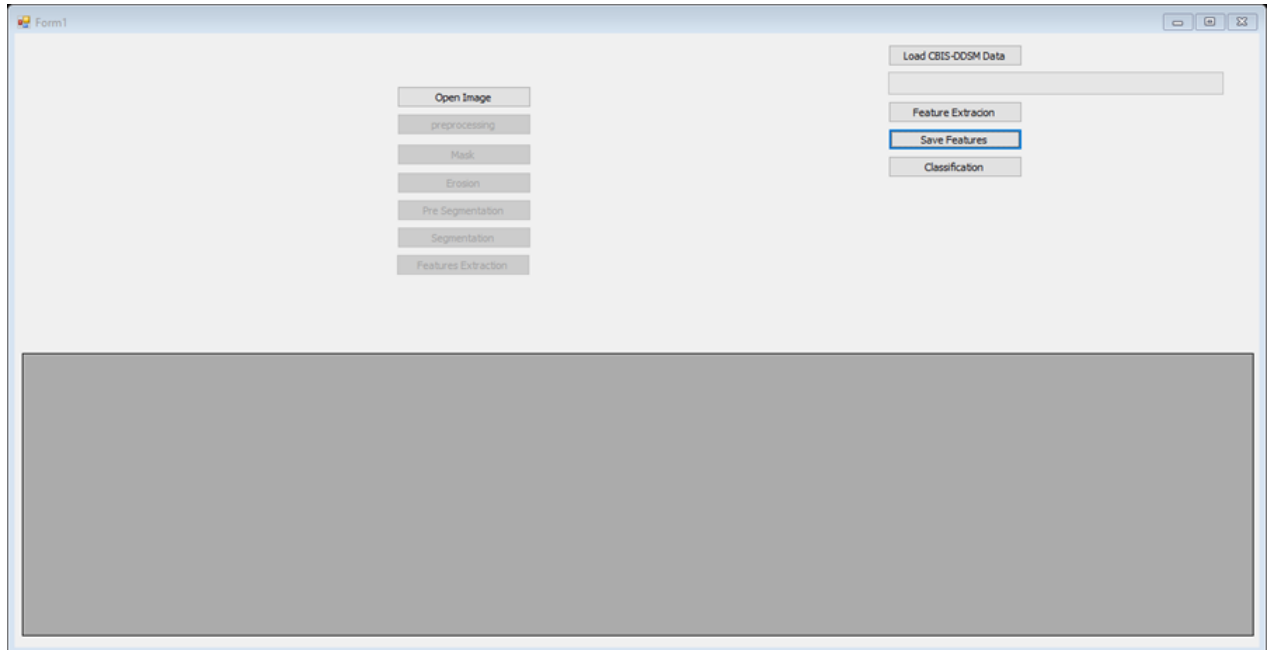


Figure A. 11 Saving the extracted features step

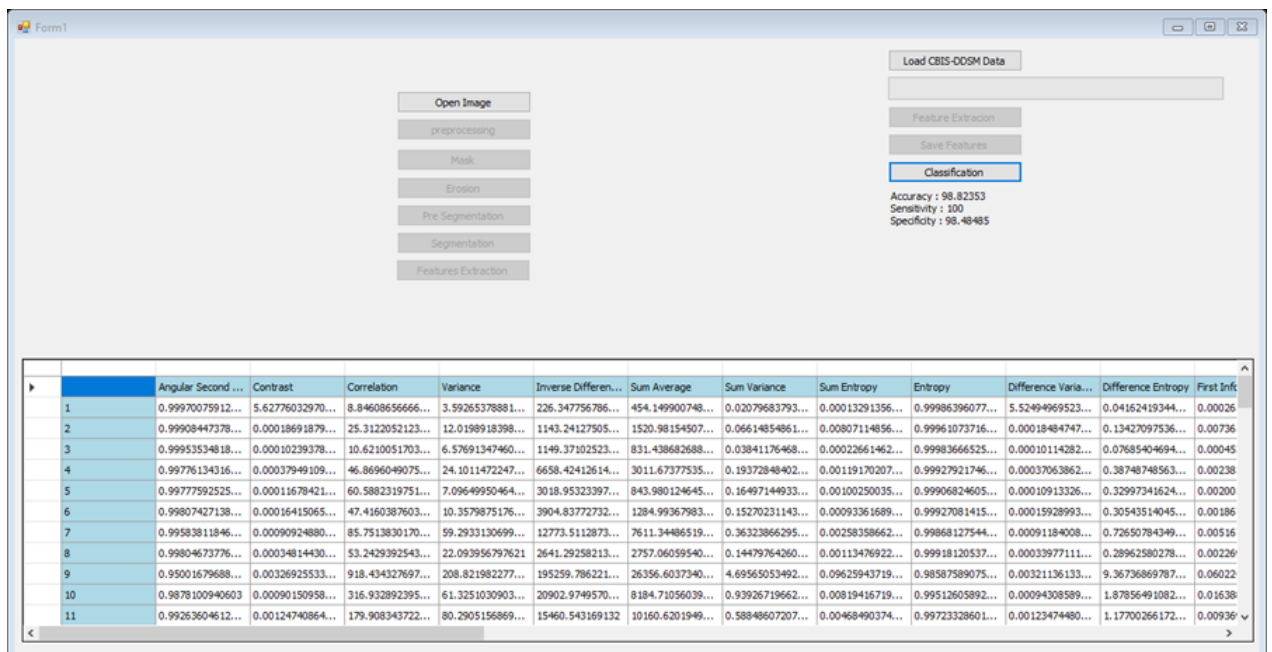
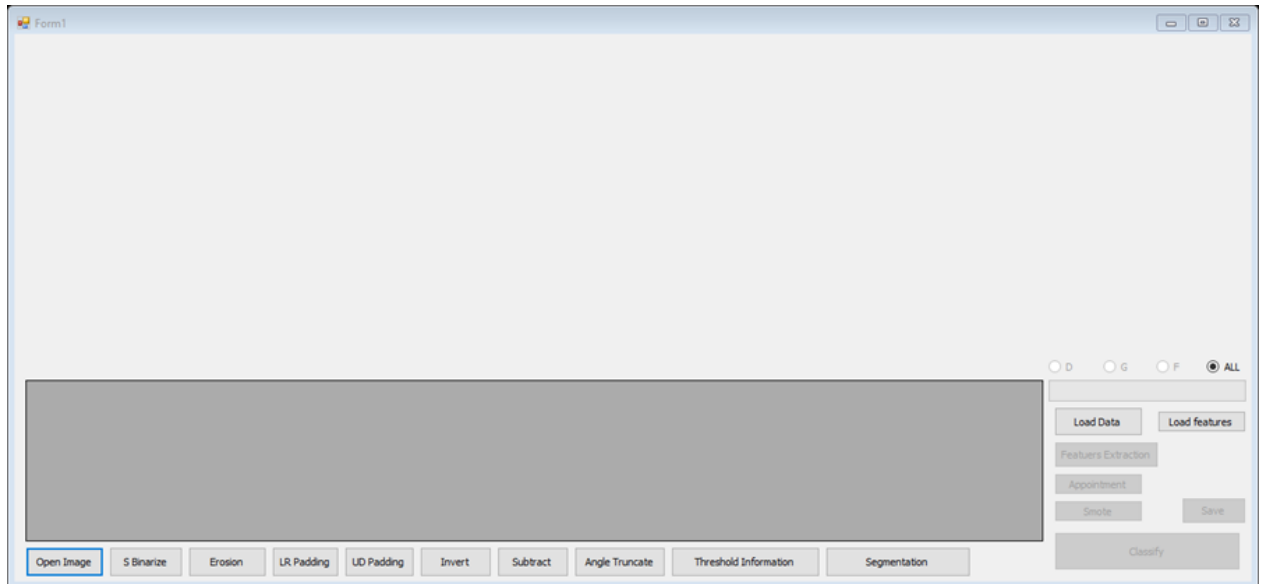


