POLITECNICO DI TORINO

Master's Degree in ICT for Smart Societies

Master's Degree Thesis

Breast Tumor Detection by Mammogram Image Segmentation

Supervisors

Prof. Michela MEO

Prof. Guido PAGANA

Prof. Alex FERWON

Prof. Robin AUGUSTINE

Candidate

Mersedeh KOOSHKI FOROOSHANI

July 2021

To Mom and Dad
who have supported me to be the person I am today

Summary

A neoplasm is an abnormal mass of cells with uncontrolled and progressive growth in the body caused by genes mutation. It can originate from any tissue in the body. Neoplasm comprises benign and malignant tumors. Cancers are malignant tumors that can spread to almost anywhere in the body. Breast cancer is the second leading cause of cancer-related death among women worldwide. The complexity and plurality of the lesions affect precise masses diagnosis by specialists. While detection of breast cancer at early stages can expand access to high-quality prevention and treatment services for the affected women. Therefore, timely and primal diagnosis of breast cancer can play a significant role to reduce mortality rates.

As an achievement in medical science, the medical imaging process has been introduced to take images of the human body for clinical purposes including diseases diagnosis. Nowadays mammography is the most efficient imaging modality used by radiologists for the screening of breast tissue. In recent decades, Machine Learning (ML) algorithms have had vast utilization in healthcare to assist physicians in more precise assessment. ML can act as a non-invasive, easy-to-use, practical, and inexpensive diagnosis procedure and even can eliminate human error and increase the accuracy of diagnosis.

Finding an accurate, robust, and efficient image segmentation technique still remains a challenging problem in digital mammography. Extraction of the region of interest (ROI) and isolating breast tissue from pectoral muscle is an essential pre-processing step before further analysis for mass detection and classification. In fact, the ROI extraction allows the search for abnormalities to be restricted to the breast tissue region.

The main purpose of this thesis is to detect abnormalities in mammography images. To achieve the desired objective, a clustering-based segmentation method is implemented. The automated segmentation technique is done through the execution of an appropriate unsupervised ML algorithm known as the standard Fuzzy C-Means clustering algorithm. In order to segment breast tissue and extract

the ROI with high accuracy an enhancement method which is the Kirsch operator applied to the mammogram images. Finally, the Thresholding method is performed to detect the tumors in ROIs. One of the most noticeable aspects of this project is to combine different ML algorithms and techniques effectively somehow that they can work in the same direction and complement each other.

Acknowledgements

I wish to express my deepest appreciation to my Supervisors **Professor Meo** for all the scientific and emotional support, **Professor Pagana** for his oversight and continued support throughout my thesis work, **Professor Ferworn** for all the invaluable advice and feedback to better my thesis and to choose the right direction, and **Professor Augustine** for his review of my hard work and for all of the opportunities I was given to further my research. I am also grateful to **Italy** and the **Polytechnic University of Turin** which gave me the opportunity to increase my knowledge at a high-quality international university in the most beautiful country in Europe.

Additionally, my appreciation goes out to my **parents** for their infinite love and their inspiration in my life: You are always there for me. My gratitude extends to my dear **Aunt Nahid** for her persistent guidance in my life, to my supportive brother, **Siamak**, to my lovely sister-in-law, **Sanaz**, and to my beloved nephew, **Hoomaan**, for encouraging me to move forward to follow my dreams. And also to my only sister, **Melika**, who is part of my soul and never allows me to be disappointed in my life, and without whose assistance, finishing this thesis would not have been possible. I would like to express my special thanks to **Reza** who is always in my heart and believes in me more than anyone else in the world.

Finally, I would like to wholeheartedly thank my friends in Italy: **Sahra**, whose kindness makes this academic journey enjoyable and who has supported me through my toughest times, and **Arman**, for his compassionate personality and his unforgettable assistance during my thesis work. Thank you for the good moments of suffering and frustration and for continuously inspiring me as we learned from each other.

Mersedeh Kooshki Forooshani
Torino 2021

Table of Contents

List of Figures

Acronyms

ML Machine Learning

CAD Computer-assisted detection/diagnosis

DIP Digital Image Processing

FCM Fuzzy C-Mean

ROI Region Of Interest

AI Artificial intelligence

ANNs Artificial Neural Networks

DTs Decision Trees

DM Digital Mammography

FSM Film-screen Mammography

2D Two Dimensional

PHI Protect Health Information

HIW Histogram-based intensity windowing

TP True Positive

FP False Positive

TN True Negative

FN False Negative

Chapter 1

Thesis Statement

Analysis of mammogram images by applying machine learning algorithms for breast segmentation, region of interest extraction, and abnormalities detection aids physicians to identify suspicious lesions with high accuracy at early stages, and this leads to saving patients' lives since breast cancer is a prevalent cause of death among females in the world.

Chapter 2

Contributions

This thesis contributes to the early detection of breast lesions in mammogram images and the result confirmed machine learning development in the healthcare system is an inexpensive, easy-to-use, and practical approach, which has a huge impact on the detection of disease with high precision, and this reduces morbidity and mortality among women suffering from breast cancer.

This research affirmed the high efficiency of machine learning and artificial intelligence methods to assist radiologists and specialists in more accurate diagnostic and eliminate human error. According to this project, the researchers can perform further analysis to identify the stage of breast cancer by focusing on tumors' size or evaluate the ratio of different tissue densities and figure out its relation to the breast cancer probability. There is also the possibility to apply this model for tumor detection in other body tissues.

Chapter 3

Introduction

A (malignant) neoplasm is an abnormal mass of cells with uncontrolled and progressive growth due to genes mutation. It can originate from almost any tissue in the body. Generally, the human cells break down via apoptosis or grow and divide into healthy new cells within a regular process which is known as cell division but during the time, mutations in genes can lead to cells that starting to grow without limitation, producing more cells, and creating a tumor [1] [2].

Breast cancer, the second most common cancer worldwide and a leading cause of cancer-related death among women, mostly originates ductal or lobular cells in the breast and invades the surrounding tissues. Breast cancer stage refers to how far the cancer cells have spread over the main tumor. Breast cancer is always caused by genome instability due to an abnormality inherited from a parent or as a result of acquired mutations [3] [4].

X-ray is electromagnetic radiation that is mostly known due to its wide benefits in medical imaging [3]. while, it is also used in some other various fields. For instance one of the other vast utilization of X-rays is in the industry because it has the capability to identify any changes or deficiencies in the internal and external parts of devices that are not visible to the naked eyes. Furthermore, the X-ray machine is an effective tool for detecting metal or even bomb at the airport during the passenger or luggage checking process. The scanner technology is also done to provide security for the government buildings [5].

However, the medical society extremely used the considerably potential of X-rays for disease diagnostic intentions [5]. In order to generate a radiograph, the desired part of the human body is located between an x-ray source and an x-ray detector. Low X-ray images of the breast are known as mammograms [3]. The Digital Database for Screening Mammography is a resource that has image analysis purposes to assist breast cancer detection [6].

In the mammogram images, the background is black, and the breast shows up in gray and white. A normal mammogram is generally mostly gray, with some white areas showing healthy dense tissue. Denser tissue shows up white. Sometimes people have more dense tissue in their breast, and that makes the detection of abnormalities on a mammogram more difficult since a tumor is also made up of dense tissue and will appear white. If an area does not look like natural tissue so it is worrying, as a matter of fact, the radiologist will check the images for white regions and high-density tissue and notice its size, shape, and edges. Nowadays, mammograms are one of the best non-invasive methods for detecting breast cancer or checking the treatment trend [7].

A highly trained radiologist or an expert is required for the interpretation of mammograms. This process can be time-consuming, prone to interpretation variability and errors, and interpreter misdiagnosis and bias, including both false positives and false negatives. In breast image analysis for cancer diagnosis, the complexity and multiplicity of the lesions possibly will occur with dense tissue interactions, so it will be difficult for radiologists to distinguish and investigate masses accurately. Due to these difficulties and constraints, considerable interest has arisen in the potential of Computer-assisted detection/diagnosis (CAD) tools, which typically rely on Machine Learning(ML) algorithms to automate medical image analysis [8]. In the past decade, many computer-aided systems and imaging techniques have been developed to assist radiologists and classified the mammogram images and to achieve robustness, efficiency, better accuracy, and precision [9] [10].

ML is certainly one of the most exciting fields in computer science due to its capability to extract useful patterns from instances, which indeed is similar to the ability of human brainpower. Thus based on this fact, it is marked as part of artificial intelligence [6]. ML algorithm is able to learn based on patterns hidden in training data and then it is possible to make predictions on the test data. The hidden patterns about a problem can be used to predict future events and implement all kinds of complex decision-making. Generally, ML mentions to tasks associated with artificial intelligence such as recognition, diagnosis, prediction, etc. Along with medical investigation, ML is also applied to many other fields, such as navigation of autonomous vehicles, and marketing [6]. ML is not only about working on data to find a formula for answering a problem but it also uses data to figure out the rules behind them. Hence, a learning process is essential for the machine to test various rules and learning how well each of them performs. Actually, a dataset is a set of data, and the important pieces of data, which help to understand a problem, are called features [3] [6]. Then during the learning process ML algorithm has learned from the data it is shown during training, therefore, the data is more critical than any other thing in the process. Finally, the output that is created is known as a model.

Indeed, the three main steps of a process in each ML project are **preparing or selecting a dataset**, **defining the appropriate ML algorithm to create a model**, **evaluating the accuracy, and refining the model** [7]. According to the algorithm and objectives, it is possible to divide ML algorithms into three main groups:

1. **Supervised learning:** labeled data is used to train machines in order to make them learn and find relationships between inputs and outputs [8].

2. **Unsupervised learning:** means there is no labeled data.

3. **Reinforcement learning:** is the field of deep learning that includes learning by reward or prediction of reward [9].

In this study, according to the following steps an unsupervised ML algorithm is implemented to perform the learning process:

- Step 1: firstly, it is required to load the unlabeled data into the system.

- Step 2: then analyzes the data.

- Step 3: afterwards, the algorithm has to search for patterns based on the features or specific behavior in the dataset.

- Step 4: finally, when a pattern is specified, the output comes out.

Unsupervised ML has some advantages with respect to Supervised ML. In fact, not having labeled data seems good in some cases and Unsupervised ML algorithms have the capability to interpret data on its own. This type of ML techniques is much faster in terms of performance in comparison to Supervised ML since no data labeling is needed here [10].

3.1 Digital Image Processing (DIP)

In computer science, digital image processing is a method to apply algorithms on digitalized image for analysis, in order to achieve an enhanced image or to extract some features or beneficial information.It can be applied in different fields such as Object detection, Diagnosis of tumors in medical imaging, Measuring tissue volumes, Traffic control systems, Face recognition [3]. The major digital image processing techniques are pre-processing, image compression, image restoration, edge detection, morphological image processing, and segmentation [3].

Nowadays, digital images have an undeniable role in different fields of study. Medical imaging processing refers to the management of images by using computer systems. Medical imaging is the method used to acquire images from different parts of the body in order to identify diseases [3] [11] [12].

3.2 Benefits of Medical Image Processing

Advances in the technology of medical image processing benefit the healthcare system over the past few decades. Physicians, for instance, are able to make the net diagnosis through a non-invasive accurate procedure, reduce mismanagement, and allow better outcomes for patients' treatment [13]. Automated image analysis will decrease errors in diagnosis, by eliminating the need for radiologist assessment in the search for anomalies.

Computer algorithms have a vital role in medical image analysis like tissue volume valuation, diagnosis, treatment planning [14]. Clustering is an unsupervised ML that is about the allocation of a set of data into subsets, which are known as clusters so that data in the same cluster are similar in some features. Cluster analysis is used in many fields, including pattern recognition, and image analysis [14] [15].

Fuzzy c-means(FCM) which is one of the most common clustering algorithms can be used for the segmentation of medical images [16] [17]. Image segmentation designs to divide an image into several regions that are corresponding to the requirements, and it is always requiring much skill or effort due to the complexity of images. It subdivides a dataset X into C-clusters. Mostly other techniques are added to FCM for improving the results [17].

In this thesis, a novel approach is proposed to analyze the breast tissue, extract the Region of Interest(ROI) and find the accurate distinction between the border of the tumor and breast. FCM technique has been applied to the clustering of the mammogram images and in order to obtain the experimental result with high accuracy, Kirsch Compass Kernel algorithm as a non-linear edge detection method [18] is performed on the mammogram images earlier. Therefore, the most possible edge points of images are detected and subsequently image segmentation with high accuracy will be achieved. As the next stage, the main breast tissue as the ROI is extracted from images and finally, the tumors can be distinguished by applying the best threshold to the most likely regions.

Chapter 4

Literature Survey

Screening mammography intends to recognize breast cancer at earlier stages of the disease, when treatment can be more successful. Although there are screening programs worldwide, the interpretation of mammograms is affected by false positives and false negatives at high rates. However an Artificial Intelligence(AI) system is able to surpass human experts in breast cancer prediction [19] [20].

4.1 Artificial Intelligence and Machine learning Review

AI is a part of computer science that tries to make more intelligent systems while one of the needful capabilities of intelligent behavior is learning. Machine Learning(ML) is one of the main subfields of AI research that rapidly developed [21]. There is significant interest in the use of ML in medicine[22]. From the beginning, machine-learning algorithms were designed and used to analyze medical data sets. Modern hospitals are equipped with essential tools for collecting and monitoring data, which is also possible to share in large information systems. ML technology is currently well suited for analyzing medical data [21], and in fact, ML techniques can 'learn' from the vast amount of healthcare data currently available, in order to assist clinical decision making [22]. In particular, it helps lots of work to be done in the medical diagnosis field. As soon as computer systems came into use in the fifties and sixties, different algorithms were introduced to provide modeling and analyzing large amounts of data, and at the start, three major branches of ML emerged.

Classical work in symbolic learning is developed by Hunt et al. (1966), in statistical methods by Nilsson (1965), and in neural networks by Rosenblatt (1962). Through the years all three branches developed advanced methods (Michie et al., 1994)[21]. Big data and machine learning have a huge effect on all aspects of modern life. Thus, it is obvious that the application of machine learning and big data had a big revolution on medicine and health care. Recent efforts have developed such machine learning algorithms that can perform on par with human physicians [23]. ML is classified into three main types: supervised, unsupervised and semi-supervised learning. For supervised learning, data are labeled, and the input data can predict the output data, while data in unsupervised learning are not labeled. Most of the data are not labeled in semi-supervised learning, and supervised and unsupervised techniques are used together [24] [25].

ML is not new to cancer research. Artificial neural networks (ANNs) and decision trees (DTs) have been used in cancer detection and tumor diagnosis for more than 20 years (Simes 1985; Maclin et al. 1991; Ciccheti 1992). Nowadays machine learning algorithms are being used in lots of applications such as detecting and classifying tumors via X-ray images. In fact, machine learning methods have extended usage in the identification, classification, detection of tumors, and other malignancies [26]. ML methods have begun to establish themselves as an alternative approach to computer-based data analysis in oncology, as this field moves gradually away from being the preserve of traditional statistical analysis [27]. Supervised and unsupervised learning are the two prominent machine-learning algorithms used in pattern recognition and classification [28].

4.2 Computer-Aided Diagnosis(CAD) Systems and Medical Imaging

Among all of the known cancers, breast cancer is a major concern among women. It is the second-most common and leading cause of cancer deaths among women. According to published statistics, breast cancer has become a major health problem in both developed and developing countries over the past 50 years, and its incidence has increased in recent years. However, early detection of breast cancer can give patients a more chance in treatment and for a full recovery. Thus, efficient diagnostic at the early stages of breast cancer can play a vital role in the reduction of morbidity and mortality rates. Computer-aided detection or diagnosis (CAD) systems use computer technologies to detect abnormalities such as calcifications, and masses in mammograms. The utilization of these results by radiologists, can have a significant effect on the early detection of breast cancer and help to deduct the death rate among women who suffer from breast cancer. Thus, in the past several years, CAD systems and related techniques have attracted the attention of both research scientists and radiologists. Radiologists, on the other hand, are attracted by the effectiveness of clinical applications of CAD systems [29], and although significant progress has been made over the last 20 years much works still needs to be done to develop more effective CAD system [30].

CAD implementation contains various fields such as enhancing the mammogram, identifying suspected regions, feature extraction from segmented mammograms, classifying the mammograms and so on [30]. During the 19th century, machines were more used for diagnosis or therapeutics in medical science (Raiser, S.J.,1978). The use of electricity resulted in the invention of the x-ray. Electromagnetic radiation in a wavelength range generally introduced as X-rays was earliest discovered by a German professor of physics, Wilhelm Roentgen. Because of the unknown nature of his discovery, he called them X-rays but they are also still known as Roentgen-rays, particularly in Germen-speaking countries. Physicians gained the ability to view the inside of the body, by using X-rays [31]. Medical imaging is the technique and process used to create images of the human body for clinical purposes including to diagnose diseases or medical science. Since the discovery of X-rays, medical imaging has undergone near continuous innovation. After the Second World War, several generations of novation and new discoveries concentrated on the relation of computerization and imaging technologies happened in X-ray, computed tomography, magnetic resonance imaging, nuclear imaging, and ultrasound-positioned medical imaging, which led to transforming healthcare science. Medical imaging has brought a high sense of vision into medical science, leading to an extensive change in the healthcare system [29].

Now medical imaging is one of the standards of new medical care for diseases such as cancer, trauma, and many others [29]. The considerable revolutions in the medical imaging industry are taking place; moving from expensive, large, static, and complicated systems to smaller, easier to use, and more available devices. The recent imaging technologies focus on combining ease-of-use with very high accuracy, allowing information to be accessed efficiently while providing higher throughput. These new solutions are cost-effective and can be used in a variety of clinical applications (MitchellMagaldi, D., 2008) [31]. Ultrasound generates images by using sound waves; MRI does it by magnetic fields. CT scanning and mammography presented new feature that makes new diagnostic power and clinical capabilities possible. The evolution of digital imaging brought a new generation in performance and speed as it offered new options for data access and transmission and develop new volumes of information(NEMA, 2006 Dec.) [31] [3]. Over the 20th century, medical imaging progress was due to technological innovation and engineering improvements in physical tools. A new imaging state driven by extension in the biological knowledge base represents a substantial change. The level of success and the economic formation of medical imaging reside firmly in the ability to implement new methods with higher diagnostic specificity and sensitivity (Fullerton, G.D., 2005; Hazle, J.D., 2005; Fullerton, G.D. and Hazle, J.D., 2008) [31].

