SAKARYA UNIVERSITY
INSTITUTE OF SCIENCE AND TECHNOLOGY

AN ANALYSIS OF MAMMOGRAM IMAGES FOR BREAST CANCER PREDICTION USING DATA MINING TECHNIQUES

M.Sc. THESIS

Mohammed I. F. MANSOUR

Department	**:**	**MECHATRONICS ENGINEERING**
Field of Science	**:**	**MECHATRONICS ENGINEERING**
Supervisor	**:**	**Assist. Prof. Mustafa KUTLU**

July 2020

SAKARYA UNIVERSITY
INSTITUTE OF SCIENCE AND TECHNOLOGY

AN ANALYSIS OF MAMMOGRAM IMAGES FOR BREAST CANCER PREDICTION USING DATA MINING TECHNIQUES

M.Sc. THESIS

Mohammed I. F. MANSOUR

Department	**:**	**MECHATRONICS ENGINEERING**
Field of Science	**:**	**MECHATRONICS ENGINEERING**
Supervisor	**:**	**Assist. Prof. Mustafa KUTLU**

This thesis has been accepted unanimously / with majority of votes by the examination committee on 29.07.2020

Assist. Prof. Mustafa KUTLU	**Assist. Prof. Abdellatif BABA**	**Assist. Prof. Serap KAZAN**
Head of Jury	**Jury Member**	**Jury Member**

DECLERATION

I declare that all the data in this thesis was obtained by myself in academic rules, all visual and written information and results were presented in accordance with academic and ethical rules, there is no distortion in the presented data, in case of utilizing other people's works they were refereed properly to scientific norms, the data presented in this thesis has not been used in any other thesis in this university or in any other university.

Mohammed MANSOUR

22.06.2020

ACKNOWLEDGMENTS

In the Name of Allah Most Gracious Most Merciful. First, I give thanks to God for protection and ability to do work. I give deep thanks to the Professors and Lecturers of the Faculty of Technology and other workers of Sakarya University. My special and heartily thanks to my supervisor, Assistant Professor Mustafa KUTLU who encouraged and directed me. With his supervision that this work came into existence.

I also thank my family and friends who encouraged me and prayed for me throughout the time of my research. I would like to thank my parents Issam Mansour and Amira Mansour; whose love and guidance are with me in whatever I pursue. They are the ultimate role models. I would like to thank my loving brothers Mansour and Abd AlNaser and my sister Mais. May the Almighty God richly bless all of you.

TABLE OF CONTENTS

CHAPTER 3.

CHAPTER 4.

CHAPTER 5.

CHAPTER 6.

LIST OF SYMBOLS AND ABBREVIATIONS

AI	: Artificial Intelligence
ANN	: Artificial Neural Network
ARM	: Association Rule Mining
BC	: Beast Cancer
CART	: Classification and Regression Trees
DI	: Digital Image
DIP	: Digital Image Processing
DM	: Data Mining
DT	: Decision Tree
FCM	: Fuzzy C Means
FN	: False Negative
FP	: False Positive
FPR	: False Positive Rate
GRNN	: General Regression Neural Network
ID	: Iterative Dichotomize
KDD	: Knowledge Discovery Dataset
KM	: K Means
KNN	: K Nearest Neighbor
MAFIA	: Maximal Frequent Item-set Algorithm
MIAS	: Mammogram Image Association Society
ML	: Machine Learning
NB	: Naive Bayes
PNN	: Probabilistic Neural Network
PSO	: Particle Swarm Optimization
ROI	: Region of Interest
SEER	: Surveillance Epidemiology End Result

SNN	: Statistical Neural Network
SOM	: Self Organizing Map
SVM	: Support Vector Machine
TN	: True Negative
TP	: True Positive
TPR	: Ture Positive Rate

LIST OF FIGURES

LIST OF TABLES

SUMMARY

Keywords: Data Mining, Digital Image, Algorithm, Clustering, Classification, Breast Cancer, Mammogram

Breast malignant cancer is one of the most dangerous diseases that women suffer from. In this research, Data Mining (DM) with machine learning (ML) and its various techniques were applied for processing the Breast Cancer (BC) digital images. MIAS dataset images were used for analysis and prediction of cancer diseases; Mammography Image Analysis Society (British research groups organization http://peipa.essex.ac.uk/).

This work analyzes breast cancer images in three main stages; preparing images using digital image processing tools, then using clustering techniques for segmentation of mammogram images and extracting the affected area and using classification techniques for cancer data classification. Before using DM techniques, digital image processing involving image enhancement techniques were applied to the images. Digital image processing represents functions and techniques which aim to improve the quality of images and prepare the data for the next processes. Preparing data is a very important stage in DM since it removes the unwanted details from data. Segmentation of BC images using clustering techniques mainly K Means (KM) and Fuzzy C Means (FCM) was achieved for detecting the abnormal region in the images based on the intensity of pixels. The algorithms were implemented in MATLAB for analysis. During these implementations, the used clustering techniques' performances were compared. Three parameters were considered; run time, a number of clusters, and memory space used for saving and storing the clustered results images. Both algorithms gave proved significant results. The run time of KM was three time less than FCM but memory space of FCM clustered images results was two time less than KM. Four images were clustered by FCM and five images were clustered by KM. For more checking and evaluating the performances of clustering algorithms' results; classification algorithms were used for classifying of extracted data from the clustered images and other BC data. The classification technique was used for categorical class label prediction of cancer disease. The main attributes for classification where the number of pixels representing cancer affected area which was found and extracted by clustering techniques. Six attributes were given to the classification algorithms. Classification algorithms; Artificial Neural Network (ANN), K Nearest Neighbor (KNN), and Support Vector Machine (SVM) were used for classification of BC data and prediction of cancer possibility. The highest accuracy was found using ANN (97%), followed by KNN (94%) and SVM (52%) in the last.