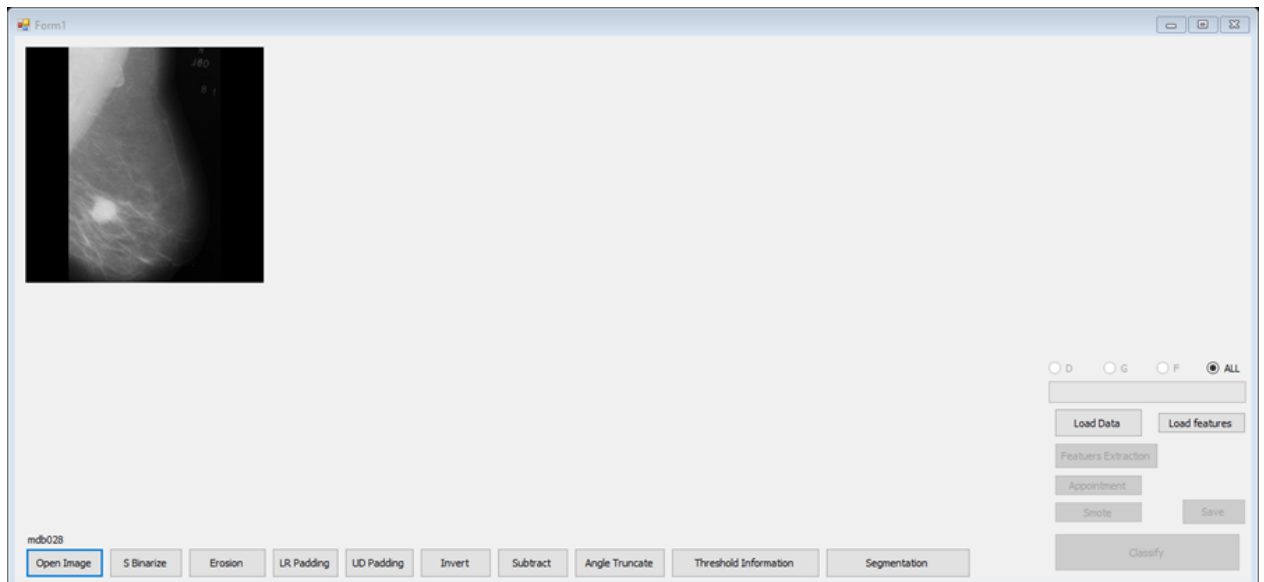
Figure A. 12 Classification step

## Appendix B

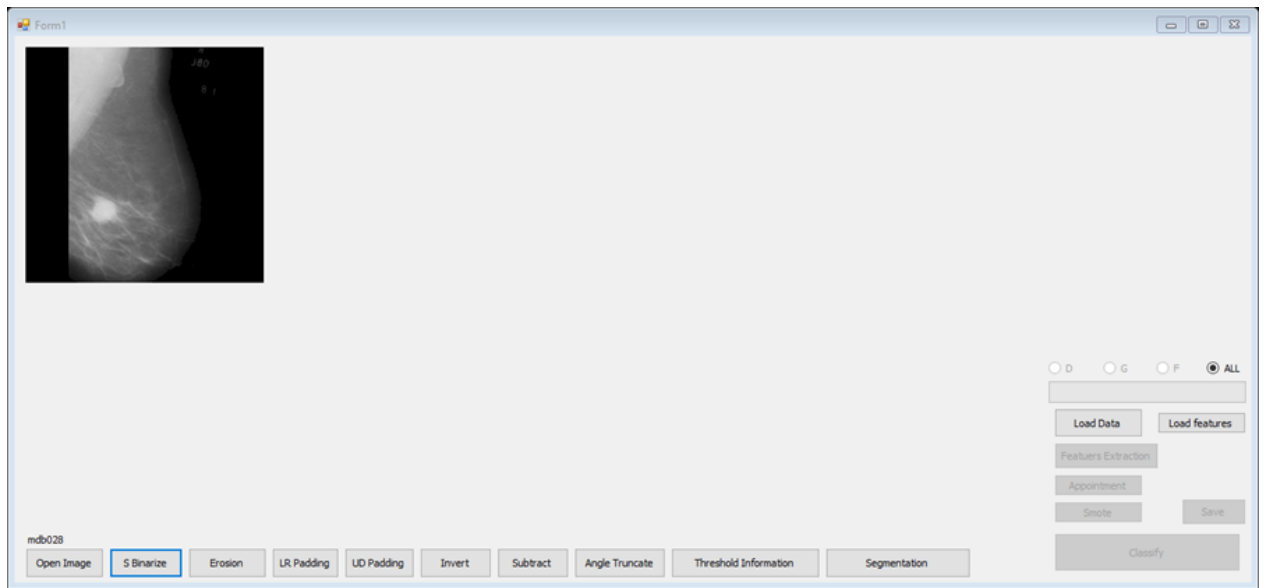
In this appendix, the second model (Pectoral Muscle Removal and Solving Data Imbalance Problem) illustrated through the following figures:



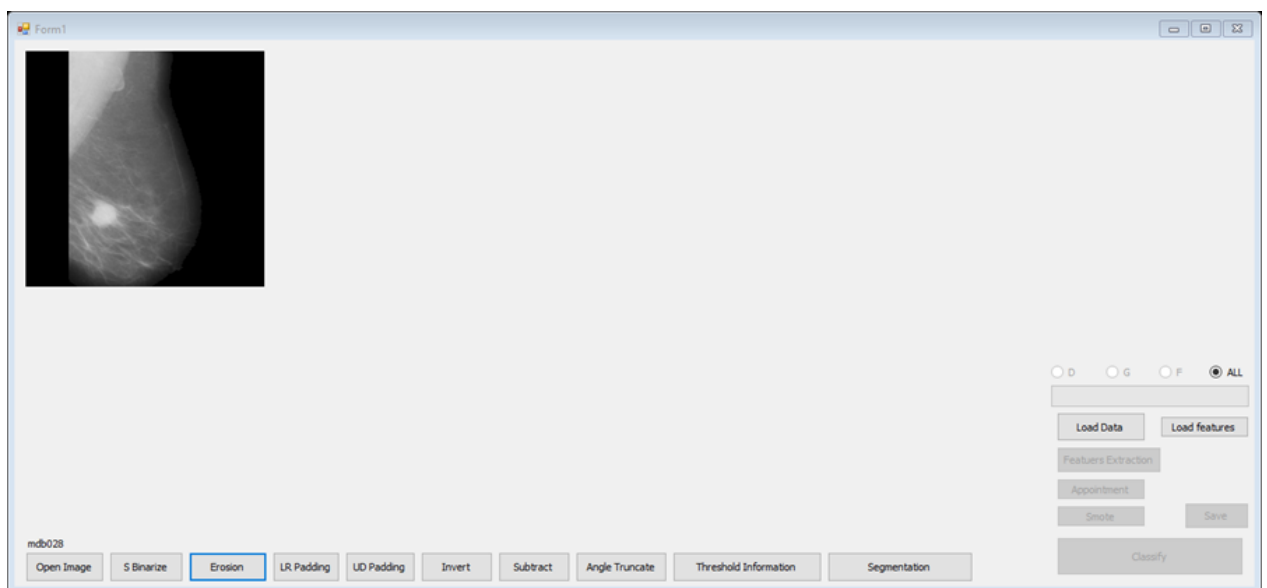
**Figure B. 1** The main form of the second model