Recently, much attention has been drawn to various imaging modalities and becoming more available for acquiring complementary information for variety anatomy [32]. The standard procedure for use in breast cancer screening is mammography [33]. Mammography is an X-ray examination of the breast used to detect breast cancer [34]. Both analogue mammography and Digital mammography(DM) have been proven to reduce mortality from breast cancer [33]. The radiologic signs of breast cancer include mass lesions, asymmetries between images of the two breasts, and architectural distortion. In order for breast cancer to be detected accurately at the earliest possible opportunity, all factors influencing the acquisition, display and interpretation of the mammogram must be optimized and those optimum conditions must be maintained over time. This process, referred to as quality assurance, requires the cooperative efforts of several individuals: the technologist (radiographer), the radiologist, the equipment manufacturer and the medical physicist [35]. Detection of breast cancer requires that the breast be properly positioned in the mammography system and appropriately compressed by the technologist (radiographer). The exposure factors must be properly selected. Furthermore, the equipment used to accumulate and illustrate the image must be correctly designed to include the whole breast tissue in a mammogram image with high quality, excellent spatial resolution, and minimum distortion. This should be accomplished at the lowest radiation dose to the breast, compatible with these aspects of image quality [35].

The successful use of screen-film mammography in breast cancer screening is one of the major achievements of medical imaging in this century [36]. However, the current standard, film-screen mammography (FSM), has several inherent limitations on image quality. DM was developed as a convenient alternative that is expected to improve the quality of breast imaging and reduce the radiation dose required [34]. In screen-film mammography, the film serves as the medium for both image acquisition and display. FSM capability is bounded for low-contrast lesions in breasts with high dense. Therefore, diagnosis often requires additional imaging, which results in more radiation exposure for the patient. When additional images fail to provide useful diagnostic information, a decision must be made as to whether the suspicious regions require biopsy or short- or long-term follow-up. Because of the expense and risk associated with additional radiation exposure and surgery, any method of image presentation that increases the diagnostic conspicuity of lesions in breast tissue, especially in dense tissue, would be a significant advance. Digital mammography systems, unlike screen-film mammography systems, allow manipulation of fine differences in image contrast by means of image processing algorithms [36].

DM was introduced in 2000 to offer new opportunities that are not provided by conventional FSM for the detection of breast carcinomas. The primary benefit comes from more reliable and efficient image management. The second one relates to the X-rays usage for breast imaging. DM also has the potential for increased efficiency in image archiving and retrieval, the possibility of avoiding the costs, complexity, and waste disposal problems associated with chemical processing of film [35]. Therefore, digital systems are a steady replacement for film mammography. Digital mammography differs from screen-film imaging in that the x-ray detector has a linear (or in some cases logarithmic) response to x-rays over a very wide range of exposure levels. Furthermore, the detector signal is digitized and stored in computer memory, thus the image can be demonstrated and manipulated independently of the acquisition operation. In spite of film in mammography, the brightness and contrast of digital images can be set independently of x-ray exposure while they are being viewed on a high-resolution monitor [35].

Today medical images represent a basic role in the diagnosis of numerous diseases. Ranging from anatomical information, functional activities, to molecular and cellular expressions, medical imaging means a direct observation from the different organs of human bodies with considering the anatomical changes and biological processes. Medical imaging usually requires skilled specialists to interpret the information revealed in the images in the best way. However, due to the various subjective factors plus limited analysis time and tools, it is prevalent that multiple doctors may prepare different interpretations, leading to several diagnoses. Moreover, for the same set of medical imaging, a medical doctor may give different diagnosis results at different times [37] [38]. Hence, the analysis of medical images is a very crucial task because imaging is a basic modality to diagnose any disease at the earliest stage. However, with the development of science and technology and the progress in medical imaging applications, manually analyzing data seems a challenging issue. Radiologists due to the inexperience or high workload may misinterpret diseases which leads to missed diagnosis, increment in false-negative or false-positive results thus non-lesions may be interpreted as lesions or benign lesions may be misinterpreted as malignant. In this case, CAD systems, which have been a promising area of research over the last two decades can be really helpful for radiologists in medical image analysis. The idea of using a computer to help in medical image diagnosis is not a new idea indeed the concept of CAD is to assist radiologists in interpreting medical images by using dedicated computer systems to provide 'second opinions' but the final medical decision is made by the radiologists. Studies on CAD technology shows that CAD aids radiologists for the improvement of diagnostic accuracy. The CAD system was originally developed for breast cancer screening from mammograms in the 1960s. Nowadays, it is one of the most important areas of research in the field of medical image analysis and It plays a supporting and final interpretative role in medical diagnosis [38] [39] [40] [41].

CAD is a technology that includes multiple elements like concepts of artificial intelligence (AI), computer vision, and medical image processing. The main application of CAD system is finding an abnormality in the human body. Among all these, detection of tumors is the typical application because if it misses basic screening, it leads to cancer [39]. CAD systems fundamentally work on highly complex patterns found in image. Sometimes CAD systems have positive impact in mammography analysis and make benefit for patients by achieving diagnosis with high accuracy and sensitivity, however there is no improvement in limited situation. Normally, CAD systems are optimized by number of images [39] [42]. Radiologists are already familiar with computer-aided detection/diagnosis (CAD) systems, which just highlight the presence or absence of image features known to be associated with a disease state. However, advances in algorithm development, combined with the ease of access to computational resources, allows AI to be applied in radiological decision-making at a higher functional level which is not a threat to radiology. It is indeed a tremendous opportunity for its improvement. In fact, similar to our natural intelligence, AI algorithms look at medical images to identify patterns after being trained using vast numbers of examinations and images [12].

4.3 Digital Image Processing Review

Digital image processing is a recent subject in computer history and it is important domain for many reasons. In the 1960s Bell Labs and the University of Maryland, and a few other places started to develop several techniques for digital image processing. However, the cost of processing was high with the computing equipment of that era. In the 1970s, image processing proliferated, when cheaper computers and dedicated hardware became available [43].

Digital image processing in medical image analysis is mostly used to detect or classify any abnormality [44]. In the 2000s, fast computers became available for signal processing and digital image processing has become the popular form of image processing. Because of that, signal image processing became a versatile method, and cheapest [43]. One of the major goal of digital image processing is to retrieve required information from the given image in a way that it will not effects the other features of that image [45]. Most image processing algorithms include steps such as **Preprocessing, Segmentation, Feature extraction, Feature**

Selection, and Classification.

Normally preprocessing is required to reduce the noise and improve the quality of the image, it has to be done on digitized images. To find Regions of Interest(ROI) the segmentation step is applied. Image segmentation is significant work because different breast tissues have different resolutions. The selection and removal of ROI are vital tasks in the segmentation stage of image processing. In general, image segmentation plays a crucial role in many medical imaging applications [46] [47]. In the feature extraction step, the desired features are determined based on specification of the region of interest. In feature selection, the best set of features are selected. Finally, on the basis of selected features, classification is performed in the classification step [46] [48].

Preprocessing is a common method to eliminate the noisy data in dataset and improve quality. It will prepare the medical images for further analysis [49]. The application of preprocessing varies depending upon the dataset [50]. Many researchers had developed many methods to perform preprocessing in different types of images such as gray-scale, color images and binary images. In radiological image, various techniques are proposed by many researchers to remove the distortion [51]. Image preprocessing can be done using different methods. Re-imaging, contrast enhancement, noise removal and mathematical operations are several preprocessing methods. Noise removal method has many techniques such as high-pass filtering, low-pass filtering, band-pass filtering, mean filtering and median filtering. Mathematical operations such as arithmetic and morphological operations are also used in the dataset to remove noise [50]. Morphology operations were used for preprocessing the input image [51]. It is a methodology for extracting shape and size information from an image. It includes a set of nonlinear operators, which effect on images by using structuring elements.

Originally, morphology represents a branch of biology that deals with the form and structure of plants and animals, and on the other hand, morphology is introduced as an image processing technique. Hence, it deals with the structure of objects. It aids to extract useful components of the image for representation of the shape of the region, boundary skeleton, etc.

All morphology algorithms are based on operations between the two sets. In digital image processing, one set is a whole image, or, some of its part, and the second one is the processing element, commonly referred to as the structuring element and also known as a kernel mask. The shape and the size of the kernel are responsible for the status of the operator on the image. In fact, morphology algorithms are derived for binary images: connected black pixels on the white background (or white on the black surrounding, depending on convention) represent sets in two-dimensional (2D) integer space [52]. Erosion and Dilation are the two basic morphological operators [50]. The invention of such mathematical tools dates back to 1964 and was meant for the filtering of binary images for mineral studies [53].

Morphological dilation is the operation that removes noise pixels, which are present in object areas by increasing objects' size. Morphological erosion is the operation that removes noise pixels that are present in the background by reducing the size of objects [54]. Due to the fact that medical images are difficult to be interpreted, therefore preprocessing techniques seem necessary to enhance the quality of the image and limit the search for abnormalities [46] [47] [55].

Segmentation is one of the most vital steps in CAD, especially for mass identification [56]. The importance of image segmentation cannot be neglected because it is used in almost every field of science [45], specifically in many signal processing techniques and their applications [43]. Segmentation is used to determine objects and boundaries in the image, and assign labels to every pixel in image such that pixels with the same label share certain visual characteristics. The result of image segmentation is a set of segments that collectively cover the entire image [57].

The segmentation process is to find the desired partitions of the image based on the required information. Algorithms based on classifiers have been widely applied to segment organs in medical images like brain images [43]. However, it is observed that there is not a perfect method for image segmentation, since each image has its own different type. It is also a very difficult task to find an appropriate segmentation technique for every group of images. Since a method applied to one image may not remain successful to another type of image, therefore segmentation techniques have been divided into different groups [45]. Obviously, the goal for all image segmentation processes is partitioning the image into regions or in other words, to cluster pixels of an image into image regions [43]. Some of the most famous image segmentation methodologies including edge-based segmentation, Fuzzy theory-based segmentation, Partial Differential Equation based segmentation, Artificial Neural Network bases segmentation, threshold-based image segmentation, and region-based image segmentation [45]. Recently there have been a number of researchers have tried to create several image segmentation algorithms as in (Krinidis et al., M Lange, et al., Mignotte, et al.)[43]. Segmentation algorithms are based on two basic properties of intensity values; one is a discontinuity and the other is similarity. Discontinuity is partitioning of an image based on abrupt changes in intensity. The similarity is partitioning of an image into regions based on a similar set of predefined criteria. Each segment of image reveals information in the form of color, intensity, or texture [58].

The anonymization of medical images and protecting data privacy is a challenging problem [59] with great importance in human life and science. During the past two decades, different methods for privacy-preserving in clinical records have been introduced to protect health information (PHI) [60] [61]. An essential issue for accessing clinical records outside of hospitals is the medical data de-identification, i.e., the removal, or replacement, of all mentioned PHI phrases entirely(PHI is any information in a medical record). Thus, a de-identified data can be made publicly available for non-hospital researchers as well, to facilitate research on human diseases [60] [62].

Data privacy-preserving means protecting the identity of patients and their sensitive information. Hence, this problem has been studied extensively in the context of medical data, by the computer science and Several models, algorithms and system designs have been proposed to protect identities of the published data [61]. Data de-identification is at the core of an AI/ML model [66], thus the use of ML algorithms has gained a lot of traction in recent years in order to make a model that achieves this task successfully [60] [63].

4.4 ML algorithm for medical image segmentation and classification

In order to improve the efficiency in disease diagnosis, some digital image processing algorithms have been investigated for detecting, segmenting, and classifying anomalies in medical images. The most popular approaches are the local and global thresholding, region growing algorithms, clustering methods, stochastic relaxation, fuzzy sets, and multi-scale techniques [64]. Clustering approaches are constructive tools to investigate data structures and have emerged as well-liked techniques for unsupervised pattern recognition and are applied in many application areas such as machine learning [65]. The aim of clustering is to assign instances into classes, which are not defined a priori, and which are supposed to somehow reflect the structure of the entities that the data represents [66]. Generally, clustering can be presented as the two main categories: hierarchical and partitioning clustering. Clustering can be either hard or fuzzy type [67].

Due to the overlapping nature of the cluster boundaries, some classes of patterns may be specified in a single cluster group or dissimilar group. This property limits the application of hard clustering in real-life. To, reduce such limitations fuzzy type clustering came into the picture and helps to provide more information about the memberships of the patterns [65] [67]. After fuzzy theory was introduced by Zadeh, the researchers use the fuzzy theory in clustering. Then in 1981, Bezdek added the fuzzy factor and he proposed FCM [67].

The Fuzzy clustering problems have been extensively considered and Bellman et al and Ruspini were proposing the basis of fuzzy clustering. Based on the objective function content, The FCM algorithm is a complete technique that is widely used. Whenever the algorithm is able to minimize an error function, it is known as C-Means which c represents the number of classes or clusters, and if the used classes are using the fuzzy technique or simply fuzzy, then it is introduced as FCM. In the FCM method, the pattern can be included in all the cluster classes, which have a certain fuzzy membership degree. Hoppner et al have made a good effort towards the survey of FCM [67] [68].

FCM methods are very well suited for segmenting the regions and have been extended and modified in many ways in order to solve the image segmentation problem [58]. As mentioned before, FCM is an important tool for image processing in segmentation of the color image. The FCM algorithm is commonly used to analyze different forms of data and images based on clustering ability. Particularly, FCM and its different methods utilized in the extraction of medical images to find the disease affected regions by means of segmenting the images on the basis of intensity [56]. It is an iterative clustering algorithm [58] and an extension of K-means clustering algorithm [69] in which a pixel can belong to more than one cluster and a set of membership levels is associated with each pixel. However, there are some limitations associated with FCM: the number of clusters must be fixed before clustering or there are lot of computational complexity [70].

Basha and Prasad proposed a method using morphological operators and FCM clustering algorithm for the detection of breast tumor mass segmentation from the mammogram images. Initially this algorithm isolated masses and microcalcification from background tissues using morphological operators and then used FCM algorithm for the intensity-based segmentation. The results indicated that this system could facilitate the doctor to detect breast cancer in the early stage of diagnosis [28]. literally, the potential of FCM clustering for diagnosing diseases like breast cancer has been proved to resolve the uncertainty and unknown noise in medical image segmentation [48].

4.5 what is edge detection method?

The vision processing identifies some features of an image that are relevant to the evaluation of the structure and objects' properties. Edges are one such feature that has a significant role in image analysis and represents the local changes of the image. Edges typically occur on the boundary between two different regions in an image [71]. The significant local change in the image intensity is usually related to intensity dissimilarity in the image or inconsistency in the first derivative of the image intensity. Discontinuities in the image intensity can be step or line discontinuities. In the first one, which is step discontinuity the image intensity suddenly changes from one value on one side of the discontinuity to another value on the opposite side, while in the second one the image intensity abruptly changes value but it returns to the starting value within some short distance. However, step and line edges happen rarely in real images. Indeed, intensity changes are not instantaneous but occur over a finite distance thus, step edges become ramp edges and line edges become roof edges. Since sharp discontinuities rarely exist in real signals due to the low-frequency components or the smoothing related to the sensing devices [72].

Standardly, the border of an object generates the step edges due to the differences between the image intensity of the object with the background intensity of the image. It is important to consider that real images are very noisy but the changes due to noise are not edges, even though the changes are local. The two main definitions of edges are: firstly, an edge detector is an algorithm that produces a set of edges from an image. Secondly, contour is a list of edges or the mathematical curve that models the list of edges [71].

Edge detection is one of the oldest and most basic operations in image processing, which is often used as a basic building block to improve the performance of more complex algorithms [73] and is even known as the first step in recovering information from images. Due to its importance, edge detection continues to be an active research area [71]. The coordinates of an edge point may be the integer row and column indices of the pixel where the edge was detected, or the coordinates of the edge location at subpixel resolution. The edge coordinates are expressed by the coordinate system of the original image, but more likely are in the coordinate system of the image produced by the edge detection filter, since filtering may interpret image coordinates. The edge detector has produced set of edges that are partitioned into two subsets: correct edges and false edges. The first subset corresponds to edges in the scene while the other one is not matched to the edges in the scene. There is also a third set of edges which is represented those edges in the scene that should have been detected. This is the set of missing edges. The false edges are called false positives, and the missing edges are known as false negatives. Edge detection uses local information to decide if a pixel is an edge [71].

The basic idea of edge detection is to detect abrupt changes in the intensity of an image. Diagnostic of those changes can be done by performing first or second-order derivatives. In the 70's, edge detection methods were developed by using small operators (such as Sobel masks), attempting for an approximate evaluation of the first derivative of the image. In 1980 Marr and Hildreth argued that intensity changes are relevant to the image scale, so different sizes of operators are required for edge detection. They also stated that a sudden change in the image intensity in the first derivative will be known as a peak (or valley) and in the second derivative as a zero crossing. Haralick's algorithm is an alternative approach based on the second derivative. This algorithm has proposed a model to approximate the image locally around a point. Thus by using this model, an approximation to the second derivative of the image can be evaluated analytically. Then by making a condition over the model parameters, the edges are achieved [73].

Edge detection is a measure in medical imaging for abnormalities detection and extraction of features. Tumors vary in size, shape, signal intensity, and contrast, hence their classification as benign or malignant requires accurate preservation of their edges and other morphological details. Hence, selection of a sensitive edge detector in medical imaging helps not only in extraction of features for diagnostic detection but also aids in deciding the future therapies and treatment patterns [74].

Although many statistical and several filtering approaches have been used for edge detection, the algorithms based on the Laplacian of Gaussian and based on the gradient of Gaussian were very popular in the 1980s. The Laplacian of Gaussian edge detection scheme is still dominant in models of biological edge detection. Edge detection based on the second directional derivative was presented as well, by Haralick, which incorporated a form of image smoothing based on approximating the image with local surface patches. Despite that many edge detectors have been developed, there is still no well-defined metric to help in selecting the appropriate edge detector for an application. Lack of a performance measure makes the judicious selection of an edge detector for a particular application a hard challenge [71].

In the past two decades' several edge detection techniques or algorithms are proposed, based on that the effective edges are evaluated or analyzed. The primary key behind the edge-detection methods is to restrict false detection in the edges and computational time. In this, the canny optimal detection algorithm aims to discover the optimal edge which reduces the probability of detecting false edges and gives sharp edges [75].

Edges and boundary detection of objects in medical image analysis is a critical issue in numerous applications such as detection of abnormalities, tissue measurement, surgical simulation, and many more. It is also a very helpful task in the recognition of different organs in the human body [76]. There are many edge detection techniques for image segmentation [76] and they began in the sixties, where the first method and simplest calculation is the method of Robert and then emerged many methods with the progress that appeared on technology, including the method of Sobel, Canny, Laplacian, Prewitt, Kirsch, and Robinson [77].

Those techniques basically include two types namely: Gradient (Roberts, Sobel, and Prewitt) and Gaussian. The Gaussian based edge detection is also two types: Laplacian of Gaussian (LoG) edge detection and Canny edge detection. The Gradient method finds the edges by searching for the most value of the first derivative of the image and lowest value of it while the Laplacian method searches for zero crossings in the second-order derivative of the image to find edges [76].