VERİ MADENCİLİĞİ TEKNİKLERİ İLE MEME KANSERİ TAHMİNİ İÇİN MAMMOGRAM GÖRÜNTÜLERİNİN ANALİZİ

ÖZET

Anahtar kelimeler: Veri Madenciliği, Dijital Görüntü, Kümeleme, Sınıflandırma, Meme Kanseri, Mamografi

En tehlikeli hastalıklardan biri, genellikle kadınların mağdur olduğu kötü huylu meme kanseridir. Farklı tekniklerle veri madenciliği; makine öğrenmesi ve algoritmaları yardımı ile kümeleme ve sınıflandırma için meme kanseri dijital görüntülerine uygulanmaktadır. MIAS (Mammography Image Analysis Society) veri kümesi görüntüleri, kanserin analizi ve tahmini için kullanılmıştır. Bu çalışma meme kanserini üç ana aşamada analiz edecektir; dijital görüntü işleme araçlarını kullanarak görüntülerin hazırlanması, daha sonra mammogram görüntülerinin bölütlemesi için kümeleme tekniklerinin kullanılması, kanser sınıflandırma tekniklerinin kullanılması için etkilenen bölgenin çıkarılması. Piksel yoğunluğuna bağlı olarak görüntülerdeki anormal bölgenin saptanması için esas olarak K Means (KM) ve Fuzzy C Means (FCM) kümeleme teknikleri kullanılarak meme kanseri görüntülerinin bölümlere ayrılması sağlanmıştır. Analiz için algoritmalar MATLAB'da geliştirilmiştir. Bu uygulamalar sırasında kullanılan kümeleme tekniklerinin performansları karşılaştırılmıştır. Temel olarak üç parametre dikkate alınmıştır; çalışma süresi, küme sayısı, kümelenmiş sonuç görüntülerinin kaydedilmesi ve saklanması için kullanılan bellek alanı. Her iki algoritma da önemli sonuçlar vermiştir. KM'nin çalışma süresi FCM'den daha az, ancak FCM kümelenmiş görüntülerin sonuçlarının bellek alanı KM'den daha azdır. Dört görüntü FCM tarafından kümelenmiş ve beş görüntü KM tarafından kümelenmiştir. Kümeleme algoritmalarının sonuçlarının performanslarının daha fazla denetlenmesi ve değerlendirilmesi için; kümelenmiş görüntülerden ve diğer meme kanseri verilerinden çıkarılan verilerin sınıflandırılması için sınıflandırma algoritmaları kullanılmıştır. Sınıflandırma tekniği, kanser hastalığının kategorik sınıf etiketi tahmini için kullanılmıştır. Sınıflandırma için ana özellikleri kanserden etkilenen piksel sayısı alanı temsil eder. Bu nedenle, sınıflandırma algoritmalarına altı öznitelik atandı. Kanser verilerinin sınıflandırılması ve kanser olasılığının öngörülmesi için Support Vector Machine (SVM), K Nearest Neighbor (KNN) ve Artificial Neural Network (ANN) gibi sınıflandırma algoritmaları kullanılmıştır. Kanser verilerinin sınıflandırılmasında en yüksek doğruluk ANN (97%), ardından KNN (94%) ve son olarak SVM (52%) kullanılarak bulunmuştur.

CHAPTER 1. INTRODUCTION

Data mining (DM) is defined as the process that concerns in extracting useful patterns and knowledge from the available set of data. Extracting comprehendible and unknown useful information was a challenging topic for many researchers and businesses [1]. Developing technology raised the sizes of databases which demand new methods to deal with and finding ways to get benefits from available sources [2]. In some applications it’s known as the knowledge discovery process, other known terms like knowledge mining or knowledge extraction are used all for defining DM. The researchers Pratiyush and Manu reviewed knowledge discovery in DM and various domains of the data gathered perspective, also they explained that the DM is using a portion of the applications of Knowledge Discovery Database (KDD) [3]. Machine Learning (ML) and its different techniques and approaches methods like Artificial Neural Network (ANN) and Genetic Algorithms give data analyzing science a huge advantage of analyzing and predicting more new information from datasets. Understanding large and complex information improve become an important issue in the field of business and science. Application sectors of DM are very wide and large; from data analysis to data discovery. Application of DM is found in different fields like space sciences, robotics and medical data applications [4,5].

Breast Cancer (BC) diseases are dangerous and are from the most diseases leading to death. Many women in every country of this world are suffering from breast cancer. BC is known as a dangerous disease with complex treatment ways and law possibilities of staying alive unless early discovering and treatment is applied to the woman. BC can be explored by self-examination, clinical breast examination, and breast diagnosis tests via medical imaging. Researches in BC have grown up quickly in the last decades due to many reasons. Within the improvement of digital data processing and analyzing which give chance to use these all technologies for investigating the BC diseases. The segmentation of BC is a required mission in medicine. Detecting cancer in early stages can decrease the possibility of dangerous but detecting cancer at an advanced level requires complex operation with the law possibility of staying alive. DM techniques give the ability to apply all these new technologies to analyze and predict breast cancer diseases. DM with different Digital Image Processing (DIP) can be applied to cluster breast images, find the affected region in the breast and build a system that may help interested people and radiologists to predict different cancer diseases. In this research, different digital image functions and processing methods were applied to mammogram digital images. Clustering algorithms; K Means (KM) and then Fuzzy C Means (FCM) were applied to cluster the mammography Digital Images (DI) and extract the affected regions. For more analyzing and checking the performance of clustering algorithms; different classification techniques from ML techniques were applied to classify the cancer data.

1.1. The Problem Statement

The problem statements of the thesis are as follow:

1. The lack of quality of breast cancer digital images.
2. The need for the analysis of breast cancer cells through segmentation.
3. The spread of BC in light of the unavailability of prediction and analysis.
4. The inadequate performance of various clustering algorithms.

1.2. Scope of The Thesis

This thesis mains to investigate the different performance of clustering and classification algorithms. Analyzing the mammography images of the breast through DM steps. The main objective of this research is to analyze two clustering algorithms and to verify their accuracies by using different classification techniques. The design of an excellent system should be implemented and analyzed by clustering the image with a minimum number of clusters. Run time and space memory of running of clustering algorithms were used to compare different clustering algorithms' results. Classification techniques were used for categorical prediction of cancer possibility for cancer data.

The proposed method has three main stages, preprocessing or preparing of the mammogram image, image segmentation through clustering, and then the classification of cancer data. For a better analysis of the implemented clustering algorithms; the classification process was applied to check the quality of clustering algorithms which is the last stage of the whole implemented system. Each step or stage in this research has its own inner steps and methods; where the general workflows through them.