**Figure B. 2** Open a mammogram image step for one sample



**Figure B. 3** Pre-processing (special thresholding) step



**Figure B. 4** Noise-removing step

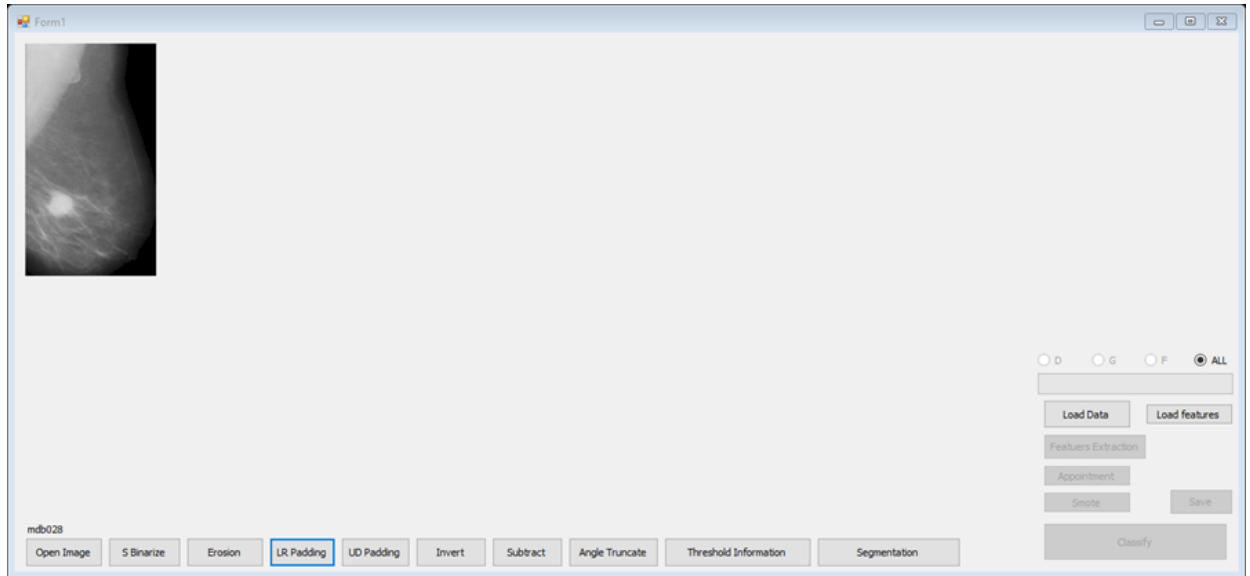


Figure B. 5 Left-Right image cropping step

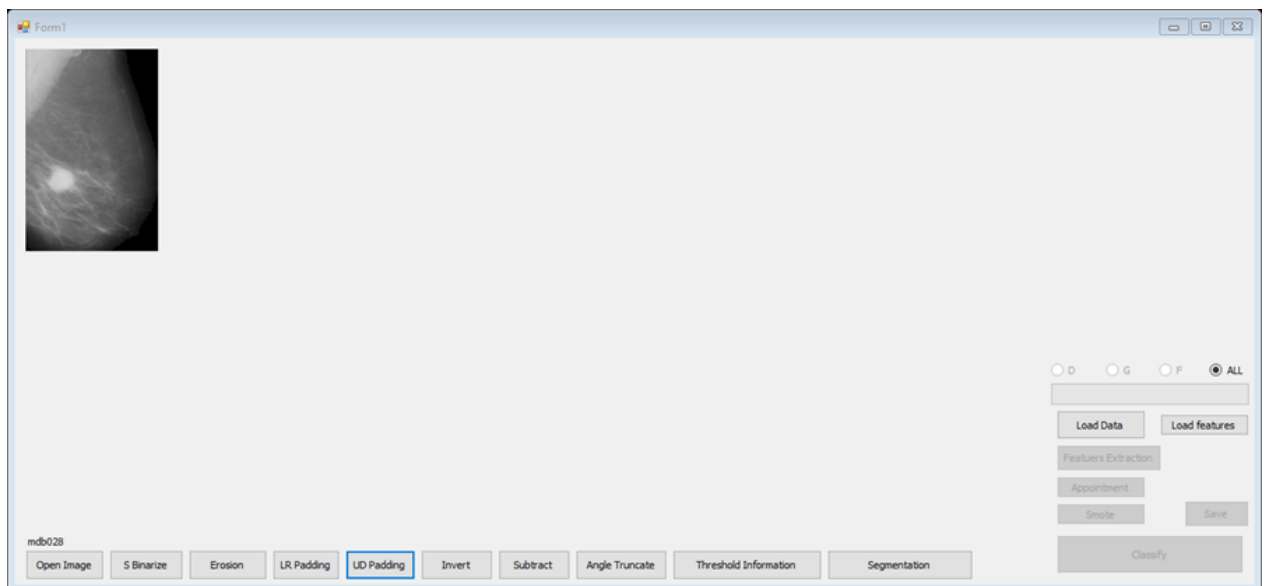
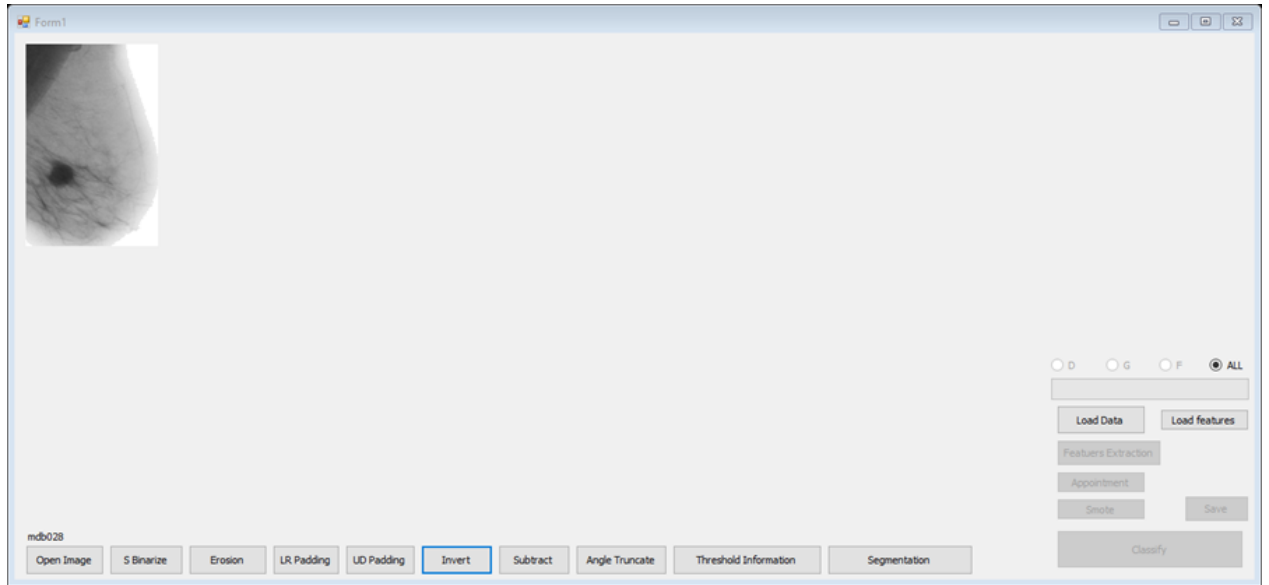
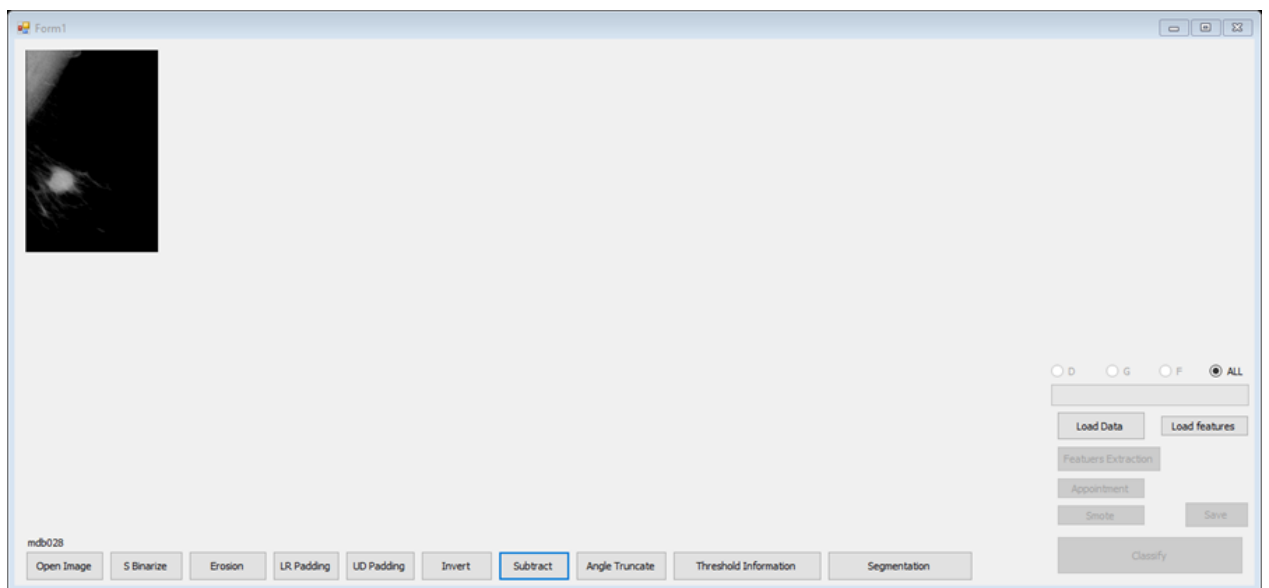