The Roberts edge detection is introduced by Lawrence Roberts in 1965 which requires a simple and quick 2-D spatial gradient. The Sobel and Prewitt edge detection methods were introduced in 1970 and Sobel works on the edges at the highest gradient of points while Prewitt uses the mask to approximate digitally the first derivatives [3] [76]. The LoG detection technique was proposed by Marr in 1982 and it is a second-order derivative that highlights regions of rapid intensity change [76]. The Canny edge detection method was earliest created by John Canny for his Master's thesis at MIT in 1983. Canny edge detection is a multi-stage algorithm with the purpose of finding an extensive range of edges and an adaptive Canny edge-detection method is proposed based on the Canny theory. In adoptive Canny, the 3*3 neighborhood is implemented to compute the gradient instead of the canny algorithm in the 2*2 neighborhood. The Otsu method is also used to attain the high and low thresholds [72] [76] [78].

Kirsch Compass Kernel is another edge detection method and is based on gradient information like Sobel, Prewitt, Roberts. One of the most important edge operators is the Kirsch filter due to its effort on achieving a good balance between preserving edge details and suppressing the noise components. Kirsch operator is a non-linear edge detector that detects the edges with maximum strength in a fixed direction and the main advantage of kirsch filter is the efficiency of detecting maximum edges. This filter takes a single kernel mask and rotates on each element of the image at an angle of 45 degrees (in all eight compass directions) like North (N), North West (NW), South(S), SouthWest (SW), West (W), East (E), South East (SE), and North East (NE) [79] [80]. The maximum value of these filters is then computed as a result of the Kirsch method [77].

Kirsch algorithm is based on the step edge, according to the image properties and the states of Kirsch values, it is possible to adjust threshold values for the achievement of the most possible edge point in the images. When the contrast between the foreground and background is very tough and centralized, the Kirsch algorithm will have very excellent performance. The variables considered in the selection of an edge detection operator include edge orientation, edge structure, and noise environment. Operators can be optimized to look at the vertical, horizontal, and diagonal edges. Hence, the objective is to do the comparison of a variety of directions and analyze the performance [81]. Despite the fact that image edge detection mainly reduces the amount of data and filters out ineffective information, it keeps the major structural features in an image. It is principal to have a good understanding of edge detection algorithms since edge detection methods are implemented to assists in object recognition during the image analysis [72].

The enhancement techniques are mostly developed to improve the quality of medical images for better diagnostic of masses; as an example, Bovis and Singh and Antonie et al used histogram equalization to augment the medical images such as mammograms. However, Schiabel et al used the histogram equalization in parallel by other techniques for quality improvement of mammogram images. While Pisano et al developed the contrast limited adaptive histogram equalization (CLAHE) to figure out the improvement process of mass detection in dense mammograms. Herminger et al compared contrast limited adaptive histogram equalization (CLAHE) and histogram-based intensity windowing (HIW) in order to determine which of them outperforms the other in the detection of simulated masses [82].

As is above-mentioned, image enhancement technology plays a vital role in medical imaging. The main purpose of image quality enhancement in healthcare is to increase the correct diagnostic chance by physicians, which leads to upgrading the effectiveness of treatment. Image improvement techniques are formed on two groups first is the Spatial-domain method that operates directly on pixels, and the next is the Frequency-domain method, which works based on the Fourier transform of an image. The spatial domain method is focused on the **Histogram Equalization** in the image enhancement field. By using histogram adjustment, the intensity distribution of the image will be improved in the whole image. Histogram equalization effectively spreading out the most frequent values of intensity. The histogram filter has advantageous for the bright and dark areas in medical imaging. This method is simple, fast, and has acceptable results for many applications; these are the main advantages of this method. Image enhancement is used to augment the visual aspect of an image. For contrast enhancement, histogram equalization technique is used [83], in other words, to obtain contrast enhancement, intensity values are added by using histogram equalization over different segments. It enhances the contrast of each pixel relative to its local neighborhood. As a result, improved contrast can be produced for all levels in the image [30]. It is also used to preserve the better brightness of the image. In this method, pixel values of the input image are matched to produce a uniform histogram of the image [83] [84].

Chapter 5

Methodology and Implementation

Mostly the problems in people's health can affect the quality of their lives. Therefore, this is a fact that the proper and timely treatment based on a correct and precise diagnostic has a high impression on the improvement of patient's lifestyle or even saving their lives. During the recent decade, development of Machine Learning (ML) algorithms for healthcare has increased rapidly. In fact, ML algorithms can provide good predictive performance on different tasks [85].

Breast cancer is one of the common and hard diseases among all women in the world, which can lead to unpleasant consequences [86] [87]. However, accurate detection of mass lesions in the breast screening at early stages makes treatment possible and even can rescue patient's life. The assist for automated breast cancer detection at an initial stage and the selected procedure of the treatment is the most significant plan in the development of CAD systems.

Due to all the above-mentioned points, this project is focused on applying ML algorithms to mammography images in order to identify any suspicious mass. The goal of this study is to use diverse ML algorithms to create a proper model for breast tumor detection by mammogram image segmentation. The presented method has been divided into four main sections and the whole idea over every step is briefly mentioned within the following lines from 1 to 4. Furthermore, the details, concerning each of the steps, are provided during the next pages.

The four main phases are Data collection, Pre-processing, Image analysis and Tumor detection.

1. **Data collection:** explains collecting data for specific study and analysis. Collection of mammogram images as a dataset is required for this project since it can be used to recognize tumors and aids to lessen the risk of death at an early level.

2. **Pre-processing:** the first step after collecting the desired data is pre-processing. This stage seems essential in order to convert the collected and pure data into a clean dataset with high quality which are used as the input data of the chosen ML algorithm for further analysis. Based on some factors such as data quality, data types and etc, the required operations on data will differ. Fortunately the prepared mammogram images for this thesis work have high-quality. The two following operations are considered to provide the whole dataset for future processes: 1) Data anonymization, 2) Noise reduction.

3. **Mammogram image analysis:** a segmentation technique is implemented to segment breast image and determine the boundaries of the Region of Interest (ROI). Moreover, before the segmentation step, an edge detection method is performed to strengthen edge points in the mammograms and this leads to higher accuracy and performance in segregation of mammogram image and specifically on breast tissue and masses. Then by removing the pectoral muscle in the image, only breast tissue will be left which result in reduction of computational cost. For the segmentation, the Fuzzy C-mean(FCM) clustering method is chosen.

4. **Tumor detection:** in the last step, all objects in clusters are labeled then by the thresholding method the objects suspected of the tumor will be detected. It is important to estimate an appropriate threshold value to decrease false-positive results.

Overall in this research, a set of mammogram images is considered as the data resource for the algorithm. However, one of the first important issues prior to process medical images is the anonymization to protect patient privacy. Hence, in this thesis a mask is created on the image. Morphological operations are performed on images to remove noise and improve image quality for better analyses. The selected model for image segmentation and features extraction is a clustering-based approach based on the standard FCM clustering algorithm. Additionally, a powerful and precise edge detection technique is known as the Kirsch operator are implemented to increase the accuracy of segmentation process while Histogram equalization is employed to enhance the contrast of the images which are resulted from the edge-detection stage. After obtaining ROIs that is the outcome of the segmentation phase, the final objective is the identifying any suspicious mass in ROIs according to the thresholding technique. The figure 5.1 represents a diagram of the methodology steps;

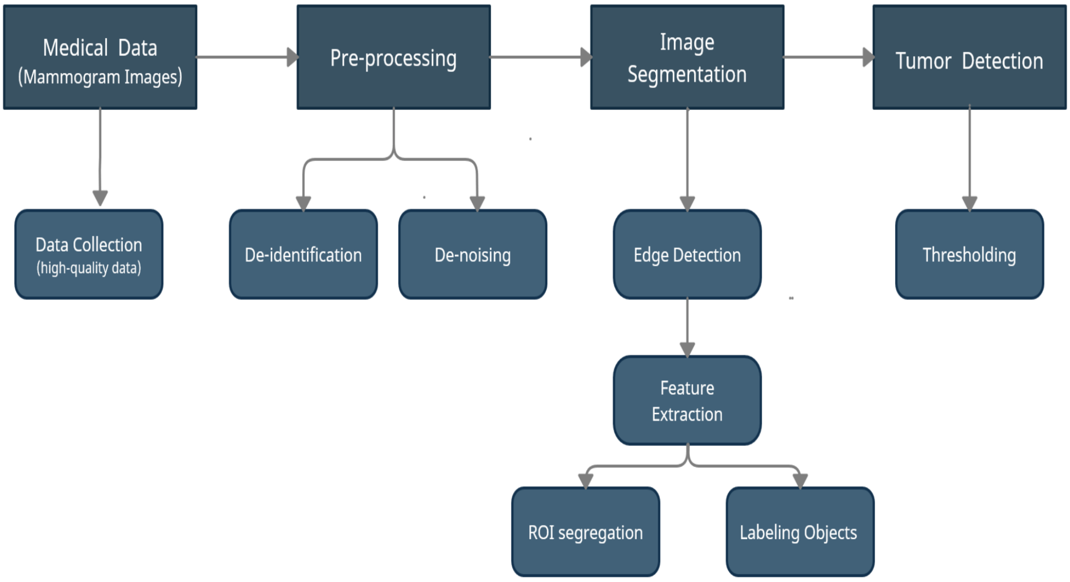


Figure 5.1: *Methodology Diagram*

In this thesis, a combination of different techniques are performed to evaluate data step-by-step and detect breast cancer.

5.1 Data Collection

Clinical data acquisition can be accomplished through different methods such as electronic medical records, voice records, paper forms. The collected data quality depends on the type and quality of equipment and it affects data analysis and the accuracy of experimental results. In order to obtain digital mammography, a stationary x-ray machine that includes two plates captures the x-ray images of the breast tissue at lower doses. Indeed, the breast compresses or flattens between the two plates to spread the tissue apart and allow to take a more complete image in terms of tissue visibility.

All of the mammograms used in this project were obtained from a database of digital screening mammography, which is provided by the radiologists and two experts labeled all achieved images. The purpose of collecting this dataset is to provide a set of mammogram images to be analyzed by the chosen ML algorithms. The dataset contains 260 cases that consist of both normal and cancerous cases (150 normal cases and 110 cancerous cases).

5.2 Pre-processing

When data is collected through digital mammography, the quality of images obtained from medical devices determines the result of processing. In some cases, obtained images have noticeable noise caused by technological features of devices' operation. Pre-processing is applied to remove unwanted information in order to get an enhanced and much clear image for the efficient use of the data set. Also, the de-identification of medical data, which is an important task, must be done in this stage.

5.2.1 De-identification

In this project, first, the data anonymization is performed on the mammogram image by creating a mask. The mask assists to remove the patients' identity or any other information on the mammogram images of the dataset. In order to remove the identification of the patient, each color domain on the RGB mammogram image is extracted. Then by comparing the Red and Blue pixels, those of them which have not the same intensity is changed to the zero value (they are removed).

5.2.2 De-noising

Then RGB color image needs to be converted into the grayscale image to reduce the image complexity followed by image segmentation and feature extraction. Pre-processing has an important role in low-level image processing. It is possible to filter out the noise present in an image using filtering. Morphological filters are mathematical operations and they are applied to the binary image, therefore a threshold value(0.03) is calculated for the binarization process to apply morphological filters and remove noise. Although, this dataset includes high-quality images these filters are very effective in de-noising binary images [86]. In this process, irrelevant image contents are eliminated selectively. This is implemented by the Morphological operations including dilate, erode, close, and open. The Morphological operations are based on the relationships between two sets: the input image, and the processing operator. These operators that are the structuring elements have usually a much smaller size than the input image. Obviously, results in the output image can be varied based on the shape and size of the structuring element [86] [88].

Morphological filtering profit is filling small holes, eliminating small projections to smooth object outlines, and simplify segmented images. Primary operations are dilation and erosion. These operations use a structuring element that determines exactly how an object will be dilated or eroded. Dilation process expanding image objects by changing pixels with a value of "0" to "1". On the other hand, the erosion process shrinking binary objects by changing pixels with a value of "1" to "0". Erosion operation reduces objects' size and eliminates small anomalies by subtracting objects with a radius smaller than the structuring element. In grayscale images, erosion lessens the brightness and therefore the size of bright objects on a dark backdrop. Erosion is denoted by:

$$I \ominus SE(x, y) = max[I(x + i, y + j) - SE(i, j)] \tag{5.1}$$

Where $I(x, y)$ is a gray-scale image, and $SE(i, j)$ is the structuring element.

Opposite of Erosion, dilation normally expands objects' size by filling holes and discontinuous regions, and linking areas that are separated by smaller regions with respect to the size of the structuring element. In grayscale images, dilation grows the brightness of objects. Dilation is denoted by:

$$I \oplus SE(x, y) = max[I(x - i, y - j) + SE(i, j)] \tag{5.2}$$

Where $I(x, y)$ is a gray-scale image, and $SE(i, j)$ is the structuring element.

When the two above morphological operations are finished hole filling as another operation that is widely used in medical image processing is required. Most of the medical image processing operations generate a binary form based on the original image at different steps. Generally, the binary images are created by an ordinary segmentation technique such as thresholding. They include foreground objects restricted by backdrop areas. Sometimes a set of backdrop areas lie within the foreground regions in the binary image. It is known as a hole within the foreground objects. Hence, a hole is a dark region (in pixels) restricted by light pixels.

5.3 Mammogram image analysis

Image analysis had significant evolution in the last two to three decades and obviously, any progress in medical imaging or medical image processing will lead to an improvement in healthcare such as disease classifying or disease diagnostic. ML algorithms are helping to tumor identification, disease classification, and patterns recognition in medical images.

5.3.1 Kirsch Compass Kernel edge-detection

Before applying a clustering algorithm to classify mammogram images into different clusters (based on the pixel's intensity) to attain the ROI which is breast tissue area (where normally cancer cells start growing), a contrast enhancement approach is performed. Contrast enhancement is a procedure that helps the image features represent more clearly by making optimal use of the colors available on the display or output device. Contrast manipulations mean changing the values' range in the image for contrast increment. In this thesis work, the contrast enhancement approach is performed using a non-linear edge detection method. Boundaries include lines in an image that have a very considerable effect on high-level pictorial tasks such as object recognition and scene understanding. Detecting boundaries have been a vital problem since the commencement of computer vision [89] [90] [91].

Edge detection significantly reduces the amount of data and takes out useless information, while maintaining the main structural properties in an image. Since edges has a crucial function in an image, edge detection is used to assist for better detection of objects in image processing and recognize the boundaries of similar areas in the image based on some attributes such as intensity and texture [88] [89] [90] [91].

In other words, the purpose of this work is to achieve more accurate boundaries surrounded by breast tissue. The edge-detection method proposed by Kirsch finds a maximum edge in a few pre-determined directions. The operator takes a single kernel mask and rotates in 45-degree increments through the 8 compass directions (N, S, E, W, NW, SW, NE, SE). Each pixel of images uses these 8 masks, and each masking has a great response to a certain edge direction, the maximal value of all 8 directions is set as the output value of this point. The masking sequence number of the greatest response forms the code of edge direction [89] [90] [91].

$$M0 = \begin{bmatrix} 5 & 5 & 5 \\ -3 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix} \quad M1 = \begin{bmatrix} 5 & 5 & -3 \\ 5 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix} \quad M2 = \begin{bmatrix} 5 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & -3 & -3 \end{bmatrix} \quad M3 = \begin{bmatrix} -3 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & 5 & -3 \end{bmatrix}$$

$$M4 = \begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & -3 \\ 5 & 5 & 5 \end{bmatrix} \quad M5 = \begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & 5 \\ -3 & 5 & 5 \end{bmatrix} \quad M6 = \begin{bmatrix} -3 & -3 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & 5 \end{bmatrix} \quad M7 = \begin{bmatrix} -3 & 5 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & -3 \end{bmatrix}$$

Figure 5.2: *The eight templates of the Kirsch operator [89]*

When the edges are identified, the previous templates are multiplied by the 3x3 region of the image and pick out a template with the greatest output value. Then take this maximal value as the edge intensity of the central pixel, and the maximal value template as its edge direction.

Supposed P as a point in an image and set $q(k = 0{,}1, ...7)$ as edge intensity of the kth direction. when an image is dealt with the kth template of Kirsch operator, then the edge intensity of $P(i,j)$ is $S(i,j)maxq(k = 0{,}1, ...7)$ and the corresponding edge direction $kD(i,j) = \{k/q \quad is \quad maximum\}$.

$P0$	$P1$	$P2$
$P7$	$P(i,j)$	$P3$
$P6$	$P5$	$P4$

Figure 5.3: $P(i,j)$ *and its grayscale of eight neighborhood unions [90]*

After extracting the edges and before adding these edges to the image, a Histogram Equalization technique is employed to enhance the quality of the final image from extracted edges. Histogram equalization is one of the most important techniques for the image quality augmentation and is widely used for image contrast enhancement in a variety of applications due to its simple function and effectiveness while medical images is one example. Histogram filter affects the operation of some useful edge detection methods. The performance of many histogram-based enhancement techniques on the X-ray images validates the effectiveness of this enhancement method [92].

By adding Histogram equalizer to the result of Kirsch operator, the gray level range in the image will be increased. The goal is visibility improvement of the obtained borders in breast tissue.

5.3.2 Mammogram image segmentation using FCM algorithm

The purpose of image segmentation is to partition images into different regions, based on given criteria for future processing. Image segmentation plays an important role in medical applications such as abnormality detection, quantitative analysis, and postsurgical assessment.

Image segmentation methods are mostly done based on the pixels' characteristics of the image. In fact, it can separate the foreground of the image from the backdrop, or cluster regions of pixels based on similarities in color or shape. For example, segmentation in medical imaging is mostly used to detect and label pixels in an image that represent a tumor in the intended part of the body. An image analysis algorithm for mammograms is offered to enable radiologists in finding out dense regions and analyzing masses, and could effectively split and enhance the intended image to stand out any unsure dense area. The algorithm behind these works is the Fuzzy C-Means(FCM) clustering. Fuzzy clustering as a soft segmentation method has been widely studied and successfully performed on image segmentation. The FCM clustering algorithm is one of the most popular algorithms among the fuzzy clustering techniques, and it is used in image segmentation because of robust characteristics for ambiguity. It is possible to preserve much more information by using the FCM algorithm in comparison to hard segmentation methods [93]. FCM is a formalized clustering technique that has been used for feature analysis and classification. It plays an important role in the field of pattern identification. Many fuzzy clustering methods are stated recently and most of them are based on distance criteria depending on distance estimation among pixels and the cluster centers. The presented algorithm works in pixel intensity and concentrates on calculating the centers that minimize the dissimilarity inside of each cluster [32].