The work was implemented and analyzed in MATLAB using mammograms MIAS images dataset; Mammography Image Analysis Society (British research groups organization http://peipa.essex.ac.uk/) [6]. Other programs; Microsoft Office Excel was used for analyzing the result of clustering and classification algorithms.

1.3. Objectives

The main objectives of the work are summarized as follow:

1. To use different DIP tools for preparing of mammogram images.
2. To segment the cancer area in mammogram digital image (DI).
3. To predict cancer possibility of cancer data using different classification algorithms.
4. To check the performance of different clustering and classification algorithms through their qualities.

1.4. Thesis Layout

This thesis consists of seven chapters. Chapter1 introduced the thesis including the problem statement, scope and the main objectives. Chapter 2 detailed background related to DM, DIP and BC. Chapter 3 provided a literature review from different researches related to DM techniques applied to mammograms and BC data. Chapter 4 discussed research methodology, the preparation of data using DIP tools and functions. Chapter 5 reviewed both clustering and classification algorithms applied to mammograms images for segmenting of mammogram images and predicting of cancer probability. Chapter 6 discussed the result of applied DM techniques; clustering and classification algorithms. Finally, Chapter 7 concluded the thesis, summarized the gathered result and draw future work.

CHAPTER 2. TECHNICAL AND MEDICAL BACKGROUND

This chapter provides an overview related to DM, DIP and BC. DM has been discovered and employed in many information mining systems. In this chapter, main stages and techniques of the DM were summarized. Then the chapter goes into DIP and the importance of processing of DM. A variety of types of DI and formats of DI were given. The last part of the chapter gave an overview about breast anatomy and BC. Different diagnoses tests for BC were summarized as well.

2.1. Introduction to DM

DM is defined as the process which concern in extracting useful patterns and knowledge from available set of data. The new discovered knowledge can be used and employed in different fields for making new systems and business. According to [7], DM has three steps of trials summarized in these three terms; exploration, pattern identification and deployment. The exploration stage clean and transform data to another format. In this stage also; the type of data and important patterns related to problem are specified. In the second stage; important pattern identification is done. This stage aims to specify patterns that help for making good prediction. In the last stage; the deployment of patterns process takes place for the desired planned outcome.

DM is known that it is a critical process in the extraction of knowledge from datasets. It takes importance in many domains. DM is an important and critical section of KDD. Some researchers treat DM as KDD since it uses the same process tools as KDD. KDD goes through different stages, including preparation, selection, and cleansing of data, as well as including prior knowledge and interpretation of results and outcomes [8]. The steps are summarized in Figure 2.1. [9]. According to [9], some of basic steps of KDD are summarized in the following steps:

1. Work on and understanding the domain.
2. Building the needed dataset.
3. Filter and processing the needed data.
4. Filtering data.
5. Matching the objectives of the KDD process.
6. Choosing the DM algorithm.
7. Applying the DM algorithm.
8. Interpreting patterns existing in the mind.
9. Amalgamating available knowledge.

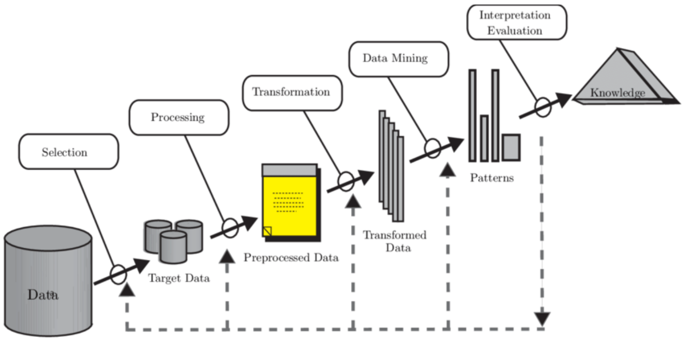


Figure 2.1. Representation of the steps comprising tile KDD process [9]

Before applying DM, pre-processing or preparing of data should be applied to make the data ready for the next stage or required purpose. Data preparing is known as it is an important process since it removes the unimportant and unwanted component from the data. The application of pre- processing varies depending on the dataset used. Data preprocessing is known as an important stage in DM deals mainly with preparation of the data, then transform the data to be used in the main stages or approaches of DM techniques. Data preprocessing aims to remove unwanted information available in the dataset.

In this research; the used data are digital mammogram images. Pre-processing or preparing of images should be applied to make them ready for main data mining processing stages. It is possible to say preparing of data or image is done by different methods and for different aims. These aims can be for improving of the quality of images or it can be just for feature extractions. For dataset of digital images; some of methods can be summarized here. Image preprocessing can be done using some methods such as image re-sampling, gray scale contrast enhancement, noise removal and some mathematical and morphological operations. In image re-sampling the numbers of pixels in the dataset are increased or reduced. In gray scale contrast, enhancement visualization is improved by brightening the dataset. The aim of these preprocessing operation is to improve the quality of the images in the dataset. As a result, making them better and ready for the feature extraction or clustering process.

After preparing of data for the required aim which mainly means to get new discoveries from the preprocessed data, DM different application of modeling techniques can be applied. This may involve Artificial Intelligence different techniques like Machine Learning (ML) and Probabilistic Reasoning. Other approaches for DM are statistical approaches and database-oriented approaches. In this proposed work, ML approaches represented by clustering will be appropriate for analyzation of digital mammogram images and classification technique for analyzation of BC data. In this research; mainly just two type of DM techniques based on ML techniques were employed. These are clustering and classification.

2.2. Introduction to DIP

Digital image is an array, its elements are arranged in columns and rows, each element of them is called a Pixel. The location of each pixel is determined by its own coordinates (x, y). One value; integer, double or logical has to be attributed to each pixel in order to determine its color for colored image or its gray level for gray scale image.

The aim of DIP can be divided into three main classes; image processing for improving its quality (image adjustment, image enhancement, image filtering, image transformations and image compression), image analysis (region of interest operations, binary operations and image statistics) and image understanding (image clustering, image classification, image registration, image indexing and target identification).

2.2.1. Types of DI

There are three types of DI:

1. Binary image: this type of image takes just two color black or white (zero or one).
2. Gray image: it contains gray level details. The pixels numbers specify the gray level.
3. Colored Image: color images contains three bands of colors; blue, red and green.