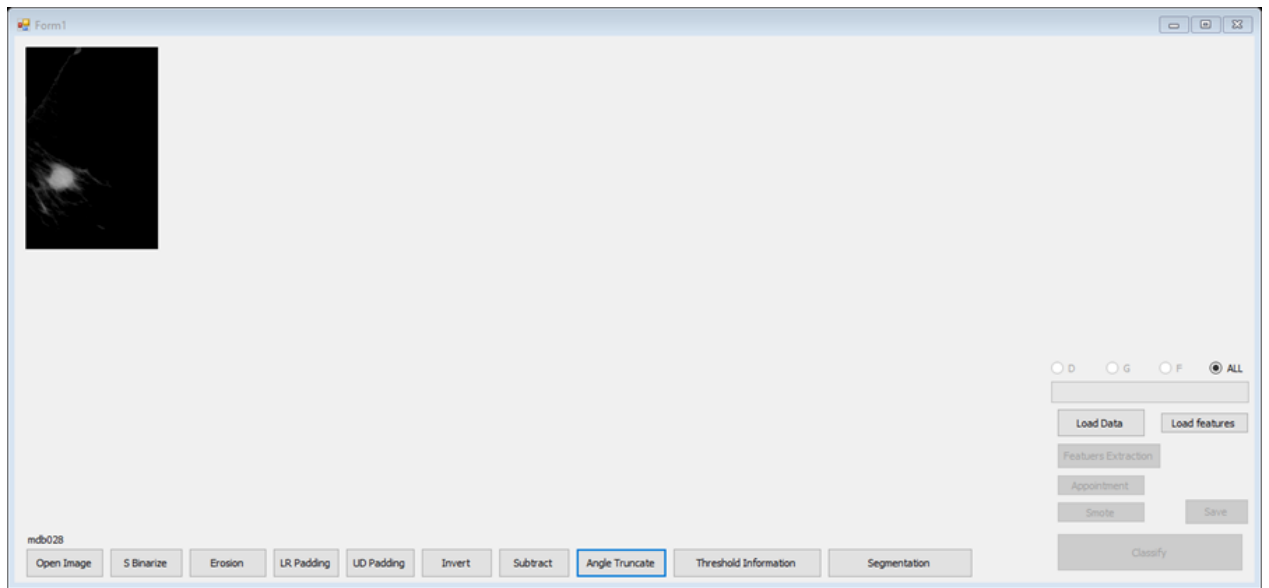
Figure B. 6 Up-Down image cropping step



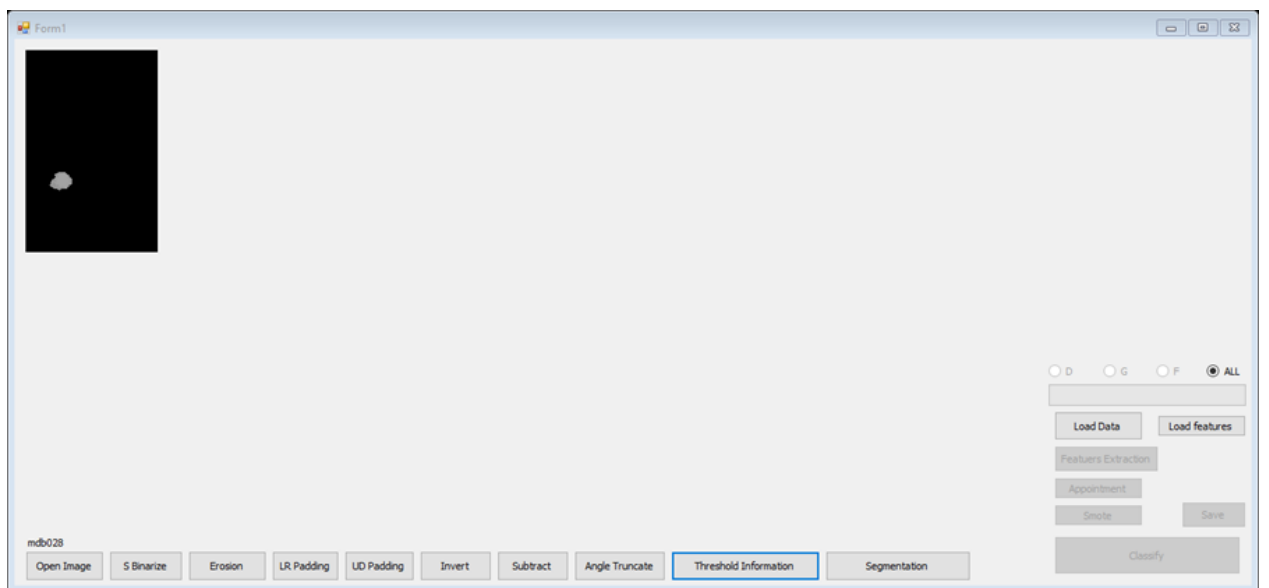
**Figure B. 7** Inverting step



**Figure B. 8** Subtracting step

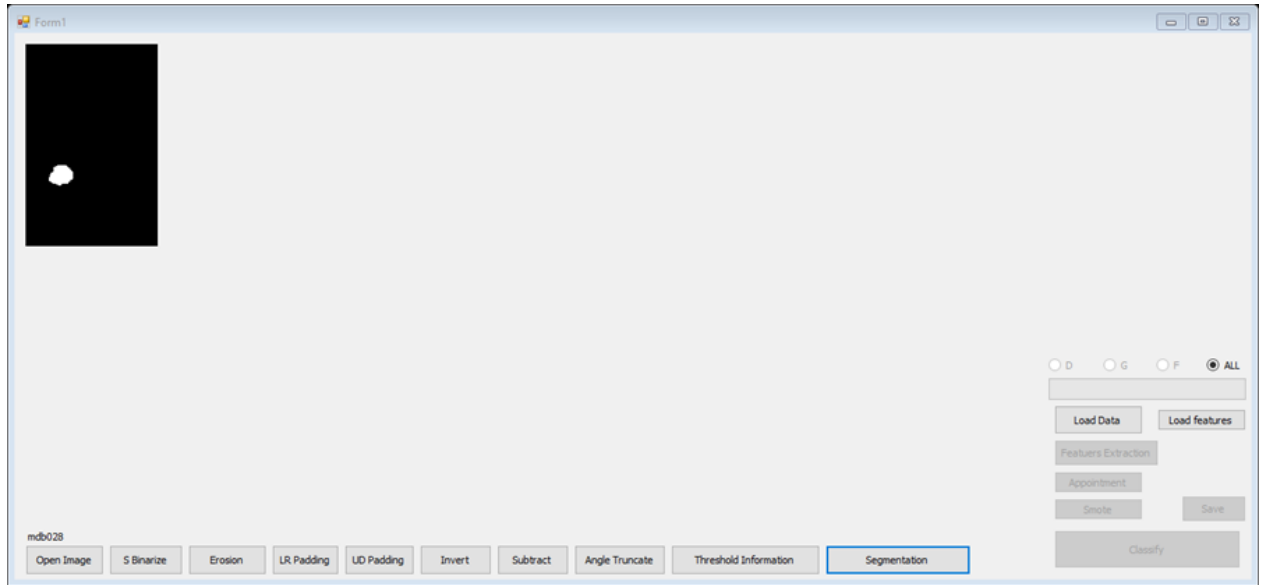


**Figure B. 9** Pectoral muscle removal step



**Figure B. 10** Pre-segmentation (ROI) step





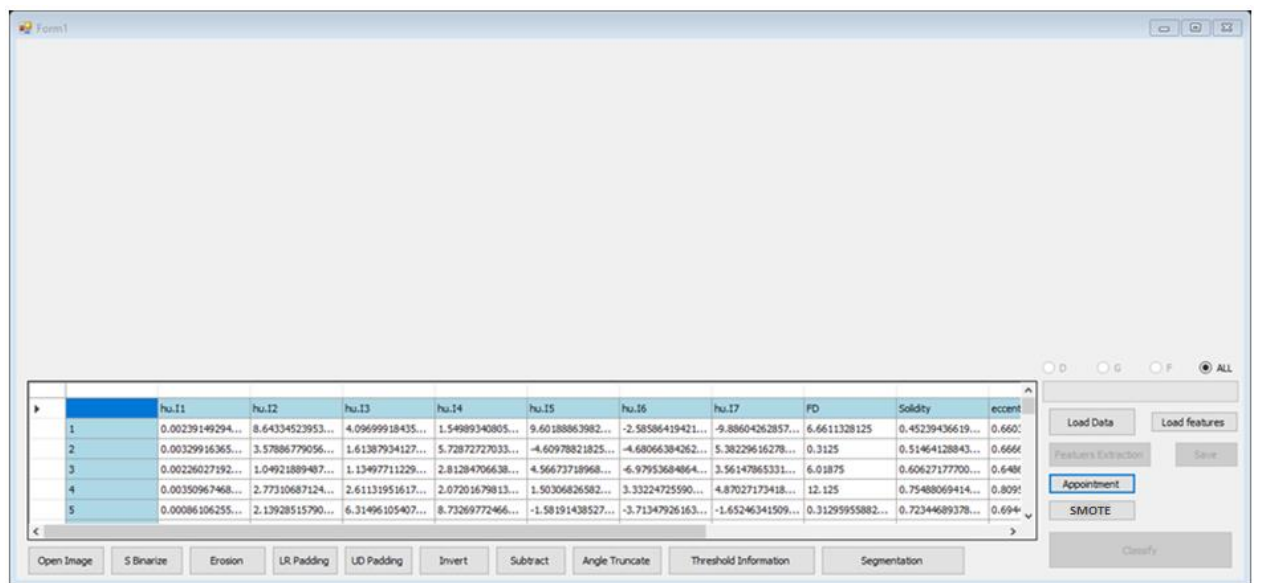
**Figure B. 11** Segmentation step



**Figure B. 12** Dataset loading step



**Figure B. 13** Features extraction step for the loaded dataset



**Figure B. 14** Minority splitting step

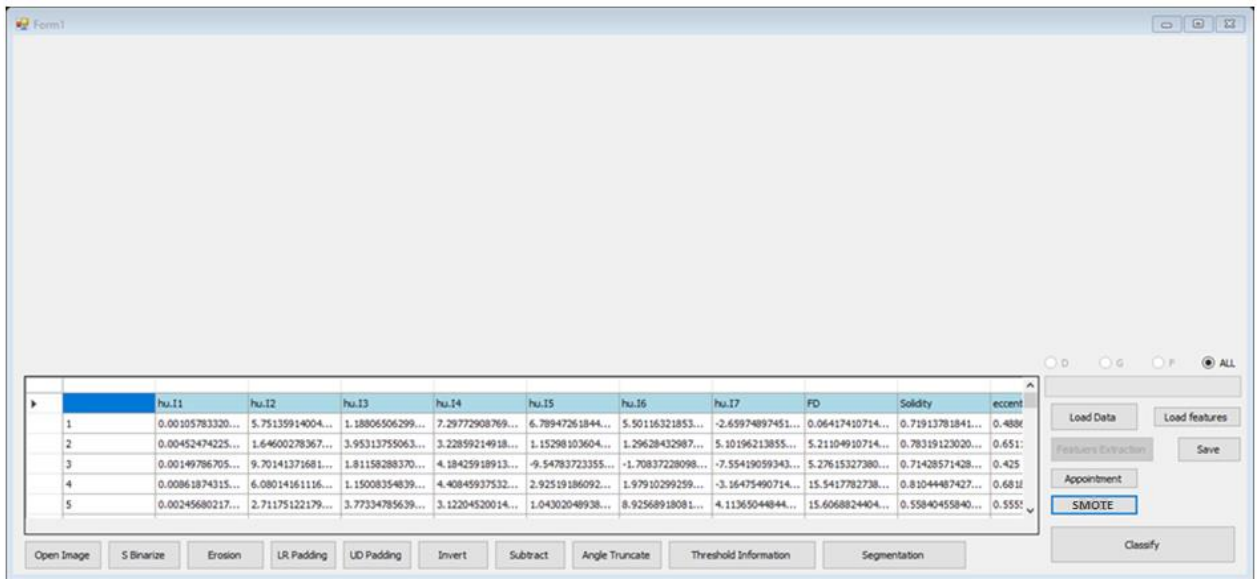


Figure B. 15 Applying SMOTE on minority samples step

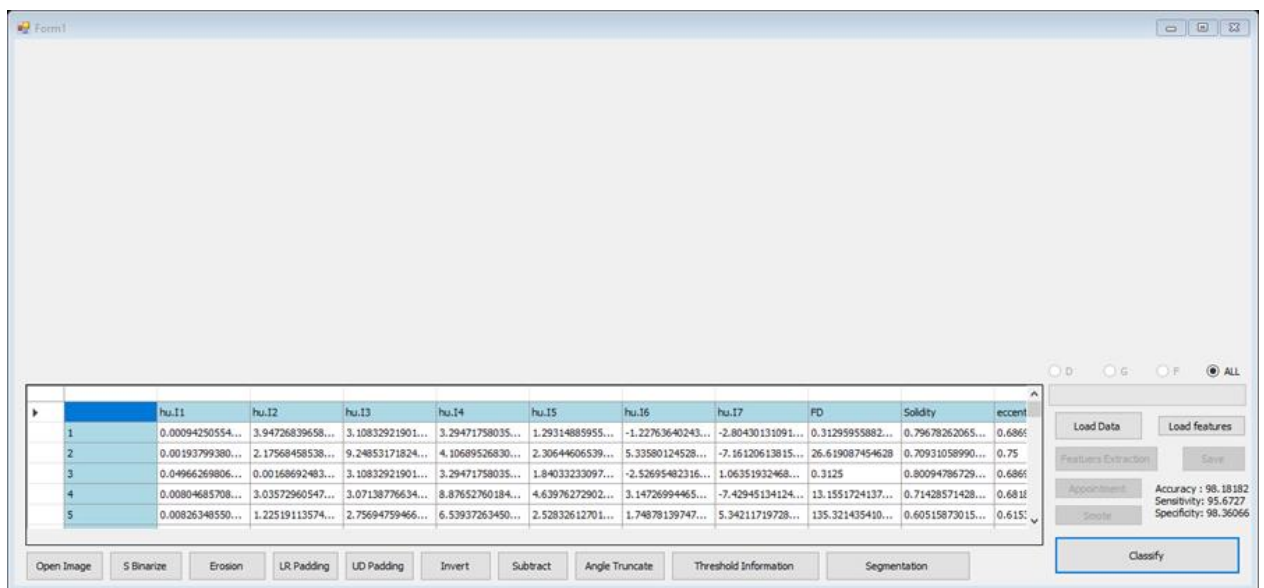


Figure B. 16 Classification step



حکومەتی هەریمی کوردستان  
وەزارەتی خویندنی بالا و توێژینەوێ زانستی  
زانکۆی سلیمانی  
کۆلیجی زانست

## باشترکردنی تەکنیکەکانی دیاریکردن بۆ پۆلینکردنی شێرپەنجە مەمک بە بەکارهێنانی فێربوونی ئامێر

تیژی دکتۆرایە  
پێشکەشکراوە بە ئەنجومەنی کۆلیجی زانست لە  
زانکۆی سلیمانی وەک بەشێک لە پێداویستیهکانی بە دەست  
هێنانی پروانامەی دکتۆرای فەلسەفە لە  
( زانستی کۆمپیوتەر )  