To identify the tumors in available mammogram datasets, the proposed approach requires the segmentation of breast image. As this proved to be an important issue, efforts were made to develop an automatic segmentation. ROI (breast tissue with highest density and apart from underarm area) is obtained by grouping pixels using a modified clustering method based on image histogram on the result of edge detection. Therefore, the input for the FCM algorithm is a set of images with high resolution which making the separation of breast tissue from the non-breast area easier [64].

In this thesis work, segmentation of breast images has been considered the most important middle step of image analysis to extract the concept behind the pixels and detect homogeneous groups of pixels that have similar intensity values. The similarity is calculated according to the distance between the pixels and the prototypes of regions, and each pixel is labeled to the nearest group or most similar prototype. This procedure considers all data to different groups, even if basically some pixels are not very representative of any group [86] [32] [57]. Basically image segmentation is classified into three different groups: region-based, edge-based, and pixel-based. There are two main kinds of classification in all these categories which are supervised and unsupervised and in addition to that, there are only two types of clustering in pixel-based technique which are hard (data items belong to only one cluster) and fuzzy clustering (data belongs to the degree of membership to each cluster)[57].

The process of identifying multiple groups of an image is known as clustering.

Data clustering is about dividing data elements into several classes or clusters thus the items in the same class are as matched as possible, and items in different classes are as dissimilar as possible. Depending on the data characteristic and the clustering purpose, different measures of similarity may be used to place items into groups, since the amount of similarity controls how the clusters are constituted. Some measures that can be used in clustering include distance, connectivity, and intensity. Fuzzy clustering is a process of assigning these membership levels and then using them to assign data elements to one or more clusters [57].

The FCM algorithm is an iterative function which is intended to divide a finite set of elements $X = x_1, x_2, ..., x_n$ into a collection of c fuzzy clusters with respect to some given scales. The algorithm returns a list of cluster centers Vi while $i = 1, 2, ..., c$, and a partition matrix U where $U = U_{ij}$ and $i = 1, 2, ..., c$, $j = 1, 2, ..., n$. However, U_{ij} is a numerical value in $[0, 1]$ that tells the degree to which the element X_j belongs to the $i - th$ cluster. As a simple description of Fuzzy Logic the following steps are stated:

- Step 1: Select the number of clusters c and initial partition matrix.
- Step 2: Calculate the fuzzy cluster centers $i = 1, 2, ..., c$ based on the initial partition matrix.
- Step 3: Calculate the new partition matrix by using $i = 1, 2, ..., c$.
- Step 4: Calculate the new partition matrix then set $l = l + 1$ and go to step 2. If the number of clusters $> c$, then stop.

After mammogram segmentation is done and ROI is obtained then it is required to extract all objects and their features in each cluster. Regionprops is a function in MATLAB that is utilized in this project to extract the property of objects in the image. There are different properties such as Area, Centroid, Boundingbox, Eccentricity, etc. However, here Area is extracted as the main object feature which will be used for the final step, tumor detection [94].

Image region property analysis:

- Step 1: Read the image file.
- Step 2: Convert the RGB image into a binary image.
- Step 3: Count and label all objects inside ROI and even in all other created clusters, then Calculate the needed property for each labeled object by using regionprops function.

- Step 4: Save all objects and its area in a matrix.

The next step is finding any suspicious lesion in extracted ROI, due to the fact that this region includes the most tissue density. Therefore, only one cluster that is much similar to the breast in terms of the size criterion is selected as ROI.

5.4 Tumor Detection

A tumor is comprised of abnormally growing regions that are dangerous for human survival. Therefore, early-stage tumor detection is useful for the increase of survival rate although it is challenging because of numerous powerful image factors including poor contrast, complex background, brightness issues, the shape of infected region, and fuzzy borders.

The last goal of this project is about detecting any kind of masses or abnormalities in mammogram images. Consequently, after ROI extraction in mammograms all the objects in that specific cluster obtaining from segmentation stage are counted and labeled, and these information are saved in a matric, the area related to each object is specified and added to the matric, as well (It is necessary to mention that although the search area should be the extracted ROI, in this thesis all the objects in other clusters are also labeled and their areas are determined in order to have a more accurate evaluation). To detect a tumor or in other words a suspicious object the thresholding method is used, which seems a simple but high-risk method due to this fact that considering a wrong value as the threshold can change the result completely. However, the appropriate value of the threshold depends on the whole dataset and in this presented work this value is estimated as 0.75, then based on this threshold and the comparison with the area of each object, any type of masses in the breast will be discovered.

Chapter 6

Evaluation

Evaluating the built model efficiency is a critical part of the machine learning structure, due to its effect on performance improvement [95]. This is a fact that an algorithm can provide a good result on a particular dataset or a class while performing poorly on the other group of data or other class, thus there is not a specific way to select the best algorithm. However, there is the possibility to evaluate the chosen algorithm's performance [96]. Evaluation metrics are able to compare the results of different models. The main feature of evaluation methods is to specify the classification performance, particularly for the final result assessment of the algorithm and the optimization of the algorithm by changing parameters [95].

There are several methods that can be used for evaluating the performance of our model in breast tumor detection such as confusion matrix, and cross-validation. However, the confusion matrix is chosen for efficiency assessment. The confusion matrix is a particular table in order to measure the ML algorithms' execution. Indeed, it is an N * N matrix, where N states the number of predicted classes.

In this project, a 2*2 matrix is considered to indicate the testing results of the made model according to the chosen ML algorithms and demonstrate four values composed of the estimated values of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) [95].

- **True Positives:** The mammograms that actually have breast tumor and correctly identified in our model.
- **False Positives:** The mammogram does not include tumor but the algorithm detect it.
- **False Negatives:** The mammogram includes tumor but the algorithm can

not detect it.

- **True Negatives:** The mammogram images that actually did not have breast tumor and correctly identified in our model.

After testing the final model overall images in the dataset the above-mentioned values as metric values are achieved. The numbers on the diagonal (TP and TN) represent that how many times the algorithm worked properly and made the correct diagnosis. However, the values which are not on the diagonal (FP and FN) indicate the number of instances that the model messed up in rightly detecting normal and abnormal cases.

Confusion Matrix		Reality	
		Presence of Tumor	Absence of Tumor
Predicted	Presence Tumor	TP=95	FP=23
	Absence of Tumor	FN=15	TN=127

Table 6.1: *Confusion Matrix*

Sensitivity means the proportion of mammogram images which consists of tumor (patient with positive index test) among all the images that have the tumor. While Specificity states the proportion of all images without tumor (patient with negative index test) among the images that not includes tumor.

According to the table 6.2, which represents some formulas the sensitivity and specificity values are computed.

Evaluation Type	Formula	Estimated values
Sensitivity	$TP/(TP+FN)$	86.36%
Specificity	$TN/(TN+FP)$	84.66%
Positive predicted value	$TP/(TP+FP)$	80.50%
Negative predicted value	$TN/(TN+FN)$	89.43%

Table 6.2: *The Metric evaluation values*

Results:
Image processing techniques plays important role in the enhancement of treatment quality. Image segmentation and tumor detection in the medical field help doctors in a more precise and fast diagnosis. However, this is not an easy task especially due to the noise in images. In general terms, noise is an inherent property of medical imaging. The disease diagnostic will face difficulty due to this

disturbance of medical imaging. Therefore, medical images need noise removal before any further processing and analysis [97].

This project presents mammograms segmentation and tumor detection based on FCM clustering and an efficient edge detection technique is additionally applied to improve the performance of algorithm.

6.1 Preprocessing

Medical images have an important role in fast diagnosing disease and monitoring the effect of the selected treatments. In order to use medical records in biomedical research and scientific studies, the anonymization of health information in medical data is an essential method for protecting patient privacy. In this project, de-identification means removing patient data in mammogram images. On the other hand, in spite of the increasing progress in the methods of capturing these images, the produced images may not pose enough quality for an accurate interpretation. Environmental noises, patients' special conditions in photography, lighting conditions, and technical constraints related to the imaging devices are the main reasons that may produce low-quality images.

In this kind of situation and specifically, when reimaging is impossible, image enhancement techniques are effective. The new techniques are used to repair the damaged images and to enhance their quality and contrast. Image noise, which is a random variation of brightness or color information in images, can affect medical imaging in terms of difficulties in diagnostic characterization or object size or even in the definition of detected object boundaries. Noise removal algorithm means the reduction or elimination of noise from images. As the first step, preprocessing is performed on the input images(dataset) for two particular reasons: **De-identification and De-noising**.

6.1.1 Data Anonymization results

Patient data needs properly de-identified in order to be publicly available for researchers for further analysis. Imaging data are especially at risk because if they contain protected health information (PHI) which releases through a public repository, one of the first people that could be held liable is the original researcher who collected and submitted the data [98].

In order to prepare clean data as the input of the algorithm, a technique for data anonymization is done and the patients' information is removed from all mammogram images.

(a) (b)

Figure 6.1: *(a) Original Image (b) After De-identification*

As it is evident in the Figure 6.1, the patient data is removed, by creating a mask on the input image and the resulting image is used for further processing without any risk of identifying.

6.1.2 De-noising results

The chosen method for noise reduction can vary depending on the type and quality of the images. In this thesis, the original images are de-noised by some mathematical morphological operations. In fact, morphology operators, through increasing and decreasing white color in different parts of an image, have an important role in improving the quality of the image in order to process or in detect various existing objects in the image. Since each organ in the body has a different structure or texture, it is possible to make some specific objects eminent in the image by using the morphological filter, hence a more appropriate image will be obtained for future analysis.

In the beginning, the shape and size of the desired mask are specified for morphological transforms. Obtaining the optimal result and reducing computation time in morphology-based techniques are dependent on the shape and size of a mask. Hence, the chosen mask which is selected arbitrarily should have an appropriate shape and size. The disk-shaped mask is independent of changes in rotation, thus it is commonly used in medical imaging in comparison to other types of mask. The mask's size also depends on the input image and can take different values for different images. According to the all above mentioned points, a disk-shaped mask is used to apply morphology transforms whose initial size is determined through trial and error and based on the input image. In this work, **Dilate**, **Erode**, and **Open** are the performed operations to acquire new images for the processing stage.

Before applying morphology operations, it is essential to convert the image into binary form according to the considered threshold. In general, the medical images in binary form are generating with simple segmentation techniques such as thresholding. The binary form of medical images contains foreground objects surrounded by backdrop regions [99]. The Figure 6.2 shows the achieved binary image from input image by considering 0.03 an appropriate value between 0 and 1as threshold for binarization.

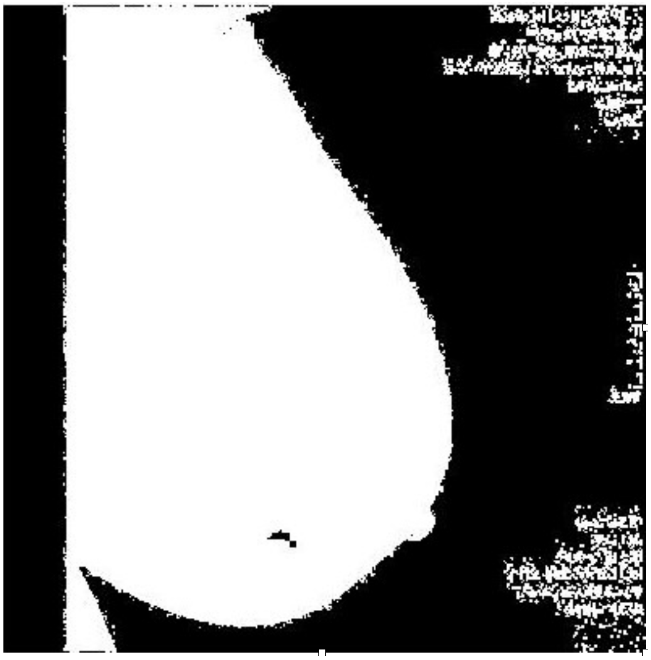


Figure 6.2: *Binarized Image*

Sometimes a set of background regions be placed into the foreground regions which happens due to weakness in selecting the optimal threshold for binary conversion. They are recognized as holes in the foreground objects. Therefore, a hole is an area of dark pixels restricted by light pixels. Although there are diverse types of hole-filling methods, morphological operations are used to achieve this purpose in this project. This method measures the area of each dark/white region in pixels and if the area of the dark region is less than the threshold value, then they will be considered as holes. Hence, the dark holes will be filled (removed) by changing their intensity into white color [99].

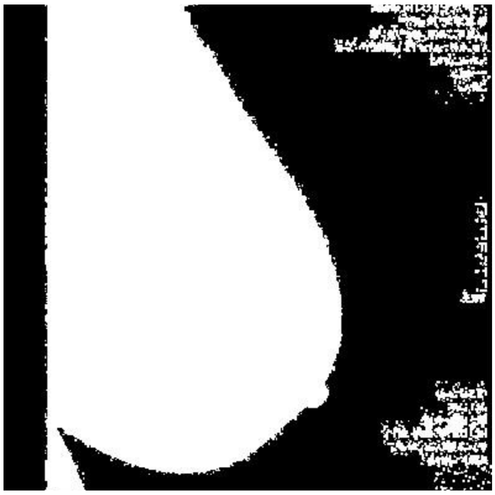


Figure 6.3: *Hole-filling result*

The Figure 6.3 displays the result of filling holes in a binary form of the mammogram image. Now the two other mentioned morphological operations have to be applied to the binary image in order to improve its quality and prepare it for the processing stage. Basically, methods of mathematical morphology act based on the structural properties of objects, and two sets of data are the input of Morphological operators. The first set contains the input image (the binary image) and the second one describes the structural element (mask). Hence, a disk-shaped mask with a radius of 1 is defined (The mask is a matrix containing zero and one values and different formats can be selected to form a mask).

Figure 6.4: *After Image Dilation*

In fact, after applying the morphological operators on the binary image, a new value for each pixel is obtained by sliding the mask on the input image. Value 1 in each mask indicates efficacy and value 0 indicates ineffectiveness in the final image. The Figure 6.4 illustrates image result in from dilation operation.

As we mentioned during previous chapters, **Literature Survey and Methodology**, Dilation is the operation of “lengthening” or “thickening” in the binary image mainly used to fill the spaces. Hence, the image after dilation will be increased in intensity. Moreover, this issue is clear in the image of the dilation result [100] [101]. The erosion process is as same as dilation, but acts in reverse the pixels are converted to ’black’, not ’white’. This operator reduces the size of objects and removes bright areas under the mask in order to make the final image looks darker than the original image. Therefore, the two main inputs for the erosion operator are the image which is to be eroded and a structuring element which was explained earlier [100] [101]. Figure 6.5 displays result of erosion operation. In addition, the combination of the two above operations (dilation and erosion) together, will produce an opening operation while erosion followed by a dilation using the same structuring element for both operations.

Figure 6.5: *After Image Erosion*

BWareaopen is an additional basic operation that morphologically open binary image (remove small objects). It removes all connected components (objects) that have fewer than P pixels(in this thesis work P=700) from a binary image and producing another binary image, BW2 [101]. Figure 6.6 shows the output of the image after performing areaopen function.

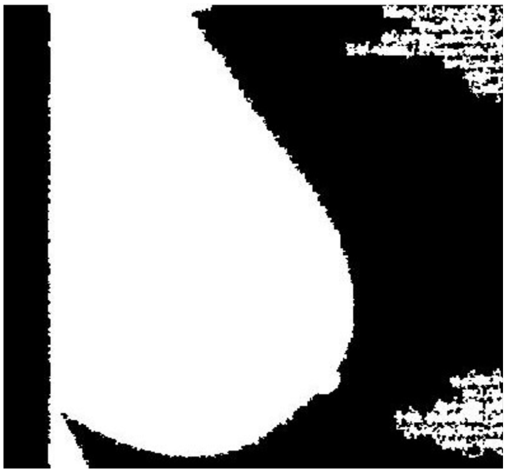


Figure 6.6: *After Removing Small Objects*

6.1.3 Conclusion of Pre-processing step

As stated at the beginning of pre-processing phase, there are medical images with poor contrast, which are difficult for analyses and information acquisition. Therefore, we should employ proper methods to enhance the quality of such images. The method applying in this thesis work is based on mathematical morphology, which attempted to enhance the image contrast and repair the damaged parts. In addition to its simplicity, this approach has a high potentiality to process poor-quality image [102].

6.2 Processing and Analysis

Mammogram images are segmented by applying FCM clustering algorithm in order to separate breast tissue from other touching organs and extract ROI from the mammogram images. Moreover, a contrast enhancement approach, which is an edge detection method, is performed to improve quality of segmentation and clustering.

6.2.1 Edge-detection (Kirsch Compass Kernel algorithm)

The edge detection technique plays an important role in image processing. It acts as a supporting tool in order to improve segmentation, enhance the identification of abnormal masses in mammography images, and subsequently reduce the false positive and false negative values. The Kirsch Compass Kernel module is a non-linear edge detector that takes a single kernel mask and rotates it in 8 directions and the maximum value over all the eight orientations is the output value for the edge magnitude image [79].

In this project, the edge detection procedure identifies sharp edges of objects in images and aids in detection of clustered with precision in digitized mammograms. The result of Kirsch module in 8 directions based on the original image for one mammogram of dataset showed in Figures 6.7, 6.8, 6.9, 6.10, 6.11, 6.12, 6.13, 6.14.