2.2.2. Creation of DI

The creation of digital image is done through two main steps; sampling and quantization. The sampling step divides an image into pixels each of them has its coordinates x and y. The Quantization step means to attribute a given intensity; gray scale or a given color to the pixel.

2.2.3. Image quality

Images qualities depend on dots per inch (dpi) which refers to the spatial resolution and the number of bits that code the pixel's intensity (Bit resolution). For example, the number of possible intensities is given: $L = 2^n$ where n represents the number of bits (8 or 16. So, L = 256 or L= 65536. Thus: [0 to 255] or [0 to 65535]. And due to these rezones; better quality for images implies more dots per inch and better quality implies more bits to code the pixel intensity.

2.3. Breast Cancer

Breasts are located in the upper ventral region in human at both sides of the stem. Each breast includes part of frontal area of a human body from the beginning of the second rib to sixth which contain the mammary gland. An illustration of the anatomy of the breast if given in Figure 2.2.

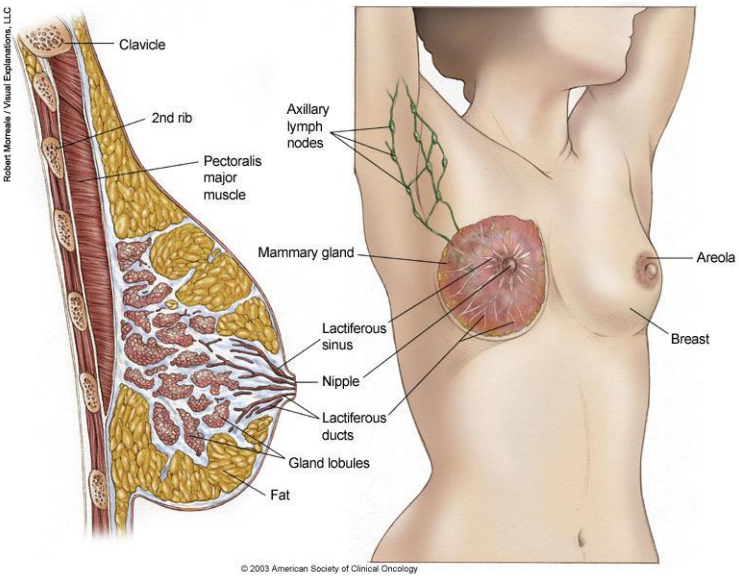


Figure 2.2. A Cross-sectional view of breast [10]

When looking at female breasts, one could spot fatty, fibrous, gland tissues, as well as ducts, nerves, and blood vessels. Lobes are small parts in the breast, amounting to 15 to 20 pieces, and made of lobules. Lactiferous ducts and alveoli are the main components of lobules. Lactiferous ducts expand in size forming small lactiferous sinuses, which are responsible for accumulating milk during lactation. Milk could be obtained through some holes in the nipples. The lobes are connected through the fibrous tissue which could be spotted on the entire surface of the breast. The fatty tissue, which is usually abundant covers the gland and determines its size, except for the areola [11][12].

The normal body cells are growing up and dividing for a specific duration of time and stop after this process. In cancer diseases cases, the body's cells continue produced cells and divided unregularly. An illustration for these cases is shown in the Figure 2.3. For medical cases, cancer is defined as malignant tumefaction. It is quite possible to move the malignant cancer to the other parts of the body through the circulation of blood and any other lymphatic action. The two types of neoplasm are benign and malignant. Benign cells do not spread uncontrollably. In the malignant neoplasm cases, cells are growing quickly and spreading to other part of the body which is known as Metastasis.

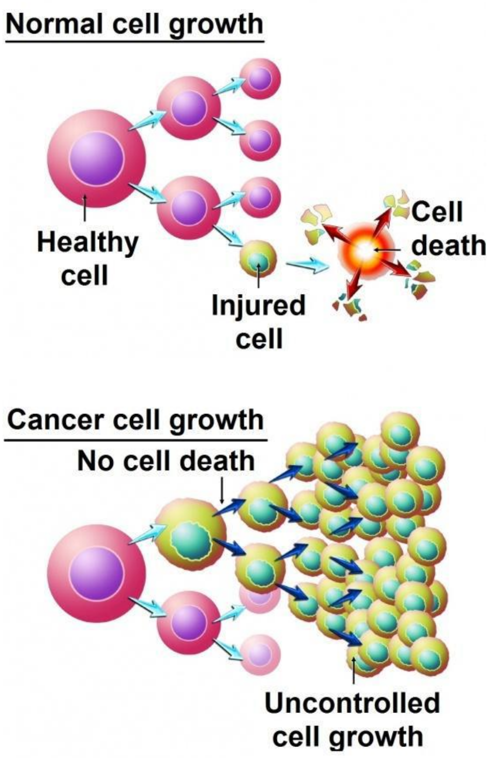


Figure 2.3. Normal and cancerous cell growth [10]

Female breast includes 15 - 20 lobes which are linked to together and terminated in the nipple through a complex structure of interrelated ducts. Each lobule has 10 - 100 terminal duct lobular unit (TDLU). The breast cancer originates in this unit. Cancer is known as in situ when the tumor has no spread via the basal membrane and only made of the lobules of the ducts. As when the cancer has broken through the basal, it is called INVASIVE, and chances on metastases increase sharply [13].

The main two main types of cancerous tumors are either malignant or benign, the characteristics of each play a main role to know which type the tumor belongs to. For instance, malignant tumor tends to be in unorganized shape with unclear borders (speculation at the boundaries), while benign frequently shows in regular shape (ovoid, spherical) with sharp borders. calcifications that are created breast cancer may appear in granular or crushed stone, casting or linear, powderish or amorphous. Breast cancer lesions are Calcification, Micro- calcification, tumor and masses. According to [14] Table 2.1. below summarizes both types of Neoplasms.

Determining the specific cause of the cancer is somehow complex because there are many known factors that increase the risk of this disease such as smoking, obesity, environmental pollution, some infections, genetic causes and lack of physical activity. These factors can cause the disease by changing the activity of body cells or damage genes indirectly or directly. There are more than 190 different known of the cancer.