پروۆسیسکردنی وێنە

لەلایەن

سروە حسن عبدالله

ماستەر لە زانستی کۆمپیوتەر (2011)، زانکۆی جواهرلال نهره حیدرآباد

بە سەرپەرشتی

د. هادی وهیسی

پروۆفیسۆری یاریدەدەر

د. علی مکی صغیر

پروۆفیسۆر

## پوختەى تېز

شېرپەنجە كۆمەللىك نەخۆشپىيە كە تىايدا خانە كانى لەش بەشپۆهە كى كۆنۆزۆل نە كراو دە گۆرپىن و گەشەدە كەن، زۆر بەى خانە شېرپەنجەپىيە كان لە كۆتايى دا بارستەپەك دروست ئە كەن كە ناسراو بە گرى يان لوو. ئەم گرى ياخود لووانە يان جۆرى بى زيانن باش (benign) وە يا جۆرى كوشندەن شېرپەنجەپىيە (malignant)، جۆرى بى زيان ياخود باش شېرپەنجەپىيە و خانە كانى ئاسايى دەرنە كەون و گەشە كەرنىيان هېواشە وە شانە كانى دراوسى داگىرناكات و بۆ ناوچە كانى ترى جەستە بلاو نابنەو. لوى جۆرى كوشندە شېرپەنجەپىيە و بەخىرايى بلاودەپىتەو بۆ بەشە كانى ترى جەستە، يە كىك لە جۆرە كانى شېرپەنجە كە دەبنە هۆى مردن بە زۆرى برىتپە لە شېرپەنجەپىيە مەمك.

شېرپەنجەپىيە مەمك يە كىكە لە هۆكارە سەرەكپىيە كانى مردنى ژنان لەسەرانسەرى جىهان ئەوئىش بەهۆى قورسى و سەختى لە دۆزىنەو و دەستنىشان كەرنى، وە بەشپۆهە كى گشتى ناتوانرپت بەر بەم نەخۆشپىيە بگىرپت ئەوئىش بەهۆى ناديارى هۆكارە كانى تووشبون. هەرچەندە دۆزىنەو و دەستنىشان كەرنى نەخۆشپە