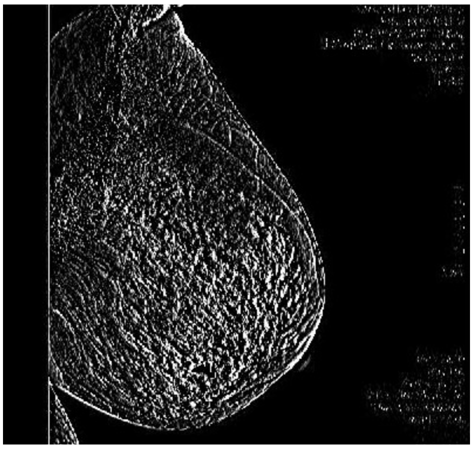


Figure 6.7: *Kirsch result in first direction*

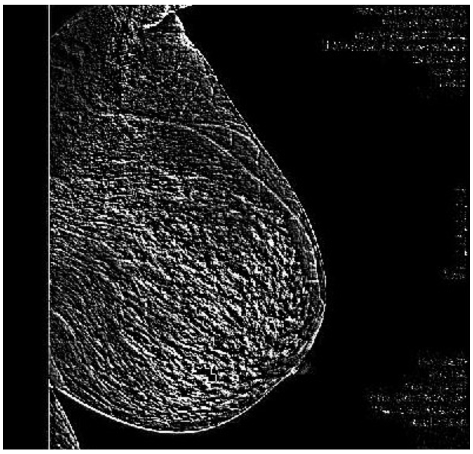


Figure 6.8: *Kirsch result in second direction*

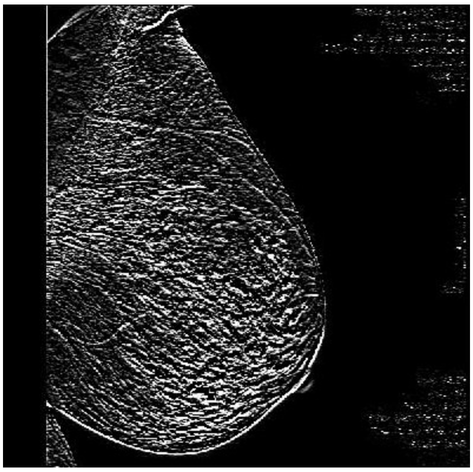


Figure 6.9: *Kirsch result in third direction*

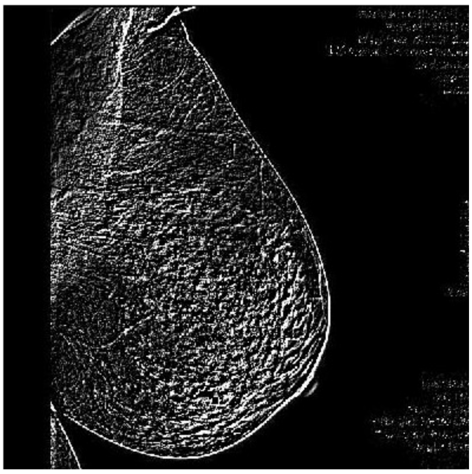


Figure 6.10: *Kirsch result in fourth direction*

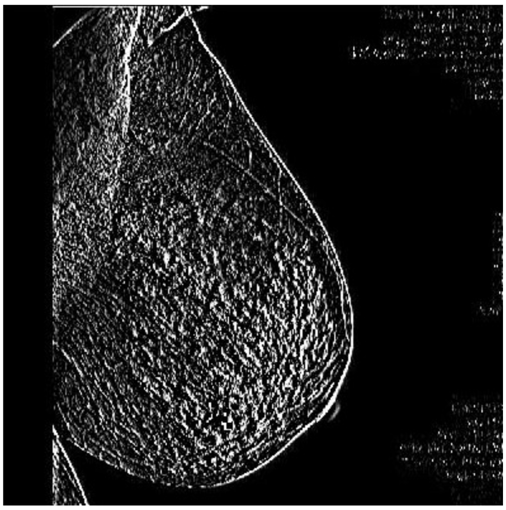


Figure 6.11: *Kirsch result in fifth direction*

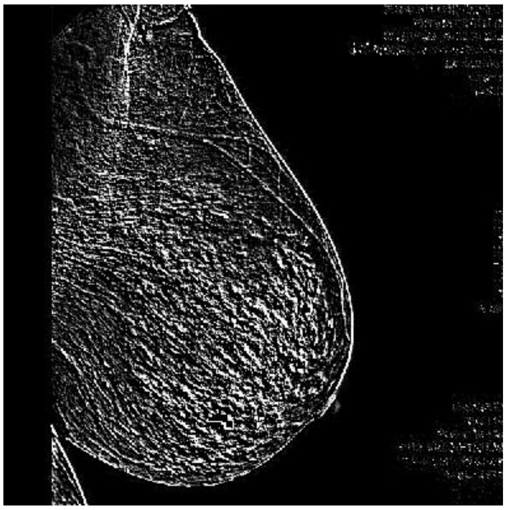


Figure 6.12: *Kirsch result in sixth direction*

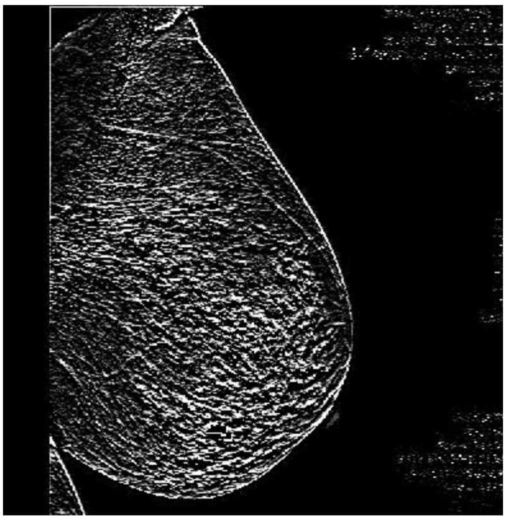


Figure 6.13: *Kirsch result in seventh direction*

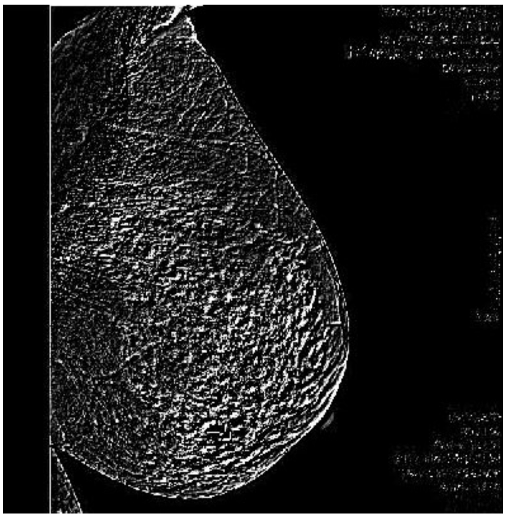


Figure 6.14: *Kirsch result in eighth direction*

By performing the edge detection method on the mammograms, the differences in pixels' intensity of image points and boundaries of different tissues are more explicit. As it is obvious from Figures 6.7, 6.8, 6.9, 6.10, 6.11, 6.12, 6.13, 6.14, the Kirsch algorithm will have very outstanding performance on boundary extraction and it can adjust threshold values to obtain the most possible edge point of images.

Figure 6.15 is shown the final image of extracting edges of 8 directions which are added together. However, it has low-intensity contrast. Therefore, the result of the Kirsch operator is added to the original image and it is illustrated in Figure 6.16 and before applying the clustering algorithm to images, an enhancement technique is performed to improve the image contrast.

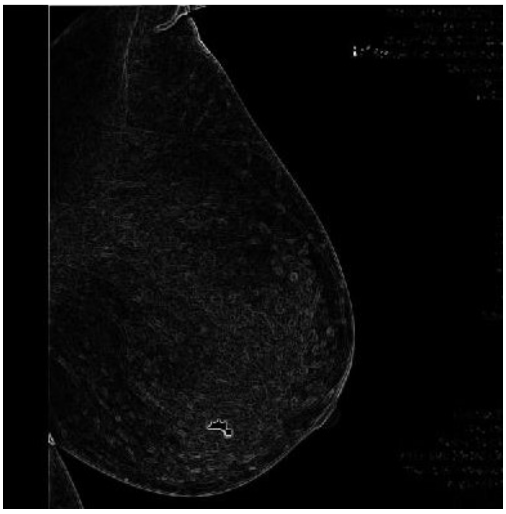


Figure 6.15: *The final Kirsch result*

Figure 6.16: *Result of edge detection output plus original image*

Histogram Equalization is an image processing technique employed for image quality augmentation [103]. It achieves contrast improvement by spreading the more frequent pixel values over those pixel values that have a smaller number of occurrences [103]. After applying this technique, the histogram image should be more visible than the input image. The result of histogram equalizer is specified in Figure 6.17.

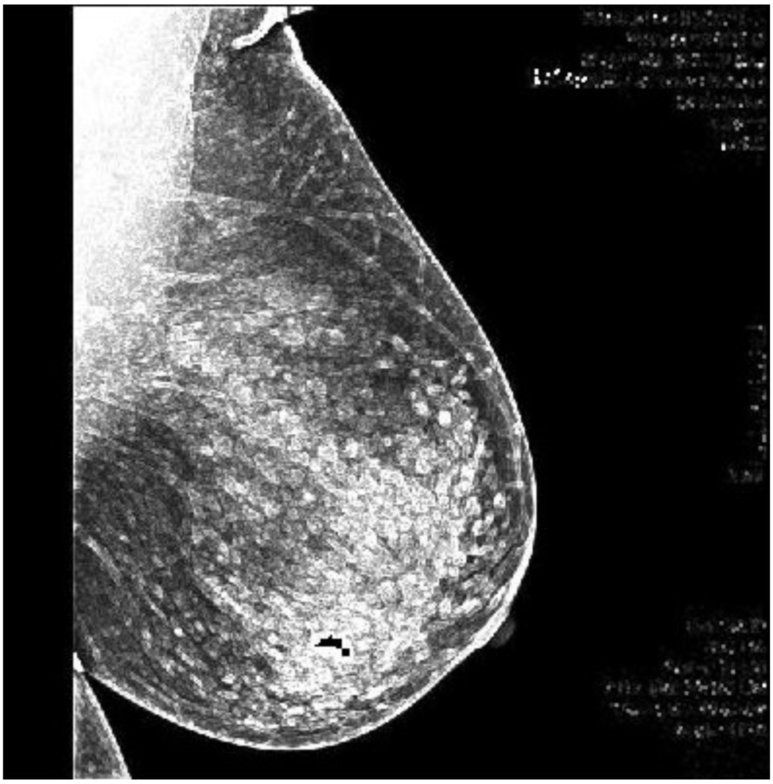


Figure 6.17: *Result of applying Histogram equalization*

6.2.2 Image Segmentation method (Clustering Algorithm)

In the proposed thesis, Fuzzy C-Means clustering as one of the most common algorithms for medical image segmentation [104] is implemented to cluster the mammogram images based on pixels' intensity values. According to the Trial-and-Test method, different numbers of clusters are considered for performing clustering and the best choice by considering the whole images in the dataset plus the enhancement technique which was applied on the mammograms seems to be 3. The Figure 6.18 shows the three clusters.

The result of clusters represents that each pixel has the potential to belong to multiple clusters. As it is specified in the result of the clustering section, the mammogram images are segmented into three images include breast tissue, background and underarm area, and to improve the visibility of each segmented image, a specific function is employed to fill the holes in each cluster. Moreover, a morphological operation is applied to remove noise from the clustered images.

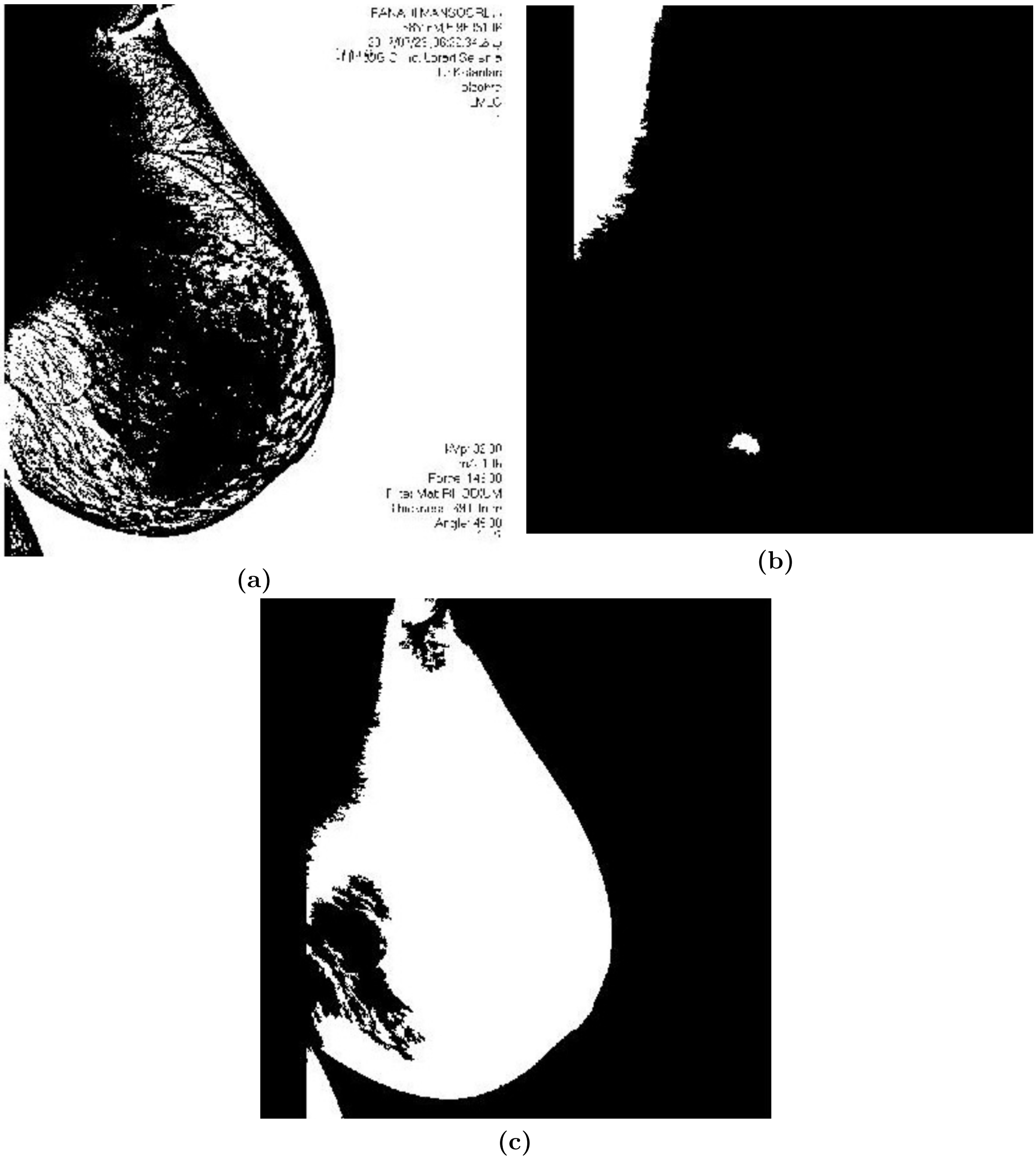


Figure 6.18: *(a) The first cluster (b) The second cluster (c) The third cluster*

Figure 6.19 displays the combination of all clusters in one image. Generally, the image segmentation technique is frequently used to find objects and boundaries in pictures. In the most cases it leads to the process of specifying a label to each pixel in a segmented image such that pixels with the same label share certain characteristics [105].

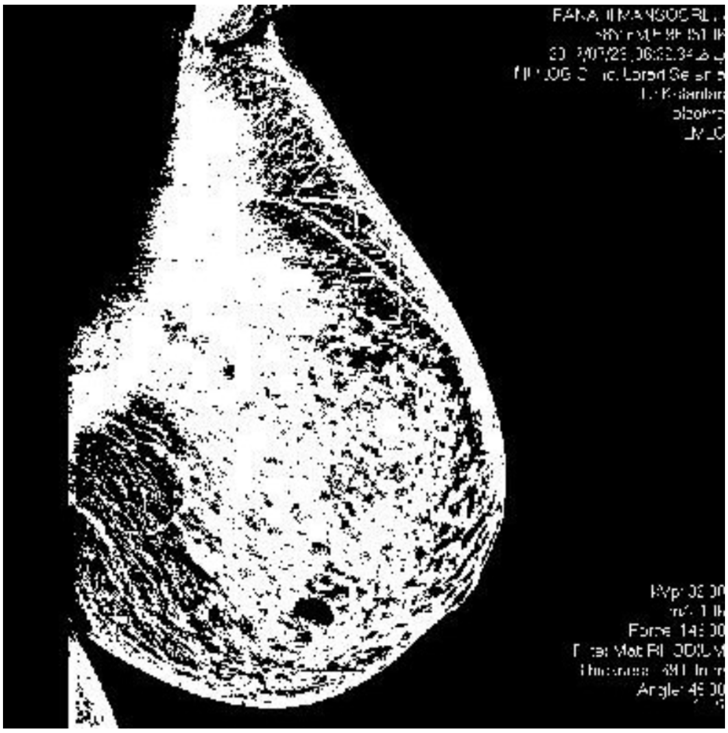


Figure 6.19: *All clusters*

A wide range of segmentation techniques is developed by researchers to analyze the objects' properties and localize tumors in medical images [106]. The goal of segmentation is to figure out the high probability regions in terms of suspicious or abnormal mass in a mammogram [49], and in this case segmentation phase is used to separate breast tissue from other touching organs.

6.3 Feature Extraction(ROIs Segregation and Labeling Objects)

After using the FCM algorithm for mammograms division and isolating desired areas then it is necessary to study on texture features or objects' properties in the ROIs. Indeed, texture feature is useful in differentiating normal and abnormal patterns. The texture is shown by color range variation of surface on the medical image. Extricating features from mammographic images is one of the ideal approaches to distinguish breast cancer.

In this section, the binary images are analyzed to extract the required features from each object in ROI and even other segmented clusters by using MATLAB software. In the presented method, the area is extracted for each object in a cluster, while all the other features such as centroid, eccentricity, bounding box, solidity, pixel list, could be used depending on the desired target. The area as a feature of the mass represents the actual number of pixels in the extracted area that specifies the area of the cancer mass. This feature is a scalar number [107].

6.3.1 ROI Segregation

In fact, the main reason to extract the ROIs is further investigation of their content and composition. One of the main objectives of this project is the separation of breast tissue from other touching organs around it (underarm area). Since breast cancer most commonly develops in cells from the lining of milk ducts and based on the whole mammograms in the dataset, the breast tissue is considered as ROI. Therefore, it is possible to limit the search area and make objects' analysis easier to detect suspicious objects or tumors, which is the final goal of this work.

The two next Figures 6.20, 6.21 illustrates the ROI extraction (breast tissue) which is the main objective of segmentation stage. As is clear, Figure 6.21 is the result of segmentation after multiplying the binary image with the gray level image.