Table 2.1. Neoplasms types [14]

Benign Neoplasms	Malignant Neoplasms
Moving mass	Fixed mass
Soft and clear round with besetment fibrous capsule.	Irregular shaped with no capsule
Cells multiply slowly.	Cells multiply rapidly.
The tumor is growing by expanding and pushing away and against surrounding tissue.	The tumor growth by invading and destroying surrounding tissue.
Mass is moving and not linked with surrounding tissue.	Mass is settled and linked with surrounding tissue and fixed in surrounding tissue pointedly
Never diffuses to remote parts	Always diffuses to other remote parts
Remove it easy and does not recur again	Remove it more difficult and may recur again

For BC there are several risk factors; lack of procreation or breastfeed, disorders in hormones levels, regimen and corpulence, lifestyle such as smoking or lack of physical activity, mutations and inheritance, etc. But not every time existence of these factors means that the woman will get cancer or not and this fully applies to all cancer types. For example, there are incidences among women without having risk factors except being woman and being older.

2.3.1. Breast cancer diagnoses tests

There are four types of different technologies for diagnoses and detecting of BC; Magnetic Resonance Imaging (MRI), Microwave Tomography, Breast Ultrasound and Mammography. The most used technique for detecting of cancer is Mammogram. In this research; digital mammogram images were used for the analysis. The four types of these technologies are described as follow:

1. MRI: this technology uses radio and magnet waves. Generally, this type is used for picturing of complex parts of the human body.

2. Microwave Tomography: it uses electric magnetic to find the cancer affected area, it works based on radar system which looks for abnormalities in the breast.

3. Breast Ultrasound: it uses waves of sound for creating images of the breast. This technique is generally used in obstetric systems due to its safety.

4. Mammography: it uses x rays for creating medical images. This type is the most used technology for diagnosing of breast cancer.

CHAPTER 3. LITERATURE REVIEW

This chapter reviewed various researches and works which had been done by various researchers related to classification and analysis of medical and breast cancer data and segmentation of breast cancer images. DM techniques are used for descriptive analysis and prediction of knowledge from an existing huge set of data. DM is used for visualization of data, association, pattern generation, clustering and classification from large set of data in different field of applications. The other known term for DM is KDD. A variety of DM models are available in the literature such as educational DM, environmental DM, web DM, image DM, medical DM and social DM. This review mainly covers the basics of medical data, breast cancer analyses, clustering of medical image and classification of breast cancer data analysis. Different researches and works are being analyzed by researchers to find abnormal and affected areas in medical images.

The identification and prevention of diabetes disease using of DM with different classification techniques were carried out to predict diseases among patients. Different classification methods like SVM, Apirori, KNN, NB and C4.5 were used in this research. The best highest accuracy result was found using of C4.5 technique. This proposed technique is known and used in various healthcare units around the world [15]. A research related to Decision tree (DT) and SVM classification techniques were proposed by researchers. They approved through their research that these both methods give the highest accuracy of results compared to other methods [16]. Various DT algorithms; classification and regression tree (CART), C4.5 and iterative dichotomize (ID3) were analyzed and compared for classification of tuberculosis patients' responses to treatment under different radiological and bacteriological responses. Their best results were gathered by C4.4 DT algorithm [17].

KM and maximal frequent item – set algorithm (MAFIA) clustering methods were used for analyzed of heart diseases database. The data was preprocessed then clustered using KM technique. The MAFIA was used for mining maximal of the data of heart diseases. Their system lead to a result that their method can predict heart attacks [18]. Diabetes diseases was classified using of SVM. Different classification measure parameter was used for fining of the performance of SVM with radial basis kernel function. The measure parameters were found to be high using of this function with SVM. A conclusion was assumed that the performance of SVM can be improved by using of feature subset selection process [19].

DT was used for analyzing of breast cancer datasets. The method was used for determining of high risk of BC. The proposed system validate that the foundation of statistical associations is possible for detection of BC through DT [20]. Different classification techniques were used for analyzing and extraction of mass disease in mammograms depending of the ages. SVM, Tree Boost ad Tree Forest was used in this research [21]. Swarm optimization classification algorithm was used for analyzing of BC data. The data was preprocessed for the selectin of features firstly using of genetic algorithm. The result of this proposed system was determined effectively [22].

DM with its different techniques were used for prediction of BC survivability in different datasets. The used component techniques in this analysis were statistical neural network (SNN), self-organizing map (SOM), radial bias function network (RBFN), probabilistic neural network (PNN) and general regression neural network (GRNN) [23]. These all techniques were used mainly in a component way for reduction of data. Hybrid algorithms including Fuzzy Decision Tree was used for analyzing of BC. Different DT methods were used like inference techniques and fuzzy membership functions in the proposed system. The hybrid DT was found to have more robust result than normal FDT [24]. SOM and FCM was used for detecting of tumor in mammogram images. Statistical features of tumors were found which can help for studying of statistical information of cancer disease. This research then gave doctors ability to provide better treatment to the patients [25].

A hybrid usage of FCM and SOM were used to cluster medial images. This proposed method could then use for categorizing the affected area in breast images [26]. An identification of the presence of BC calcification and mass in breast images using a KM and FCM research was proposed. This method was done using a combination of both these clustering methods and the result was successfully gathered for determining of BC diseases [27]. Three clustering methods; KM, region level set and KM were used for segmentation of breast thermo gram. The used level set method gave high accuracy and efficiency result compared to other used methods [28].

A new modified KM technique were proposed, and the result were compared with general KM and FCM. The result of the new modified method is better the other compared methods results [29]. Different DM techniques; ANN, association rule mining (ARM) was used for detecting of detection and classification of abnormal region in mammogram images. The result of these techniques was good compared to other methods [30]. A research was done for analyzing of particle swarm optimization (PSO) algorithm for clustering of satellite image and MRI. The gathered results show that the proposed method performs better than other classifiers KM, FCM, K Harmonic Means and genetic algorithms [31].