كە لەقۇناغە سەرەتايپە كانى دا يارمەتيدەرە بۆ كەمكەرنەو و رپژەى مردن و هەلئار دەى چارەسەرى باشتر بۆ نەخۆشە كان داين دەكات. ئەگەر شېرپەنجەپىيە مەمك لە قۇناغە سەرەتايپە كانى دا دەستنىشان بگىرپت، دەتوانرپت رپژەى چاكبونەو و چارەسەرى زياد بگىرپت.

لەئىستادا تەكنىك و شپۆزى وپنە گرتنى جۆراو جۆر هەپە لەوانە سۆنەر (ultrasound)، تۆمۆسپىتپىننى دىجىتالى مەمك (DBT)، وپنە گرتن بە لەرپنەو و مۆگناتپسى (MRI) وە مامۆگرافى (mammography) كە لەئىستادا بە كاردپن بۆ دەستنىشان كەرنى و دۆزىنەو و شېرپەنجەپىيە مەمك. يە كىك لە بەسوودترىن و بەرچاوترىن تەكنىكە كان بۆ دەستنىشان كەرنى شېرپەنجەپىيە مەمك لە قۇناغە سەرەتايپە كاندا برىتپە لە مامۆگرافى، بە بەكارهپننى رپكارپكى ووزەى نزم بۆ بەدەستپىننى وپنە بىنراوى پىكها تە كانى بەشى ناو وەو و مەمك كە ئەمەش وەك تەكنىكپكى وپنە گرتنى جىپى متمانە و گرنىگ نىشان دراو. مامۆگرافى ئامرازپكى باو كە تەكنەلۆجىپى وپنە گرتن بەكار دەهپىت بۆ دۆزىنەو و دەستنىشان كەرنى هەر نااسايى بوونپك وە فەراھەم كەرنى ئاسان بۆ ناسپنەو و هەر لوىە كى شېرپەنجەپىيە خراب. هەرچەندە پىداچونەو و سەپر كەرنى مامۆگرامە كان بە شپۆهە كى دەستى يان كەردارە كى كاتى ئەوئىت وە هۆكارە مۆپىيە كانپش بەرپرسن لە رپژەى هەلە.