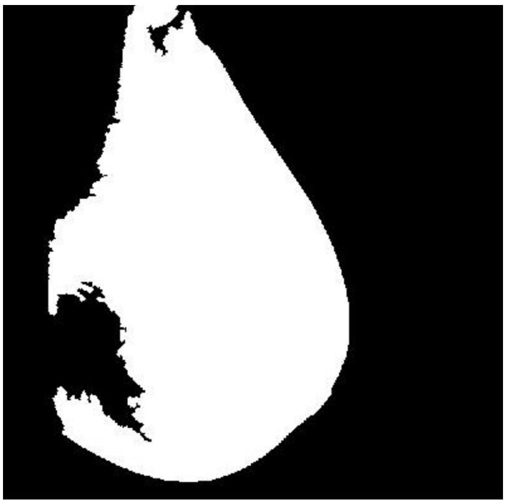


Figure 6.20: *Binary image of breast segmentation result*

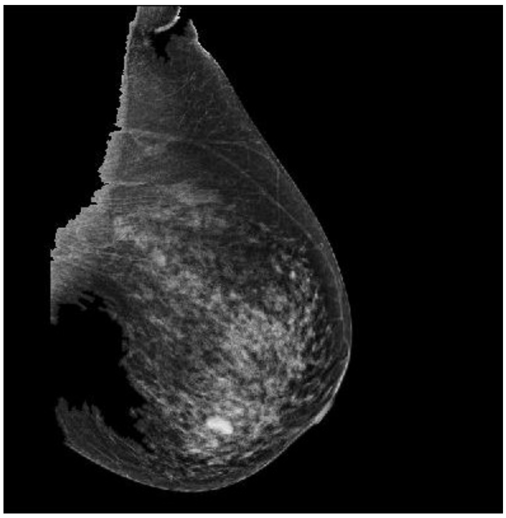


Figure 6.21: *Breast Segmentation result*

6.3.2 Labeling Objects

Labeling of connected components in a binary form of medical images is one of the most fundamental operations since it is necessary to recognize objects in an image [108]. After labeling, a binary image will be known as a labeled image. Thus, it is possible to extract each object in the (labeled) image by its label, and then calculate its shape features such as area, perimeter, circularity, centroid, etc [109].

A simple and efficient procedure is applied in this thesis to assign labels to every object in binary images and extract the useful information of the objects for further analysis. As mentioned in the previous section and although in this dataset, the ROI is the main search region for detecting tumors however in order to check all clusters for any suspicious object the labeling method is performed for all clusters.

1x3 double

	1	2	3
1	1	1	360000
2			

Figure 6.22: *Objects in first cluster*

3x3 double

	1	2	3
1	1	1	360000
2	2	1	19522
3	2	2	522

Figure 6.23: *Objects in first and second cluster*

8x3 double

	1	2	3
1	1	1	360000
2	2	1	19522
3	2	2	522
4	3	1	120438
5	3	2	539
6	3	3	2107
7	3	4	550
8	3	5	357

Figure 6.24: *Objects in first, second and third clusters*

The Figures 6.22, 6.23, and 6.24 state the labeled objects in three clusters. The first column in each table represents the number of clusters, the second column indicates number of each object related to that cluster and the third column specifies the area of each object (one of the required information).

6.4 Tumor Detection

Since the mammogram segmentation and ROIs extraction have been done and additionally the objects in ROIs have been labeled, therefore the upcoming step of this thesis work is to detect tumors in ROIs. The objective of the tumor detection phase is to detach cancerous and non-cancerous mammograms in the whole dataset.

The proposed method to identify tumors (cancerous cells) in the ROIs is Thresholding technique. The appropriate threshold value depends on all mammograms in the dataset. The optimal value to discover tumors in ROIs is estimated based on the Trial-and-Test technique and the size of tumors in the whole dataset.

Obviously, the assessment of threshold value is a challenging task for a huge dataset with various ranges in size and shape of cancerous cells, and it may cause the increment in False Positive and False Negative values.

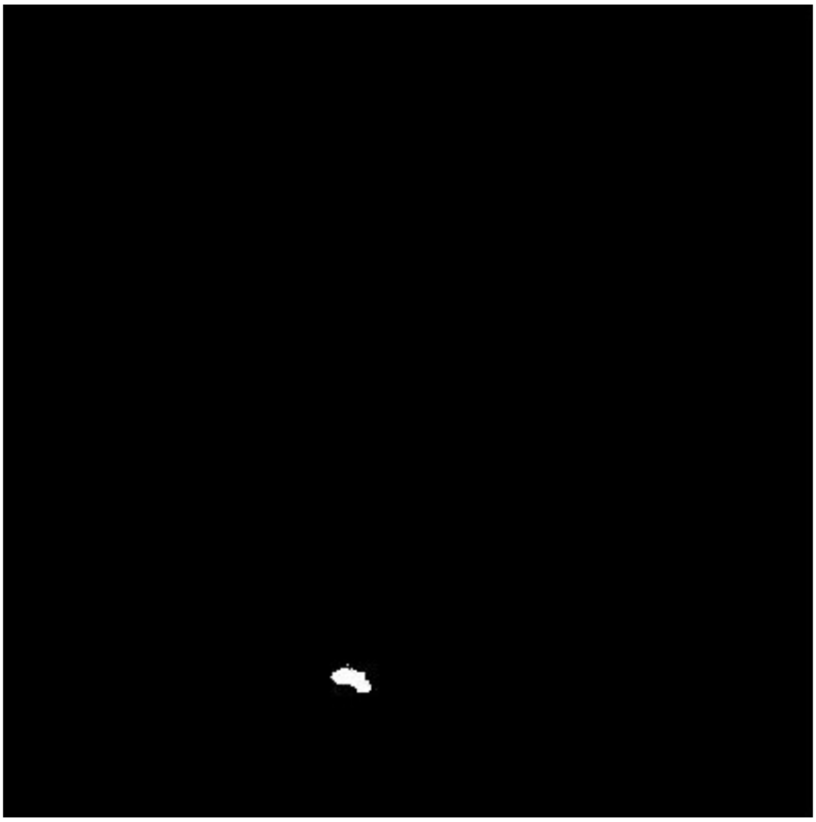


Figure 6.25: *Tumor Detection result*

Figure 6.25 displays the detected tumor in the ROI based on the thresholding technique. According to entire dataset, mainly the suspicious object in ROIs has a brighter color in the binary image (the pixels' intensity of tumor are more white) in comparison of normal tissues and this expresses compressed tissue of them.

6.5 Conclusion of Processing step

Breast cancer is one of the most common causes of cancer death in women, while the mortality rate can be greatly reduced by early detection. Thus, different computer-aided detection (CAD) techniques are applied to digital mammography to provide a fast diagnosis with high accuracy [56]. The main goal of this thesis is primarily to distinguish the boundary of breast tissue with high accuracy and considering it as the desired region after separating it from the underarm area. Furthermore detecting abnormal masses that helps for the classification and separation of the cancerous and non-cancerous images in the dataset, and all by using image processing and machine learning strategies. In this study, the FCM clustering algorithm has been implemented for segmenting mammogram images, which is done based on the intensity values of pixels. Moreover, for accuracy increment in the image segmentation part, a non-linear edge detection technique is performed. The thresholding method is applied to detect tumors in the suspicious region of mammograms in order to isolate normal and abnormal breast tissue. Generally, the image segmentation and features extraction perform well in identifying mass regions in mammograms.

Chapter 7

Conclusion and future work

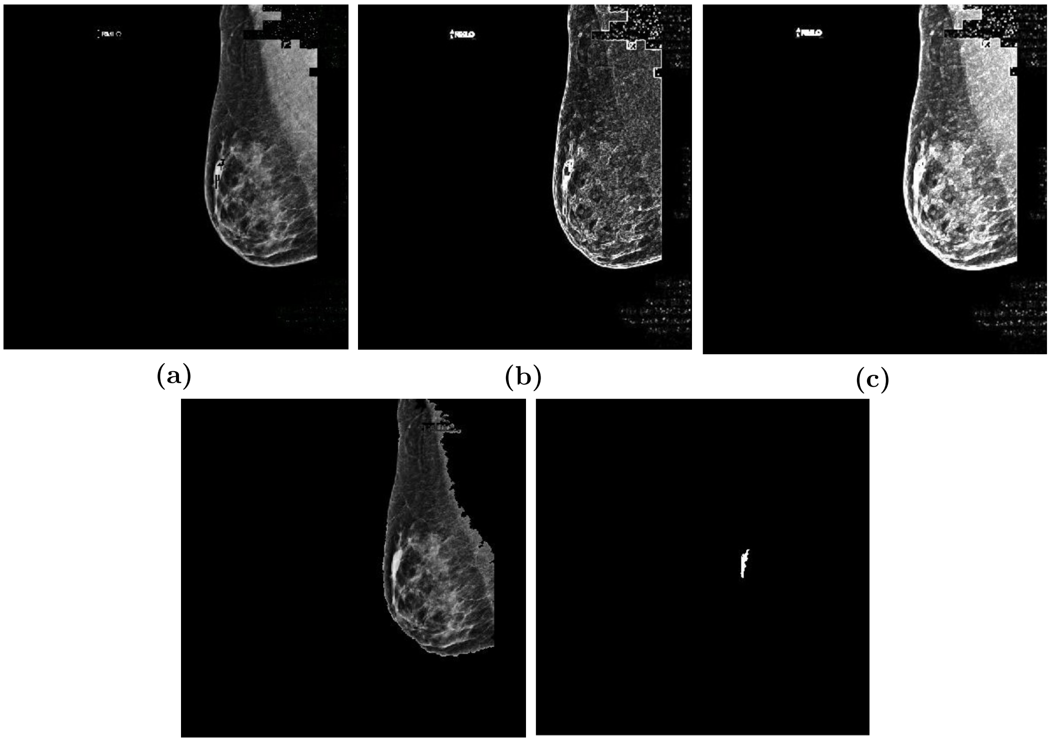
Breast cancer is the most common cancer among women in the world but diagnostic at an early stage makes treatment more effective. Due to the fact that breast tissue has many variations from person to person, therefore it is a challenging problem to characterize mammograms.

Hence in this dissertation, a full-field analysis method is ended, which focuses on mammograms' segmentation and tumor identification in order to help radiologists in the mammography analysis.

The Fuzzy C-Means(FCM) clustering algorithm, as an unsupervised machine learning (ML) technique, is applied to the dataset with the aim of mammogram image segmentation to automatically determine accurate regions of interest (ROIs) and identify characteristics from breast lesions. Although FCM has relatively good performance on segmenting images, in order to increase the accuracy of segregation results, an edge detection method (Kirsch Compass Kernel algorithm) is added to this approach. Furthermore, the edge detection result is also enhanced by performing a Histogram Equalizer.

The achievements related to the study on mammogram images in the dataset indicate that mostly the underarm area and masses have a lighter color in comparison to the normal breast tissue thus their pixels will include greater intensity values with respect to the other pixels in the image. Moreover, tissue density is independent of the breast size, and tumor detection in the breast with higher density seems more challenging.

In order to develop the research in this field and for further investigation, the study on characteristics of various mammogram images by the evaluation of some breast properties such as size or volume and tissue density could be considered as an upcoming plan for advanced analysis.



(a) (b) (c) (d) (e)

Figure 7.1: *Results of the presented method respectively for one mammography of dataset: de-identified Mammogram, edge-detection result, histogram equalization, ROI segregation, and tumor detection.*

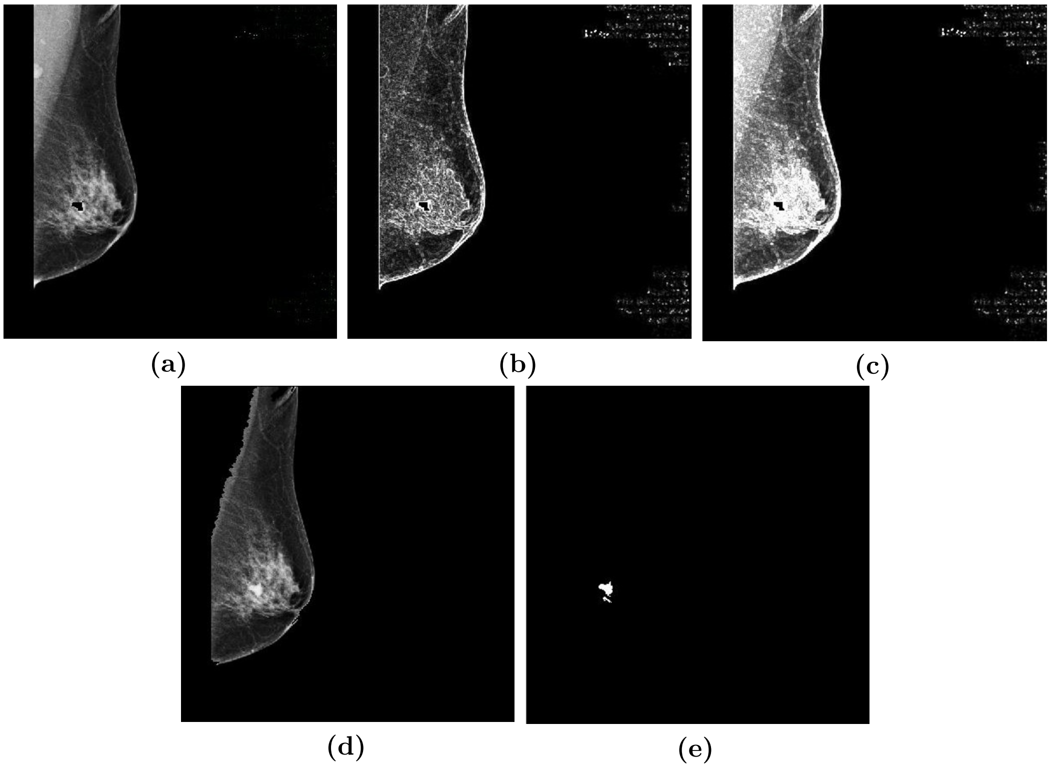


Figure 7.2: *Results of the presented method respectively for one mammography of dataset: de-identified Mammogram, edge-detection result, histogram equalization, ROI segregation, and tumor detection.*

Chapter 8

Discussion

Our presented method in the project will be compared with some of the other recent studies in order to assess the performance of the different techniques in achieving mammogram image segmentation and tumor detection.

Mostly, the assessment is made based on differentiation in terms of obtained precision or other evaluation metrics for each method. However, it must be noted that it is hard to make a real comparison due to this reason that some researchers used visual estimation of the radiologist to evaluate their methods. In addition, an obtained result of an algorithm is in high relation to some major features in the data set such as the number of data samples, or the data quality and homogeneity. Obviously, there are limited number of researches that use same dataset for the same reason. According to all the stated points, the accurate comparison of the performance of multiple algorithms requires further knowledge.

8.1 First Article

Foundation and methodologies in computer-aided diagnosis systems for breast cancer detection [110].

Up to now, several algorithms have been developed by many researchers to improve the performance of CAD systems in healthcare. In this paper, the authors discussed some methods of mammography analyses in cancer detection such as fuzzy logic, Genetic algorithm, and neural network. Based on this research, preprocessing is required as the initial step for the quality improvement of mammograms and data preparation to use in other investigations such as the segmentation process. The noise will remove by using different enhancement methods such as a median filter, morphological operation, and thresholding method with the use of MATLAB functions. Contrast enhancement can be performed by histogram equalization by increasing intensity values.

The next issue discussed in the paper is image segmentation, which aids to extract important regions for advanced analysis of the structures such as organs or tumors. The selection of adequate segmentation techniques is depending on image quality. Edge-based segmentation methods are a systematic technique to identify pixels among the regions that have abrupt intensity variation. Image segmentation and analyses also can be done through thresholding-based technique, which is one of the most used methods to separate objects in an image or extract desired regions for further analysis. The important part of this method is the selection of a proper value of thresholding based on image properties, which significantly depends on the homogeneity of data.

Finally, an evaluation is done to compare thresholding techniques to the region-based methods and clustering-based algorithms, which all are wildly used for image segmentation and lesion detection. Clustering-based segmentation algorithms are common methods in medical image segmentation and divided into two groups:

- Hierarchal clustering which is a recursive process
- partitioning clustering is considered as an iterative procedure that can be divided into hard clustering and fuzzy clustering.

Both clustering and region-based algorithms are effective methods in breast segmenting and cancer detection by extracting ROIs. Additionally, artificial intelligence algorithms and support vector machines have been extensively used in tumor classification and breast cancer diagnosis. However, data inconsistency is a critical problem in pattern recognition and tumor detection, which needs to be solved through different approaches.

Our accomplished research is focused on detecting any abnormality by mammogram image segmentation. The collected dataset includes 260 X-ray images of the breast that are divided into 150 normal and 110 abnormal mammogram images and a highly advantageous issue in our dataset is to some extent the data homogeneity. During the pre-processing stage, the mammogram images are converted into binary images based on the thresholding technique. Then morphological filters in MATLAB are performed on mammogram images to remove any noise specifically noise of conversion. In the segmentation stage, a clustering-based approach is considered due to the sufficient efficiency and in order to improve the accuracy of this process an edge detection technique is applied on the mammogram images and in addition a histogram equalization is employed to enhance the pixel intensity contrast. As the final phase of our work and after object feature extraction, the thresholding method is selected to be applied to the homogenous dataset. An optimal threshold makes the possibility to identify the tumor or any suspicious mass in the mammogram images.

Although our project is composed of several techniques however we can claim that the totality of our work is based on the FCM clustering algorithm hence it is a clustering-based model for image segmentation.

8.2 Second Article

Image Enhancement and Edge-based Mass Segmentation in Mammogram [111].

Generally, mammograms have low contrast, hence detecting suspicious mass or splitting ROI will be a challenging task. In the second chosen research, the focus is on mass segmentation and border identification of suspicious mass in the region of interest (ROI). In fact, an ROI is part of the mammogram image that contains a suspicious mass and the ROI has been detected by CAD systems through the edge-based segmentation process.

As a brief explanation of the whole work, firstly an image enhancement technique, Gaussian filter, was applied only on the breast tissue as ROI section of the image for contrast improvement and removing noise. Next, the optimal thresholds were selected to obtain initial segmented mass regions. Indeed the edges were detected to assist in identifying mass regions. Furthermore, morphologic dilation operation was applied to the binary image to obtain the initial contour of the mass region, and in the end, the mass boundary of the region was indicated. The suspicious areas were extracted using iterative thresholding through developing a region-based mass segmentation algorithm. In the mass segmentation method by using an edge-based approach, the "optimal" contour of a mass was detected from the edge candidate points.

An edge-based segmentation technique is one of the methods presented in this paper for recognizing the tumor in ROI of mammography. In this method, edges are identified by a pixel-level energy texture image of a mass ROI. During this experiment, it became clear that image enhancement has a huge effect on the quality of segmentation results.

The entire work in this paper is to extract the ROI then enhance the ROI partition contrast and indicate tumor boundary in ROI, while all these main tasks are done through edge-based segmentation method with the addition of thresholding techniques. According to this approach, the quality of mammogram images, data homogeneity, and the contrast of the images are important features that have a significant effect on the results. Therefore, the preprocessing step has a critical role in obtaining accurate results, particularly in the edge-based segmentation methods.

In our work, the focus is on the implementation of several appropriate and diverse techniques to explore and present different aspects in mammogram image analysis.

8.3 Third Article

Mammogram image segmentation for improving the diagnosis of dense breast issues [32].