Different classification methods were used and applied to classify medical data in different domain of applications. Surveillance epidemiology end results (SEER) BC data were classified using different techniques. The aim of this classification was to check out the stage of cancer diseases in the beginning or forward stages. Different classification rules were used to build the proposed method for better and high accuracy of result [32]. Logistic regression technique was used for classification of BC different diseases; mass, architectural distortion and calcification [33]. DT with CART technique was used for analyzation of BC datasets. The proposed methods were checked with feature selection first then without feature selection. It was concluded that the feature selection enhances the result of CART method [34]. Classification methods; C4.5, ID3 and simple CART were used for classifying of BC and predicting of BC diseases. Among the results; it was found that C4.5 is the best to discover the type of tumor [35]. A study for determining of BC masses was proposed for utilizing the morphological operators for segmentation and FCM clustering of images. The results indicate that this system can help doctors to detect breast cancer in the early stages [36]. BC Wisconsin dataset was analyzed using three classification for predicting of breast cancer disease. Three classification were used in this research; KNN, NB and SVM. The highest performance and accuracy were found using of SVM [37].

This chapter reviewed some researches and works done by various researchers related to analysis of medical and BC data, segmentation of BC images and other works for classification of BC data. The research methodologies carried out in this work are discussed in the rest chapters.

CHAPTER 4. RESEARCH METHODOLGY

This chapter discusses the research methodology of the work and the mammograms images preparing. Some information related to dataset used in this research was given as well. This research work analysis mammogram digital images using DM techniques. Clustering algorithms; KM and FCM algorithms were used to segment the mammogram images. Classification techniques were used to check the accuracy of clustering technique's results. These classification algorithms were used to classify the extracted data from the images. The work has three main stages; the first one is preparing of mammogram images using digital images processing tools, the second one is clustering of mammogram images and the last one is classification of cancer data extracted from mammogram images. This research aims to investigate the clustering algorithms in terms of their quality; run time and space complexity and the classification algorithms in terms of accuracy. The main steps of research work are summarized as follow:

1. Preprocessing or preparing of the images using different DIP tools.
2. Apply KM and FCM clustering algorithms on the images.
3. Checking the affected regions and find the number of pixels of the clustered parts from the images.
4. Compare the clustering algorithms; run times and memory spaces, finding the best implemented algorithms.
5. Apply classification algorithms to cancer data in order to check the clustering algorithm performances.
6. Check the accuracy of classification algorithms.

Figure 4.1. gives a summarization of the whole implemented system.

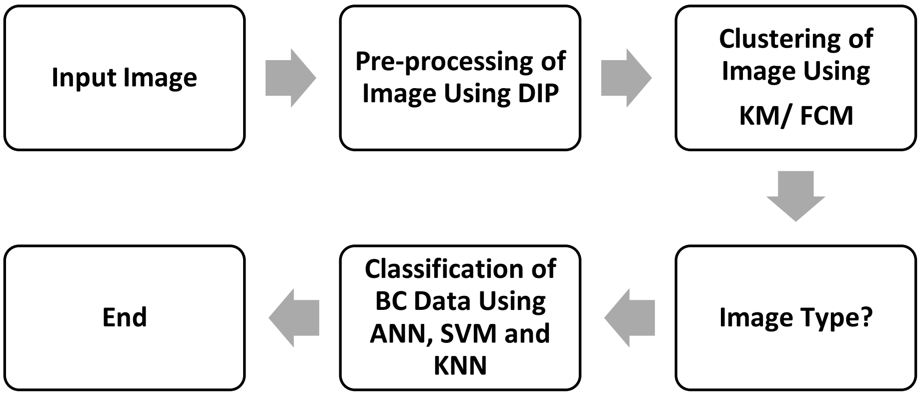

Input Image
Pre-processing of Image Using DIP
Clustering of Image Using KM/ FCM
Image Type?
Classification of BC Data Using ANN, SVM and KNN
End

Figure 4.1. Main steps of the research

4.1. Dataset Used in the Research

The dataset used in this research was taken from Mammography Image Analysis Society; (British research groups organization http://peipa.essex.ac.uk/) [6]. This dataset is known as digital mammography database (MIAS). The data set includes both sides (right and left) of breast images. The dataset has normal, malignant and benign images types. Some information includes the position and the class of abnormality, and the character of the background tissue were given with images.

4.2. Preparing of Mammogram Images

Before using of DM; preparing of data is required. The dataset used in this research is digital images. Some techniques of DIP were used for preparing of data. The possibility of noise and unimportant contents in the background of images is high. Techniques and methods of DIP were used for improving the quality of images.

Preprocessing or preparing of data before using of DM main techniques play an important and essential role since it enhances the data for the upcoming steps (clustering and classification). After changing the format and resizing of images to a standard image size 256*256, techniques such as low pass filters, high pass filters, morphological operations, and some enhancement techniques of histogram were used and applied to images. These are the most methods used in the preparing of mammogram digital images making them transformed to the second main step of the research which is clustering.

In this research; the number of images choose for the work was as follow: 10 normal images, 35 benign images and 35 malignant images. The mammogram digital images used in the research are shown in Figure 4.2., Figure 4.3 and Figure 4. 4..