لەگەل ئەو وى كە (CAD) سىستىمى دۆزىنەو و دەستنىشان كەرنى بەيارمەتى كۆمپىوتەر گەشە كەرنىپكى بەرچاوى نىشاندا كاتپك بەكارهپنرا بۆ دەستنىشان كەرنى لە وپنە مامۆگرامە كان، ئەم سىستىمانە تەكنەلۆجىپى كۆمپىوتەرى بە كاردپن بۆ دۆزىنەو و هەر نااسايى بوونپك لە مامۆگرامە كاندا. وە سوودى بەكارهپننى ئەم ئەنجامانە لەلاپن

پسپورانی تیشکەوہ رۆلئیکی گرنگ دەبینی بۆ دەستئیشان کردنی نەخۆشیە کە لە رینگە شیکردنەوہی وینەکان بە شیوہیەکی ئوتوماتیکی. رەنگە کارایی ئەنجامدانی (CAD) بگۆرپت لە نەخۆشیە کەوہ بۆ یەکیکی تر وە ناسینەوہی ئاسان نەبیت بەھۆی بوونی لیکچوون لە شانە ئاساییەکانی مەمک. ئەم جۆرە سیستمانە کەمتر باون بەلام لە بارودۆخی نەبوونی کەسانی شارەزا بۆ خۆئیندەنەوہی وینەکان دەتوانریت پەنای بۆ بېریت وەکو باریکی لەناکاو. ئەم تیزە دوو شیوازی کارکردن دەخاتە روو کە بە م شیوہیەکی لای خواروہ پۆلئینکراوہ:

لە شیوازی یە کەم دا رینگەکی نوێ پئشنیار کراوہ بۆ دۆزینەوہی ھەموو کلسە مایکروویہ ووردەکان لە وینە ماموگرافی دا ، پروسە پارچەکردنی پئشنیارکراو لەم رینگایەدا پئکھاتووہ لە دوو ئاست: لە ئاستی یە کەم، کۆکراوہی (k-means) بە کارھاتووہ بۆ جیاکردنەوہی ناوچە مەمک لە وینە کە. لە ئاستی دووہم، بە بە کارھێنانی رینگای باشتر کردنی ناوچە گەشەکردن (ORG) بۆ دەرھێنانی ناوچە خوارا (ROI) کە بۆ بە ئامانج گرتنی کلسە مایکروویہ ووردەکانی مەمک (MCs). دواتر ناوچە خوارا (ROI) پروسە دەکریت کە گروویک لە (26) تاییەتمەندی پئکھاتە (Haralick) دەرھێنران، لە گەل ئەوہشدا سی جۆری تری تاییەتمەندی (ھاوکۆلکە ی پەیوہندی ھاوبەش ، پەیوہندی پیرسون ، تیکرایی خالەکانی رووبەرە پارچە کراوہکان) بە دەست کەوت لە وینە پارچە کراوہکان. داتای ماموگرافیای سکرینی دیجیتالی (DDSM) بە کارھێنرا بۆ ھەلسەنگاندنی کارایی سیستمە کە بە سوود وەرگرتن لە پۆلئیکەری نامیری بریکاری پئشنگیری (SVM) بۆ جیاکردنەوہی شانە جۆری بی زیان باش و شانە جۆری شیرپەنجەیی. ھەستیاری سیستمی پئشکەشکراو گەیشتە سەرو 98.2% وە رپژە تاییەتمەندی دەستکەوتوو 100% بوو ھەر وھا رپژە ووردی و دروستی گەیشتە 98.82% ھەر ئەم ئەنجامانەش دەریدەخەن کە سیستمی دۆزینەوہ و دەستئیشانکردن بە یارمەتی کۆمپیوتەر باریکی بەلئندەرە بە کەمکردنەوہی رپژە مردن لە رینگە دەستئیشانکردنی پئش وەختە شیرپەنجە مەمک.

شیوازی دووہم بریتە لە رینگە دۆزینەوہ و دەستئیشانکردن بە شیوہی ئوتوماتیکی بۆ دیاریکردنی نااسایی بوونیک لە ماموگرامەکان دا. پئش ناسینەوہی نااسایی ، تەکنیکەکانی پروسە کردنی وینە بە کارھێنرا بۆ پارچەکردنی ناوچە گومانای ویسترا (ROI) بە شیوہیەکی دروست ، وە رینگە گەشە سەندووی ناوچەیی (region growing) بە کارھێنرا بۆ جیاکردنەوہی ماسوولکەکانی سنگ، دواي ئەوہ ناوچە گومانای ویسترا (ROI) پارچە کرا بە بە کارھێنانی (K-means) لە گەل لۆگاریتمی (thresholding). وە دواتر تاییەتمەندی (شیوہی بنچینەیی، نەگۆرە ساتییەکان، وە لە گەل دووریە کەرتییەکان) دەرھێنران لە ناوچە پارچە کراوی ویسترا (ROI) بۆ دۆزینەوہ دەستئیشانکردنی ھەر نااسایی بوونیک لە ماموگرامەکان. داتابەسی کۆمەلە شیکردنەوہی وینە ماموگرافی (Mini-MIAS) کە زۆر بە پئکھاتووہ لە نمونە جۆری بی زیان یاخود باش (benign) لە گەل رپژە کی کەم لە نمونە جۆری شیرپەنجەیی (malignant) بە کارھێنرا بۆ ھەلسەنگاندنی تەکنیکی پئشنیارکراو. ھەر وھا تەکنیکی

زیاده نمونه گرتنی که مینه دروستکراوه کان (SMOTE) به کارهات بۆ دابینکردنی نمونهی نوی له چینه که مینه کان بۆ به دهستهینانی کۆمهله داتایه کی هاوسهنگ بۆ نهجامدانی کارایی باشتری پۆلینکهه. پاشان پۆلینکهه داری دارستانی ههرمه کی (RF) به کار هینرا بۆ مه بهستی پۆلینکردنی ناوچهی بهش بهش کراو بۆ جوری بی زیان یا خود باش (benign) یان شیرپهنجهیی (malignant). ووردی و دروستی، ریژهیی ههستیاری، ریژهیی تایهتهندی نهجامه تاقیکاریه کان بریتی بو له 97.1%، 95.8%، و 98.4% بهدوایی به کدا. بهم شیوهیه، سیستمه که بهو ده رنهجامه ده گات که پۆلینکردنی ماموگرافی به شیوهیه کی کارا به به کارهینانی مۆدیل و شیوازی پیشنیارکراو نهجامدراوه وه هاوکاری و یارمهتیه کی گرنگ پیشکدهش به پسیورانی تیشک و سۆنه ر نهکات له دهستنیشانکردن و دۆزینه وهی شیرپهنجهی مه مک.