Breast cancer is one of the major causes of death among women in the whole world. However, today the most impressive technique for early detection of breast lesions is using mammography screening. Radiologists analyze mammograms by using CAD systems, to identify any abnormality in the breast. Although detection of suspicious masses or segmentation of body organs in medical images is a difficult job, different image segmentation techniques were developed for this purpose and the Fuzzy c-means clustering (FCM) as an unsupervised algorithm is a widespread method. The noise-free images are successfully segmented by using the standard FCM method.

The mammogram segmentation method proposed in this article is for the enhancement of mass diagnosis in the medical images and it is applied by the FCM clustering with and without adding spatial information on the three mammogram images with low tissue dense, denser, and high density. Additionally, the applied segmentation algorithm was enhanced by using the Wiener filter to increase the contrast of noisy pixels.

The results state that the presented algorithms for dividing images into several regions were able to produce the desired ROI (regions of high density) based on the clustering approach. However, the optimum centers for each cluster of images were obtained faster and with fewer iterations in spatial FCM in contrast to traditional FCM, thus the spatial FCM introduced segmentation results with relatively higher quality. The methods were successful in diagnostic the dense breast areas. Although the FCM clustering algorithm is qualified to perform mammogram image segmentation very efficiently, as regards the FCM is very sensitive to noise. Obviously, the high precision in medical image analysis is a critical issue due to the effect on better diagnosis and treatment. Due to the fact that most medical images are noisy, therefore a good approach to improve assessing and accuracy increment is to implement an adaptive FCM algorithm or using an additional method with the standard FCM which is also employed in our project.

8.4 Forth Article

Mammography Images Segmentation via Fuzzy C-mean and K-mean [112].

In this paper, the segmentation is done based on two different clustering algorithms: **the K-mean and the FCM algorithm**, and as an assistance method to improve the performance of segmentation in ROI extraction results another algorithm was applied on both mentioned algorithms.

According to the statistics assessment on mammograms, the FCM algorithm provided fairly high results (for dataset with sufficient quality) in terms of accuracy in segmenting mammograms and detecting abnormal region boundaries which is due to the flexibility related to the fuzzy logic technique. While, it is a difficult task to detect tumor boundaries accurately due to some factors like image quality, tumor size, tumor volume and shape, and even the lesion location. As it became clear earlier, the same as other stated techniques, it is beneficial to use an extra technique on the FCM algorithm to improve the obtained results of mass area diagnosis.

However, one of the major works that were not mentioned in this article, while it can be relatively helpful in decreasing the False Positive and Negative rates is the pre-processing step before any other analyses on mammogram images and the selected preprocessing method mainly depends on image quality.

8.5 Fifth Article

Suspicious Lesion Segmentation on Brain, Mammograms and Breast MR Images Using New Optimized Spatial Feature Based Super-Pixel Fuzzy C-Means Clustering [113].

Image segmentation is proposed in healthcare, for medical image analyses. The FCM clustering method is a very appropriate algorithm for segmenting images into separate regions, which helps in lesion detection. However, the disadvantage of this technique is sensitivity to noise hence it is needed to perform an additional method combined with FCM or a modified version of standard FCM to achieve better segmentation or tumor detection results. In this study, the author tried to compare the obtained result of several adaptive FCM algorithms such as FCM_S, KFCM, and SPOFCM to the standard FCM. The outcomes indicate that noise can largely affect the FCM algorithm and leads to poor performance while the SPOFCM represented the best proficiency among the other mentioned algorithms in segmentation of the mammograms. Therefore, super-pixel-based FCM clustering offered superior efficiency rather than FCM due to the robustness in terms of noise.

The assessments of experimental results on mammogram images state that the presented algorithm SPOFCM was provided acceptable performance in comparison to the standards offered by the specialists. Obviously, a super-pixel has the capability to prepare more local information than Spatial FCM, and it is more robust to noise, therefore it is proposed an appropriate algorithm for the image segmentation quality improvement in comparison to the standard FCM or FCM-S.

Since sharp edges identification, which denotes as abrupt variation in points intensities in an image conducts a considerable performance in object detection. In our performed approach, in order to improve the result of the FCM clustering algorithm in the segmentation stage, which leads to ROI extraction more precisely, an edge detection method based on the Kirsch compass kernel algorithm is proposed to aid in finding the maximum edge in eight compass directions.

8.6 Sixth Article

Machine learning and Region Growing for Breast Cancer Segmentation [114].

The discussion of this study is concentrated on mammogram image segmentation for lesion diagnosis in the breast tissue, and its complexities. Based on the research most difficulties are due to breast tissues and pixels' intensities. A starting point, for detecting any abnormality in the images is to extract the ROIs. Hence, a trainable segmentation method is applied to find the ROIs and then region growing as the other segmentation technique is implemented to obtain tissue partitioning with high precision. The stated model was made based on a two-layer backpropagation neural network and consists of two parts: **training and testing**. The whole dataset contains 250 mammogram images. This experiment indicated that the presented method is good enough to prepare acceptable segmentation results in comparison to the manual measurements or some of the automatic techniques such as Otsu thresholding or K-means clustering.

Our implemented method is an efficient method and although it is generally based on an unsupervised FCM clustering algorithm, a powerful edge-based technique and a histogram equalization are implemented to improve the quality of the breast segmentation stage. Finally, a thresholding method is applied in order to detect any suspiciously mass. While our presented approach is composed of several algorithms and methods, it still preserves its simpleness and performance proficiency and illustrates a very high quality of this hierarchical combination.

Bibliography

[1] A.G. Rivenbark and William Coleman. «An Introduction to the Conspicuous and Distinguishing Characteristics of Neoplasms». In: Dec. 2014, pp. 349–366. ISBN: 9780123864574. DOI: 10.1016/B978-0-12-386456-7.01901-8 (cit. on p. 3).

[2] Amir Hosein Jafarian, Melika Kooshkiforooshani, Abdolshakor Rasoliostadi, and Nema Mohamadian Roshan. «Vascular Mimicry Expression in Invasive Ductal Carcinoma, A New Technique for Prospect of Aggressiveness». In: *Iranian Journal of Pathology* 14.3 (2019), pp. 232–235. ISSN: 1735-5303. DOI: 10.30699/ijp.2019.94997.1939. eprint: http://ijp.iranpath.org/article_36275_61ba2a7b70818d55e294b24469da00b2.pdf (cit. on p. 3).

[3] Rafael Gonzalez, Richard Woods, and Barry Masters. «Digital Image Processing, Third Edition». In: *Journal of biomedical optics* 14 (Mar. 2009), p. 029901. DOI: 10.1117/1.3115362 (cit. on pp. 3–6, 9, 19).

[4] N. Mohamadian Roshan, S. Salehabadi, M. Kooshki Forooshani, and A. Nikoozadeh. «Accuracy of Digital Image Analysis (DIA) of Borderline Human Epidermal Growth Factor Receptor (HER2) Immunohistochemistry in Invasive Ductal Carcinoma». In: *International Journal of Cancer Management* 13.9 (2020), e101179. ISSN: 2538-497X. DOI: 10.5812/ijcm.101179. URL: https://sites.kowsarpub.com/ijcm/articles/101179.html (cit. on p. 3).

[5] Victor Mikla and Victor Mikla. «Advances in Imaging from the First X-Ray Images». In: Dec. 2014, pp. 1–22. ISBN: 9780124170216. DOI: 10.1016/B978-0-12-417021-6.00001-0 (cit. on p. 3).

[6] Bradley Erickson, Panagiotis Korfiatis, Zeynettin Akkus, and Timothy Kline. «Machine Learning for Medical Imaging». In: *RadioGraphics* 37 (Feb. 2017), p. 160130. DOI: 10.1148/rg.2017160130 (cit. on p. 4).

[7] Michael Rowe. «An Introduction to Machine Learning for Clinicians». In: *Academic Medicine* 94 (May 2019), p. 1. DOI: 10.1097/ACM.0000000000002792 (cit. on pp. 4, 5).

[8] Shahadat Uddin, Arif Khan, Md Hossain, and Mohammad Ali Moni. «Comparing different supervised machine learning algorithms for disease prediction». In: *BMC Medical Informatics and Decision Making* 19 (Dec. 2019). DOI: 10.1186/s12911-019-1004-8 (cit. on pp. 4, 5).

[9] Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. «Deep Reinforcement Learning that Matters». In: (Sept. 2017) (cit. on pp. 4, 5).

[10] Nari Arunraj, Robert Hable, Michael Fernandes, Karl Leidl, and Michael Heigl. «Comparison of Supervised, Semi-supervised and Unsupervised Learning Methods in Network Intrusion Detection System (NIDS) Application». In: *Anwendungen und Konzepte der Wirtschaftsinformatik* 6 (Nov. 2017), pp. 10–19 (cit. on pp. 4, 5).

[11] Muhammad Akram, Mehwish Iqbal, Muhammad Daniyal, and Asmat Khan. «Awareness and current knowledge of breast cancer». In: *Biological Research* 50 (Dec. 2017). DOI: 10.1186/s40659-017-0140-9 (cit. on p. 6).

[12] Filippo Pesapane, Francesco Sardanelli, and Marina Codari. «Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine.» In: *European Radiology* 24 (Oct. 2018), p. 35. DOI: 10.1186/s41747-018-0061-6. (cit. on pp. 6, 12).

[13] Carlos J. Gulo, Antonio Sementille, and Joao Tavares. «Techniques of Medical Image Processing and Analysis accelerated by High-Performance Computing: A Systematic Literature Review». In: *Journal of Real-Time Image Processing* 16 (Dec. 2019), pp. 1891–1908. DOI: 10.1007/s11554-017-0734-z (cit. on p. 6).

[14] Huiyu Zhou, Gerald Schaefer, and Chunmei Shi. «Fuzzy C-Means Techniques for Medical Image Segmentation». In: vol. 242. Jan. 1970, pp. 257–271. DOI: 10.1007/978-3-540-89968-6_13 (cit. on p. 6).

[15] Baxtyar Ahmed and Muzhir Al-Ani. «Digital Medical Image Segmentation Using Fuzzy C-Means Clustering». In: *UHD Journal of Science and Technology* 4 (Feb. 2020), p. 51. DOI: 10.21928/uhdjst.v4n1y2020.pp51-58 (cit. on p. 6).

[16] Tao Lei, Xiaohong Jia, Yanning Zhang, Lifeng He, Hongying Meng, and Asoke Nandi. «Significantly Fast and Robust Fuzzy C-Means Clustering Algorithm Based on Morphological Reconstruction and Membership Filtering». In: *IEEE Transactions on Fuzzy Systems* PP (Jan. 2018), pp. 1–1. DOI: 10.1109/TFUZZ.2018.2796074 (cit. on p. 6).

[17] Venketkumar Hariraj, Wan Khairunizam, Vikneswaran Vijean, Zunaidi Ibrahim, Shahriman Abu bakar, Zuradzman Mohamad Razlan, Rajendran Thavasimuthu, and R. Sathiyasheelan. «Fuzzy multi-layer SVM classification of breast cancer mammogram images». In: *International Journal of Mechanical Engineering and Technology* 9 (Aug. 2018), pp. 1281–1299 (cit. on p. 6).

[18] Kalaiselvi Thiruvenkadam, J Prasanna, and Karthigaiselvi Sakthivelu. «Modified Kirsch's Method for Edge Based Image Segmentation». In: Mar. 2005 (cit. on p. 6).

[19] Scott McKinney et al. «International evaluation of an AI system for breast cancer screening». In: *Nature* 577 (Jan. 2020), pp. 89–94. DOI: `10.1038/s41586-019-1799-6` (cit. on p. 7).

[20] Yashar Moharamzad, Amir Davarpanah, Ali Yaghobi Joibari, fatemeh shahbazi fatemeh, Leila Toosi, Melika Koshkifrooshani, Ali Ansari, and Morteza Sanei Taheri. «Diagnostic performance of apparent diffusion coefficient (ADC) for differentiating endometrial carcinoma from benign lesions: a systematic review and meta-analysis». In: *Abdominal Radiology* (Mar. 2021). DOI: `10.1007/s00261-020-02734-w` (cit. on p. 7).

[21] Igor Kononenko. «Machine learning for medical diagnosis: History, state of the art and perspective». In: *Artificial intelligence in medicine* 23 (Sept. 2001), pp. 89–109. DOI: `10.1016/S0933-3657(01)00077-X` (cit. on pp. 7, 8).

[22] Laura Mcdonald, Sreeram Ramagopalan, and Andrew Cox. «Unintended consequences of machine learning in medicine?» In: *F1000Research* 6 (Sept. 2017), p. 1707. DOI: `10.12688/f1000research.12693.1` (cit. on p. 7).

[23] Prableen Kaur, Manik Sharma, and Mamta Mittal. «Big Data and Machine Learning Based Secure Healthcare Framework». In: *Procedia Computer Science* 132 (June 2018), pp. 1049–1059. DOI: `10.1016/j.procs.2018.05.020` (cit. on p. 8).

[24] Gayathri B.M, C.P. Sumathi, and Santhanam T. «Breast Cancer Diagnosis Using Machine Learning Algorithms - A Survey». In: *International Journal of Distributed and Parallel systems* 4 (May 2013), pp. 105–112. DOI: `10.5121/ijdps.2013.4309` (cit. on p. 8).

[25] Saudagar AKJ Alkhathlan L. «Machine Learning Techniques for Breast Cancer Analysis: A Systematic Literature Review.» In: 2020 (cit. on p. 8).

[26] Joseph Cruz and David Wishart. «Applications of Machine Learning in Cancer Prediction and Prognosis». In: *Cancer informatics* 2 (Feb. 2007), pp. 59–77. DOI: `10.1177/117693510600200030` (cit. on p. 8).

[27] Alfredo Vellido and Paulo Lisboa. «Neural Networks and Other Machine Learning Methods in Cancer Research». In: June 2007, pp. 964–971. ISBN: 978-3-540-73006-4. DOI: 10.1007/978-3-540-73007-1_116 (cit. on p. 8).

[28] Rajkumar K. K. «Detection and classification of mammogram images using K-Means and Fuzzy C-Means clustering algorithm with texture features». In: 2019 (cit. on pp. 8, 16).

[29] Jinshan Tang, Rangaraj Rangayyan, Jun Xu, Issam El Naqa, and Yongyi Yang. «Computer-Aided Detection and Diagnosis of Breast Cancer With Mammography: Recent Advances». In: *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society* 13 (Feb. 2009), pp. 236–51. DOI: 10.1109/TITB.2008.2009441 (cit. on p. 9).

[30] S. Meenalosini, Dr. J. Janet, M. Tech, and Dr. E. Kannan. «Segmentation Of Cancer Cells In Mammogram Using Region Growing Method And Gabor Features». In: 2012 (cit. on pp. 9, 21).

[31] Marjan Laal. «Innovation Process in Medical Imaging». In: *Procedia - Social and Behavioral Sciences* 81 (June 2013), pp. 60–64. DOI: 10.1016/j.sbspro.2013.06.388 (cit. on pp. 9, 10).

[32] Marwah Ali, Raad Chisab, and Faiez Al-Rufaye. «Mammogram image segmentation for improving the diagnosis of dense breast issues». In: *International Journal of Engineering and Technology* 8 (Jan. 2019), pp. 44–52 (cit. on pp. 10, 29, 56).

[33] Cristy H. «Digital Breast Tomosynthesis». In: (cit. on p. 10).

[34] C. Ho and Hailey. «Digital Mammography versus Film-Screen Mammography : Technical , Clinical and Economic Assessments». In: 2002 (cit. on p. 10).

[35] Martin Yaffe. «Developing a quality control program for digital mammography: Achievements so far and challenges to come». In: *Imaging in Medicine* 3 (Feb. 2011), pp. 123–133. DOI: 10.2217/iim.10.63 (cit. on pp. 10, 11).

[36] Etta D. Pisano et al. «Image Processing Algorithms for Digital Mammography: A Pictorial Essay». In: *RadioGraphics* 20.5 (2000), pp. 1479–1491 (cit. on pp. 10, 11).

[37] Chung-Ming Chen, Yi-Hong Chou, Norio Tagawa, and Younghae Do. «Computer-Aided Detection and Diagnosis in Medical Imaging». In: *Computational and mathematical methods in medicine* 2013 (Sept. 2013), p. 790608. DOI: 10.1155/2013/790608 (cit. on p. 11).

[38] Hiroshi Fujita et al. «Computer-aided diagnosis: The emerging of three CAD systems induced by Japanese health care needs». In: *Computer methods and programs in biomedicine* 92 (June 2008), pp. 238–48. DOI: `10.1016/j.cmpb.2008.04.003` (cit. on pp. 11, 12).

[39] Bhagirathi Halalli and Aziz Makandar. «Computer Aided Diagnosis - Medical Image Analysis Techniques». In: Jan. 2018. ISBN: 978-953-51-3732-0. DOI: `10.5772/intechopen.69792` (cit. on p. 12).

[40] Jun Gao, Qian Jiang, Bo Zhou, and Daozheng Chen. «Convolutional neural networks for computer-aided detection or diagnosis in medical image analysis: An overview». In: *Mathematical Biosciences and Engineering* 16 (July 2019), pp. 6536–6561. DOI: `10.3934/mbe.2019326` (cit. on p. 12).

[41] Syed Anwar, Muhammad Majid, Adnan Qayyum, Muhammad Awais, Majdi Alnowami, and Khurram Khan. «Medical Image Analysis using Convolutional Neural Networks: A Review». In: *Journal of Medical Systems* 42 (Oct. 2018), p. 226. DOI: `10.1007/s10916-018-1088-1` (cit. on p. 12).

[42] Afsaneh Jalalian, Syamsiah Mashohor, Rozi Mahmud, M Iqbal, B Saripan, Abdul Rahman, B Ramli, and Babak Karasfi. «Computer-aided detection/-diagnosis of breast cancer in mammography and ultrasound: a review». In: (June 2020) (cit. on p. 12).

[43] Ashraf Aly H., Safaai deris, and Nazar Zaki. «Research Review For Digital Image Segmentation Techniques.» In: *International Journal of Computer Science and Information Technology* 3 (Oct. 2011). DOI: `10.5121/ijcsit.2011.3509` (cit. on pp. 12, 14).

[44] Devakumari DD Punithavathi V. «A Framework on Classification of Mammogram Images for Breast Cancer Detection using Image Processing with Data Mining Techniques.» In: 2020 (cit. on p. 12).

[45] Waseem Khan. «Image Segmentation Techniques: A Survey». In: *Journal of Image and Graphics* 1 (Jan. 2014), pp. 166–170. DOI: `10.12720/joig.1.4.166-170` (cit. on pp. 12, 14).

[46] Ali Yadollahpour and Hamed Shoghi. «Early Breast Cancer Detection using Mammogram Images: A Review of Image Processing Techniques». In: *Biomedical and Pharmacology Journal* 12 (Mar. 2015), pp. 225–234. DOI: `10.13005/bbra/1627` (cit. on pp. 13, 14).