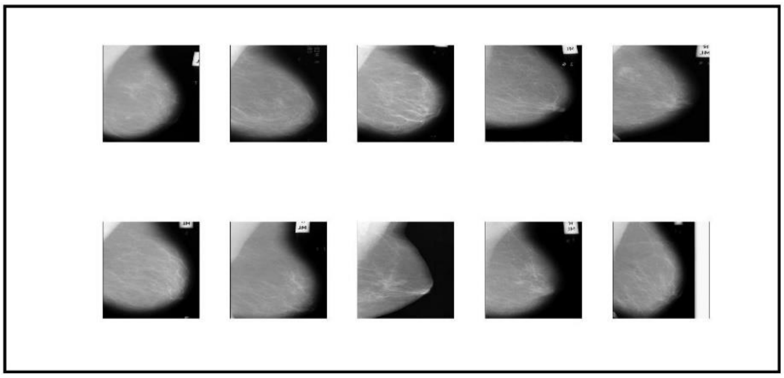


Figure 4.2. Normal images

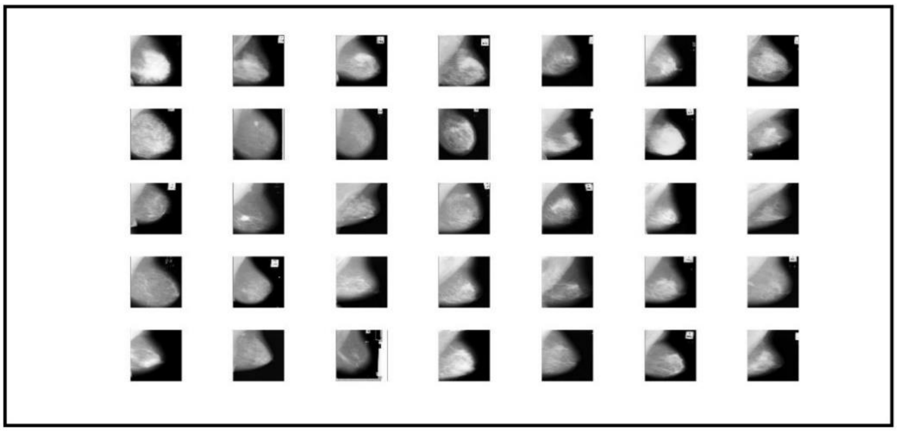


Figure 4.3. Malignant images

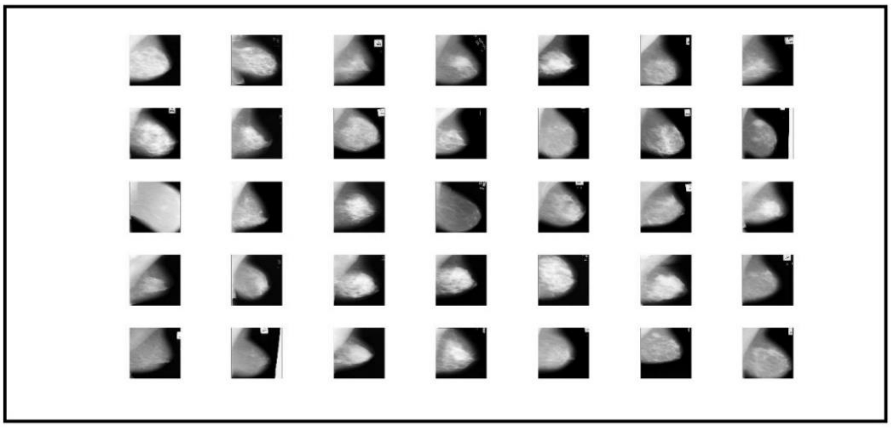


Figure 4.4. Benign images

Figure 4.5. views the pre-processing steps of mammogram digital images using various DIP tools.

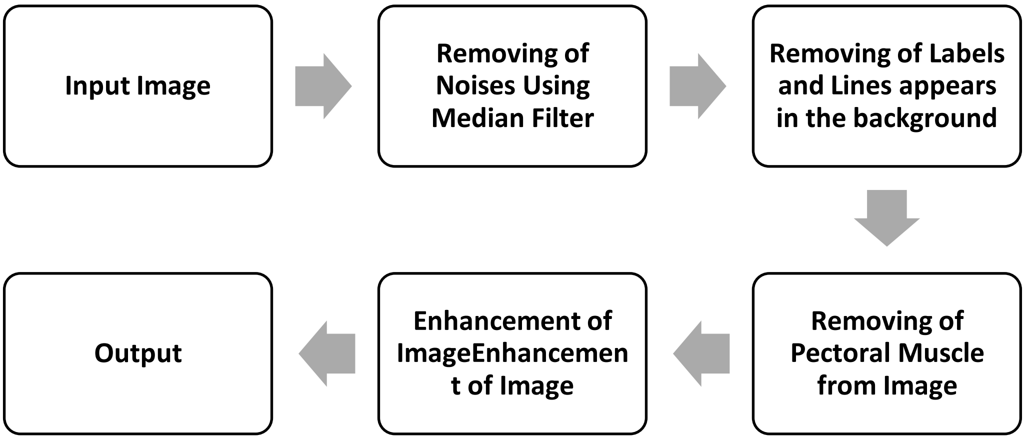

Input Image
Removing of Noises Using Median Filter
Removing of Labels and Lines appears in the background
Removing of Pectoral Muscle from Image
Enhancement of ImageEnhancement of Image
Output

Figure 4.5. Pre-processing steps of mammogram images

4.2.1. Removing of noises using median filter

Removing of unwanted noise is an important issue during the preparing of images and for this affair; median filter technique was used. Some images contain brighter or darker cell values, which is represented as noise and they were removed by median filter. Through it pixels are replaced by the median of their neighbors. This filter keeps the edges preserved as much as possible. Median filter a is non-linear filter technique used to remove or eliminate isolated points or lines from an image without reducing its details and it is an optimal solution to get rid of salt and pepper noise. The usage of this filter in the processing of digital image is a high demand since it preserves the edges during the removing of different noises. Median filter function is given by

$$f(x,y) = Median(s,t)eS(X,Y\{g(s,t)\} \quad (4.1)$$

Where (x, y) is the total number of pixels in the neighborhood (s, t) [38,39]. f is the output image and s is the input image. The results of Median filter applied on normal, malignant and benign images are shown in Figure 4.6., Figure 4.7. and Figure 4.8. respectively.