حكومة اقليم كوردستان  
وزارة التعليم العالي و البحث العلمي  
جامعة السليمانية  
كلية العلوم

# تحسين تقنيات الكشف لتصنيف سرطان الثدي باستخدام التعلم الآلي

أطروحة

مقدمة الى مجلس كلية العلوم في جامعة السليمانية كجزء من متطلبات نيل  
شهادة الدكتوراه فلسفة في (علوم الحاسبات)  
معالجة الصورة

من قبل

**سروه حسن عبدالله**

ماجستير في تكنولوجيا المعلومات (2011)، جامعة جواهرلال نھرو حيدرآباد

بأشراف

**د. هادي ويسی**

أستاذ مساعد

**د. علي مكي صغير**

أستاذ

أكتوبر 2022

شوال 1444



## المستخلص

السرطان عبارة عن مجموعة من الاضطرابات التي تتغير فيها الخلايا بشكل غير مسيطر عليه . تشكل معظم الخلايا السرطانية في النهاية كتلة تعرف باسم الورم. يمكن أن تكون الأورام حميدة أو خبيثة، الأورام الحميدة ليست سرطانية لأن خلاياها تبدو طبيعية فهي تنمو ببطء ولا تغزو الأنسجة المجاورة أو تنتشر إلى مناطق أخرى من الجسم، الأورام الخبيثة سرطانية تنتشر بسرعة إلى مناطق أخرى من الجسم. من أنواع السرطانات التي تسبب الوفاة هي سرطان الثدي.

يعد سرطان الثدي من الأسباب الرئيسية للوفيات بين النساء في جميع أنحاء العالم بسبب صعوبة اكتشافه. لا يمكن منعه تماماً لأن السبب غير معروف. ومع ذلك ، فإن اكتشاف سرطان الثدي في مرحلة مبكرة يساعد على تقليل الوفيات ويوفر للمرضى خيارات علاجية أفضل ويمكن زيادة معدلات الشفاء من سرطان الثدي إذا تم اكتشافه في مرحلة مبكرة.

تُستخدم حالياً تقنيات التصوير المختلفة ، بما في ذلك الموجات فوق الصوتية (ultrasound)، التركيب الرقمي للثدي (DBT)، التصوير بالرنين المغناطيسي (MRI)، والتصوير الشعاعي للثدي (mammography)، للكشف عن سرطان الثدي وتشخيصه. يعد التصوير الشعاعي للثدي من أكثر التقنيات فائدة وأهمية للكشف عن سرطان الثدي في مرحلة مبكرة. باستخدام إجراء منخفض الطاقة للحصول على صور مرئية للبنية الداخلية للثدي ، فقد تم إثباته كطريقة فحص مهمة وجديدة بالثقة. يمثل التصوير الشعاعي للثدي أداة شائعة تستخدم تقنية الفحص للكشف عن العيوب وتمكين التعرف على الأورام الخبيثة بسهولة. ومع ذلك ، فإن المراجعة اليدوية لعدد كبير من صور الثدي بالأشعة السينية تستغرق وقتاً ، والعوامل البشرية مسؤولة عن معدل الخطأ. بالإضافة إلى ذلك ، يُظهر الكشف/التشخيص بمساعدة الكمبيوتر (CAD) تحسينات عند استخدامه في تشخيص التصوير الشعاعي للثدي القائم على الصور (mammogram).

تستخدم هذه الأنظمة تقنية الكمبيوتر لاكتشاف التشوهات في تصوير الثدي بالأشعة السينية ، ويلعب استخدام هذه النتائج من قبل أطباء الأشعة للتشخيص دوراً رئيسياً بعد أن تتميز الآفات بالتحليل التلقائي للصور. قد يختلف أداء (CAD) لأن بعض الآفات يصعب تحديدها أكثر من غيرها بسبب أوجه التشابه مع أنسجة الثدي الطبيعية. هذه الأنظمة أقل شيوعاً ، ولكن عندما لا يتوفر مراقبون بشريون خبراء ، يمكن استخدامها في حالات الطوارئ. قدمت هذه الرسالة طريقتين رئيسيتين يمكن تصنيفهما على النحو التالي:

في الطريقة الأولى تم اقتراح طريقة جديدة لاكتشاف التكتلات الدقيقة في صور التصوير الشعاعي للثدي. تتكون عملية تجزئة الطريقة المقترحة من مستويين: في المستوى الأول ، تم استخدام مجموعة الوسائل (k-means) لعزل منطقة الثدي من الصورة ، ثم تم استخدام طريقة منطقة نمو المحسنة المقترحة (ORG) لاستخراج منطقة الاهتمام (ROI) ، والتي تستهدف التكتلات الدقيقة للثدي (MCs). لاستخراج الميزات ، تتم معالجة المنطقة ذات الأهمية بعد ذلك حيث يتم استخراج مجموعة من (26) ميزة نسيج باستخدام خصائص نسيج (Haralick). بالإضافة إلى ذلك ، يتم الحصول على ثلاث ميزات (معامل الارتباط المتبادل ، ارتباط بيرسون ، ومتوسط مساحة النقاط المجزأة) من الصورة المجزأة. تم استخدام مجموعة بيانات التصوير الشعاعي للثدي

(DDSM) لتقييم كفاءة النظام باستخدام مصنف (SVM) للتمييز بين الأنسجة الحميدة والخبيثة. بلغت حساسية النظام المقترح 98.2% والنوعية 100% والدقة 98.82%. تظهر النتائج أيضاً أن التشخيص بمساعدة الكمبيوتر هو مجال واعد لتقليل الوفيات من خلال الكشف المبكر عن سرطان الثدي.

توفر الطريقة الثانية طريقة تشخيص آلية لاكتشاف أي خلل في تصوير الثدي بالأشعة السينية تلقائياً. قبل تحديد الاضطرابات ، تم استخدام تقنيات معالجة الصور لتقسيم منطقة الاهتمام المشبوهة (ROI) بشكل صحيح. تم إجراء طريقة منطقة الاهتمام (ROI) لعزل العضلات الصدرية. بعد ذلك ، تم تجزئة منطقة الاهتمام المشبوه باستخدام خوارزمية (K-means) و (thresholding) ، بعد ذلك ، تم استخراج الخصائص القائمة على الشكل ، وثوابت اللحظة ، والأبعاد الكسورية من منطقة الاهتمام (ROI) الجزأ من أجل اكتشاف التشوهات في تصوير الثدي بالأشعة السينية. تم استخدام مجموعة بيانات (Mini-MIAS) ، والتي تتكون في الغالب من عينات حميدة مع نسبة صغيرة جداً من العينات الخبيثة ، لتقييم التقنية المقترحة. تم استخدام تقنية جمعية تحليل صورة الثدي (SMOTE) لتوفير عينات جديدة من فئات الأقليات للحصول على مجموعة بيانات متوازنة وتحقيق كفاءة مصنف أفضل. تم استخدام مصنف الغابة العشوائي (RF) لتصنيف المنطقة الجزأ على أنها حميدة أو خبيثة. كانت دقة وحساسية ونوعية النتائج التجريبية 97.1% و 98.4% و 95.8% على التوالي. وهكذا ، يستنتج النظام أن تصنيف التصوير الشعاعي للثدي يتم بكفاءة باستخدام النموذج المقترح ، مما يوفر المساعدة الأساسية لأخصائيي الأشعة في الكشف عن سرطان الثدي.