[47] Susama Bagchi, Tay Gaik, Audrey Huong, and Sanjoy Debnath. «Image processing and machine learning techniques used in computer-aided detection system for mammogram screening - a review». In: *International Journal of Electrical and Computer Engineering (IJECE)* 10 (June 2020), p. 2336. DOI: `10.11591/ijece.v10i3.pp2336-2348` (cit. on pp. 13, 14).

[48] Samir Bandyopadhyay. «Pre-processing of Mammogram Images». In: *International Journal of Engineering Science and Technology* 2 (Nov. 2010) (cit. on pp. 13, 16).

[49] Sneha Sable. «Implementation and Analysis of K-Means and Fuzzy C Means Clustering Techniques». In: 2015 (cit. on pp. 13, 45).

[50] Yoshitaka Kimori. «Mathematical morphology-based approach to the enhancement of morphological features in medical images». In: *Journal of clinical bioinformatics* 1 (Dec. 2011), p. 33. DOI: 10.1186/2043-9113-1-33 (cit. on pp. 13, 14).

[51] M Sukassini and Velmurugan T. «Noise removal using Morphology and Median filter Methods in Mammogram Images». In: Jan. 2016 (cit. on p. 13).

[52] Tomislav Stojić, Irini Reljin, and Branimir Reljin. «Local contrast enhancement in digital mammography by using mathematical morphology». In: vol. 2. Aug. 2005, 609–612 Vol. 2. ISBN: 0-7803-9029-6. DOI: 10.1109/ISSCS.2005.1511314 (cit. on p. 14).

[53] G. Boato, Duc Tien Dang Nguyen, and F.G.B. Natale. «Morphological Filter Detector for Image Forensics Applications». In: *IEEE Access* PP (Jan. 2020), pp. 1–1. DOI: 10.1109/ACCESS.2020.2965745 (cit. on p. 14).

[54] R. Pandey and S. Mathurkar. «A Review on Morphological Filter and its Implementation». In: 2017 (cit. on p. 14).

[55] D Ponraj, M Jenifer, P. Poongodi, and Samuel Manoharan. «A Survey on the Preprocessing Techniques of Mammogram for the Detection of Breast Cancer». In: *Journal of Emerging Trends in Computing and Information Sciences* 2 (Jan. 2011) (cit. on p. 14).

[56] Velmurugan T and E Venkatesan Venkatesan. «Effective Fuzzy C Means Algorithm for the Segmentation of Mammogram images of Identify Breast Cancer». In: *Journal of Control Theory and Applications* 9 (Aug. 2016), pp. 4647–4660 (cit. on pp. 14, 16, 49).

[57] R. Boss, Dr. Thangavel Kuttiyannan, and Arul Daniel. «Mammogram image segmentation using fuzzy clustering». In: Mar. 2012, pp. 290–295. ISBN: 978-1-4673-1037-6. DOI: 10.1109/ICPRIME.2012.6208360 (cit. on pp. 14, 29, 30).

[58] M.P. Sukassini and Velmurugan T. «A survey on the analysis of segmentation techniques in mammogram images». In: 8 (Jan. 2015). DOI: 10.17485/ijst/2015/v8i22/79105 (cit. on pp. 14, 16).

[59] Mingyan Li, Radha Poovendran, and Sreeram Narayanan. «Protecting patient privacy against unauthorized release of medical images in a group communication environment». In: *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society* 29 (Aug. 2005), pp. 367–83. DOI: 10.1016/j.compmedimag.2005.02.003 (cit. on p. 15).

[60] György Szarvas, Richárd Farkas, and Róbert Busa-Fekete. «State-of-the-art Anonymization of Medical Records Using an Iterative Machine Learning Framework». In: *Journal of the American Medical Informatics Association : JAMIA* 14 (Sept. 2007), pp. 574–80. DOI: 10.1197/j.jamia.M2441 (cit. on p. 15).

[61] Aris Gkoulalas-Divanis and Grigorios Loukides. «Introduction to Medical Data Privacy». In: Jan. 2015, pp. 1–14. DOI: 10.1007/978-3-319-23633-9_1 (cit. on p. 15).

[62] Kumar G Saluja B. «Anonymization of Sensitive Information in Medical Health Records». In: (2019) (cit. on p. 15).

[63] Mithun Gaur. «Privacy Preserving Machine Learning Challenges and Solution Approach for Training Data in ERP Systems». In: *INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING and TECHNOLOGY* Volume 11 (Aug. 2020), pp. 31–36. DOI: 10.2139/ssrn.3679275 (cit. on p. 15).

[64] Damian Valdes Santiago, Raul Martínez, Angela Leon Mecias, and Marta Díaz-Romanach. «Mammographic Mass Segmentation Using Fuzzy C–means and Decision Trees». In: July 2018. ISBN: 978-3-319-94543-9. DOI: 10.1007/978-3-319-94544-6_1 (cit. on pp. 15, 29).

[65] R. Muthaiyan, M. Priyadharshini, C. Sujatha, and A. Deepika. «Analysis of Brain Tumor Classification by using Multiple Clustering Algorithms». In: 2018 (cit. on pp. 15, 16).

[66] Ulrike Luxburg, Robert Williamson, and Isabelle Guyon. «Clustering: Science or Art?» In: *J Mach Learn Res* 27 (July 2011) (cit. on p. 15).

[67] Janmenjoy Nayak, Bighnaraj Naik, and Dr. H. Behera. «Fuzzy C-Means (FCM) Clustering Algorithm: A Decade Review from 2000 to 2014». In: (Jan. 2015), pp. 133–149 (cit. on pp. 15, 16).

[68] Miin-Shen Yang. «A survey of fuzzy clustering. Math Comput Model». In: *Mathematical and Computer Modelling* 18 (Dec. 1993), pp. 1–16. DOI: 10.1016/0895-7177(93)90202-A (cit. on p. 16).

[69] Prakasha S. Divya D. J. «Clustering Techniques for Medical Imaging». In: (2019) (cit. on p. 16).

[70] Monika Xess. «SURVEY ON CLUSTERING BASED COLOR IMAGE SEGMENTATION AND NOVEL APPROACHES TO FCM ALGORITHM». In: *International Journal of Research in Engineering and Technology* 02 (Dec. 2013), pp. 346–349. DOI: `10.15623/ijret.2013.0212059` (cit. on p. 16).

[71] Ramesh Jain, Rangachar Kasturi, and Brian Schunck. *Machine Vision*. Jan. 1995. ISBN: 978-0-07-032018-5 (cit. on pp. 17, 18).

[72] Sukhwinder Kaur and Saurav Bansal. «Review On: Analysis of Edge Detection with Color Control». In: 2015 (cit. on pp. 17, 19, 20).

[73] Haldo Spontón and Juan Cardelino. «A Review of Classic Edge Detectors». In: *Image Processing On Line* 5 (June 2015), pp. 90–123. DOI: `10.5201/ipol.2015.35` (cit. on pp. 17, 18).

[74] Vikrant Bhateja, Swapna Devi, and Shabana Urooj. «An Evaluation of Edge Detection Algorithms for Mammographic Calcifications». In: vol. 222. Jan. 2013, pp. 487–498. ISBN: 978-81-322-1000-9. DOI: `10.1007/978-81-322-1000-9_46` (cit. on p. 18).

[75] K. Krishnan, Shiva Ranga, and Nageswara Guptha. «A Survey on Different Edge Detection Techniques for Image Segmentation». In: *Indian Journal of Science and Technology* 10 (Jan. 2017). DOI: `10.17485/ijst/2017/v10i4/108963` (cit. on p. 18).

[76] Samir, Ranita Banerjee, Dr. Indra Maitra, Samir Bandyopadhyay, and B Poddar. «A SURVEY ON RELEVANT EDGE DETECTION TECHNIQUE FOR MEDICAL IMAGE PROCESSING». In: (Oct. 2018) (cit. on p. 19).

[77] Ahmed Abdullah, Majid Ali, and Hasan Jawad. «Evaluation of the Performance of Kirsch and Robinson Methods on Different Color Spaces». In: (2019) (cit. on pp. 19, 20).

[78] Kumari S Rani R. «A Survey on Edge Detection Using Different Techniques». In: () (cit. on p. 19).

[79] Kumaratharan N Venmathi AR Ganesh EN. «Kirsch Compass Kernel Edge Detection Algorithm for Micro Calcification Clusters in Mammograms». In: *Middle East Journal of Scientific Research* 24 (Jan. 2016), pp. 1530–1535. DOI: `10.5829/idosi.mejsr.2016.24.04.23384` (cit. on pp. 19, 40).

[80] Shajun Nisha, Mohamed Sathik, and Nagoor Meeral. «A Comprehensive analysis of edge detectors in SD-OCT Images For Glaucoma Diagnosis». In: 6 (Mar. 2019) (cit. on p. 19).

[81] Perivar Nagar. «Number Plate Detection of Vehicle using Modified Kirsch Compass Kernel Edge Detection». In: (2017) (cit. on p. 20).

[82] MT Gopalakrishna. «Mammogram Image Enhancement Techniques for Detecting Breast Cancer: A Critical Review». In: Jan. 2013 (cit. on p. 20).

[83] Kaur EJ. Kaur K Kaur EI. «A Survey on Image Enhancement by Histogram equalization Methods». In: 2016 (cit. on p. 21).

[84] Garima Vyas. «Advanced Image Enhancement Based on Wavelet and Histogram Equalization for Medical Images». In: *IOSR Journal of Electronics and Communication Engineering (IOSRJECE)* Volume 2 (Sept. 2012), pp. 12–16 (cit. on p. 21).

[85] Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew Beam, Irene Chen, and Rajesh Ranganath. «A Review of Challenges and Opportunities in Machine Learning for Health». In: *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science* 2020 (May 2020), pp. 191–200 (cit. on p. 22).

[86] Joel Quintanilla, Benjamín Ojeda-Magaña, Guadalupe Cortina-Januchs, Rubén Ruelas, Antonio Vega-Corona, and Diego Andina. «Image segmentation by fuzzy and possibilistic clustering algorithms for the identification of microcalcifications». In: *Scientia Iranica* 18 (June 2011), pp. 580–589. DOI: 10.1016/j.scient.2011.04.009 (cit. on pp. 22, 25, 29).

[87] Amir Hossein Jafarian, Melika Kooshki forooshani, Farzane Farzad, and Nema Mohamadian Roshan. «The Relationship Between Fibroblastic Growth Factor Receptor-1 (FGFR1) Gene Amplification in Triple Negative Breast Carcinomas and Clinicopathological Prognostic Factors». In: *Iranian Journal of Pathology* 14.4 (2019), pp. 299–304. ISSN: 1735-5303. DOI: 10.30699/ijp.2019.96713.1952. eprint: http://ijp.iranpath.org/article_36631_820ee301199347039e370ef20128dbfd.pdf. URL: http://ijp.iranpath.org/article_36631.html (cit. on p. 22).

[88] Mohamed Roushdy. «Comparative study of edge detection algorithms applying on the grayscale noisy image using morphological filter». In: *ICGST, International Journal of Graphics, Vision, and Image Processing GVIP* 6 (Dec. 2006), pp. 17–23 (cit. on pp. 25, 27).

[89] Ahmed Abdullah, Majid Ali, and Hasan Jawad. «Evaluation of the Performance of Kirsch and Robinson Methods on Different Color Spaces». In: *International journal of simulation: systems, science and technology* (Aug. 2019) (cit. on p. 27).

[90] Kumaratharan N Venmathi AR Ganesh EN. «Kirsch Compass Kernel Edge Detection Algorithm for Micro Calcification Clusters in Mammograms». In: *Middle East Journal of Scientific Research* (Jan. 2016), pp. 1530–1535 (cit. on pp. 27, 28).

[91] Shajun Nisha, Mohamed Sathik, and Nagoor Meeral. «A Comprehensive analysis of edge detectors in SDOCT Images For Glaucoma Diagnosis». In: (Mar. 2019) (cit. on p. 27).

[92] Yuanmin Xie, Lichuan Ning, Mingqi Wang, and Chengcheng Li. «Image Enhancement Based on Histogram Equalization». In: *Journal of Physics: Conference Series* 1314 (Oct. 2019), p. 012161. DOI: `10.1088/1742-6596/1314/1/012161` (cit. on p. 28).

[93] Yogita Dubey and Milind Mushrif. «FCM Clustering Algorithms for Segmentation of Brain MR Images». In: *Advances in Fuzzy Systems* 2016 (Mar. 2016), pp. 1–14. DOI: `10.1155/2016/3406406` (cit. on p. 29).

[94] Satyajit Mondal and Joydeep Mukherjee. «Image Similarity Measurement using Region Props, Color and Texture: An Approach». In: *International Journal of Computer Applications* 121 (July 2015), pp. 23–26. DOI: `10.5120/21833-5092` (cit. on p. 30).

[95] Awni Hammouri, Mustafa Hammad, Mohammad Alnabhan, and Fatima Alsarayrah. «Software Bug Prediction using Machine Learning Approach». In: *International Journal of Advanced Computer Science and Applications* 9 (Jan. 2018). DOI: `10.14569/IJACSA.2018.090212` (cit. on p. 32).

[96] A. Luque, Alejandro Carrasco, Alejandro Martín, and Ana de Las Heras. «The impact of class imbalance in classification performance metrics based on the binary confusion matrix». In: *Pattern Recognition* 91 (Feb. 2019), p. 6829. DOI: `10.1016/j.patcog.2019.02.023` (cit. on p. 32).

[97] Kansal M Thakur I. «A Review on Noise Reduction from Medical Images». In: 2016 (cit. on p. 34).

[98] Stephen Moore, David Maffitt, Kirk Smith, Justin Kirby, Kenneth Clark, John Freymann, Bruce Vendt, Lawrence Tarbox, and F. Prior. «De-identification of Medical Images with Retention of Scientific Research Value». In: *Radiographics : a review publication of the Radiological Society of North America, Inc* 35 (May 2015), pp. 727–35. DOI: `10.1148/rg.2015140244` (cit. on p. 34).

[99] K. Somasundaram and T. Kalaiselvi. «A Method for Filling Holes in Objects of Medical Images Using Region Labeling and Run Length Encoding Schemes». In: 2010 (cit. on p. 36).

[100] A.M Raid, Wael Khedr, Mohamed El-dosuky, and Mona Aoud. «Image Restoration Based on Morphological Operations». In: *International Journal of Computer Science, Engineering and Information Technology* 4 (July 2014), pp. 9–21. DOI: `10.5121/ijcseit.2014.4302` (cit. on p. 38).

[101] N. Senthilkumaran and J. Thimmiaraja. «An Illustrative Analysis of Mathematical Morphology Operations for MRI Brain Images». In: 2014 (cit. on p. 38).

[102] H. Hassanpour, Najmeh Samadiani, and Seyed Mahdi Salehi. «Using morphological transforms to enhance the contrast of medical images». In: *Egyptian Journal of Radiology and Nuclear Medicine* 46 (Feb. 2015), pp. 481–489. DOI: `10.1016/j.ejrnm.2015.01.004` (cit. on p. 39).

[103] P. Suganya, S Gayathri, and Mohanapriya Natesan. «Survey on Image Enhancement Techniques». In: *International Journal of Computer Applications Technology and Research* (Sept. 2013), pp. 623–627. DOI: `10.7753/IJCATR0205.1019` (cit. on p. 42).

[104] Deepali Aneja and Tarun Rawat. «Fuzzy Clustering Algorithms for Effective Medical Image Segmentation». In: *International Journal of Intelligent Systems and Applications* 5 (Oct. 2013), pp. 55–61. DOI: `10.5815/ijisa.2013.11.06` (cit. on p. 43).

[105] «Analysis of Mammographic Images for Early Detection of Breast Cancer Using Machine Learning Techniques». In: 2015 (cit. on p. 44).

[106] Aleksandr Zotin, Konstantin Simonov, Mikhail Kurako, Yousif Hamad, and Svetlana Kirillova. «Edge detection in MRI brain tumor images based on fuzzy C-means clustering». In: *Procedia Computer Science* 126 (Jan. 2018), pp. 1261–1270. DOI: `10.1016/j.procs.2018.08.069` (cit. on p. 45).

[107] Naser Safdarian and Mohammadreza Hedyehzadeh. «Detection and Classification of Breast Cancer in Mammography Images Using Pattern Recognition Methods». In: *Multidisciplinary Cancer Investigation* 3 (Oct. 2019), pp. 13–24. DOI: `10.30699/acadpub.mci.3.4.13` (cit. on p. 46).

[108] Lifeng He, Yuyan Chao, Kenji Suzuki, and Kesheng Wu. «Fast connected-component labeling». In: *Pattern Recognition* 42.9 (2009), pp. 1977–1987. ISSN: 0031-3203. DOI: `https://doi.org/10.1016/j.patcog.2008.10.013` (cit. on p. 47).

[109] Lifeng He, Xiwei Ren, Qihang Gao, Xiao Zhao, Bin Yao, and Yuyan Chao. «The Connected-Component Labeling Problem: A Review of State-of-the-Art Algorithms». In: *Pattern Recognition* 70 (Apr. 2017). DOI: `10.1016/j.patcog.2017.04.018` (cit. on p. 47).

[110] A. Jalalian, S. Mashohor, R. Mahmud, B. Karasfi, M. Saripan, and A. Ramli. «Foundation and methodologies in computer-aided diagnosis systems for breast cancer detection». In: *EXCLI Journal* 16 (2017), pp. 113–137 (cit. on p. 53).

[111] Yu Zhang, Noriko Tomuro, Jacob Furst, and Daniela Raicu. «Image Enhancement and Edge-based Mass Segmentation in Mammogram». In: *Progress in Biomedical Optics and Imaging - Proceedings of SPIE* 7623 (Mar. 2010). DOI: 10.1117/12.844492 (cit. on p. 55).

[112] Mohammed y kamil and Ali Salih. «Mammography Images Segmentation via Fuzzy C-mean and K-mean». In: *International Journal of Intelligent Engineering and Systems* 12 (Jan. 2019), pp. 22–29. DOI: 10.22266/ijies2019.0228.03 (cit. on p. 57).

[113] S. Kumar, A. Fred, and P. S. Varghese. «Suspicious Lesion Segmentation on Brain, Mammograms and Breast MR Images Using New Optimized Spatial Feature Based Super-Pixel Fuzzy C-Means Clustering». In: *Journal of Digital Imaging* 32 (2018), pp. 322–335 (cit. on p. 57).

[114] Diyar Zeebaree, Habibollah Haron, Adnan Mohsin Abdulazeez, and Dilovan Zebari. «Machine learning and Region Growing for Breast Cancer Segmentation». In: Apr. 2019, pp. 88–93. DOI: 10.1109/ICOASE.2019.8723832 (cit. on p. 58).