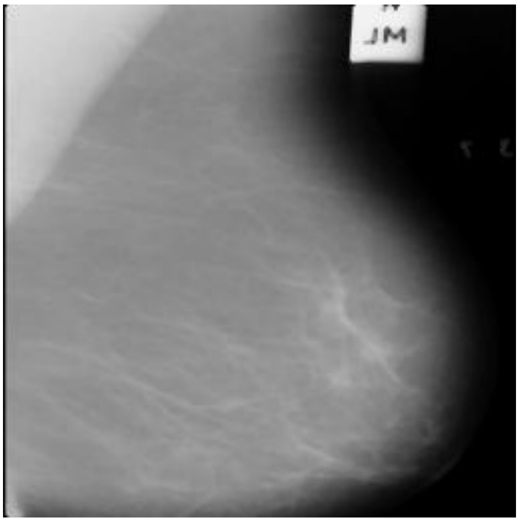


Figure 4.6. Normal image case median filtered

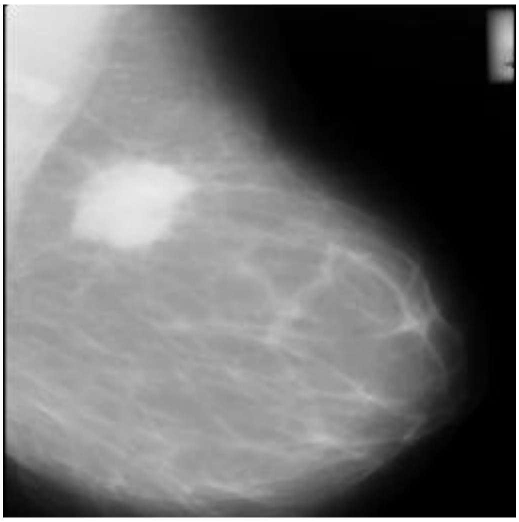


Figure 4.7. Malignant image case median filtered

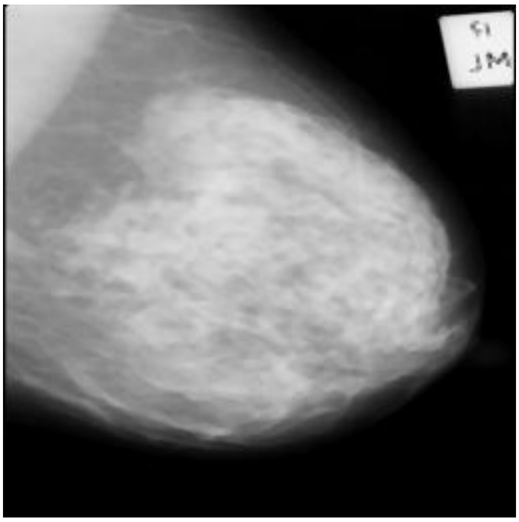


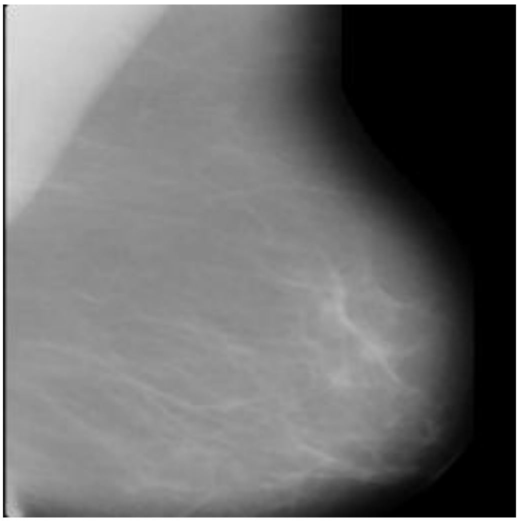
Figure 4.8. Benign image case median filtered

4.2.2. Removing unwanted parts from image using morphological tools

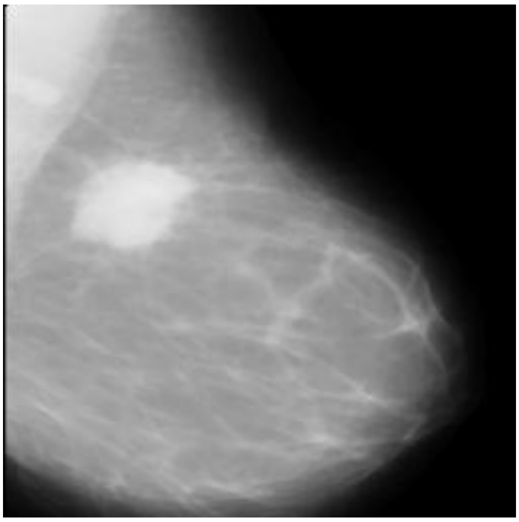
Mammogram images of MIAS dataset have objects like labels and lines appears on the background of most of the images. Morphological operations provide few technical for removing these appearing objects; labels and lines. Dilation and erosion methods were used for removing of these objects. Morphological operation used to space away collected objects; filling in gaps or spaces or rectifying noisy image. These methods use structuring element for adding or deleting of pixels of the objects. This element used to scan the original image and to build the new one. It is an array of zeros and ones, it has an origin point and it may take several shapes (disk, square, diamond, line). Dilation technique add pixels to the boundary of an object and erosion technique erode pixels from the boundary of an object.

Morphological operation has more important techniques which collects both dilation and erosion together and they are applied on images successively. The first one is opening; it is the process of applying the two operations successively (erosion then dilation) on an image to rectify the edges of detected object or to fill the thin links between different shapes in an image. The second one is closing; it is the process of applying the two operations successively (dilation then erosion) on an image to make the thin links in it wider or to fill some gaps.

For the mammogram's images used here, global thresholding was applied to the image. Then opening, closing, erosion and dilation were applied successively to the thresholded versions of images. Extra objects; labels and lines were removed from the backgrounds successively. The results of this process applied to normal, malignant and benign images are shown in Figure 4.9., Figure 4.10. and Figure 4.11. respectively.



4.9. Normal image case label removed



4.10. Malignant image case label removed

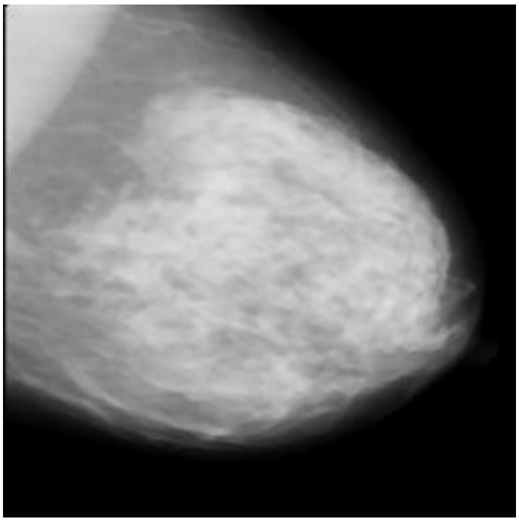


Figure 4.11. Benign image case label removed

4.2.3. Removing muscle part from the images

During taking of mammogram images, sometimes the muscle of breast body appears on the images. Removing of this part is an important need since it affects the segmentation technique. The used method is described in these steps; change the image's types from gray to binary, later scan the image to check the change of pixel value from 1 to 0 horizontally and vertically. Then the coordinates of the changed pixels were saved. In the last step; the white pixels were changed to black color. The results of applied method to a normal, malignant and benign type images are shown in Figure 4.12., Figure 4.13. and Figure 4.14. respectively.

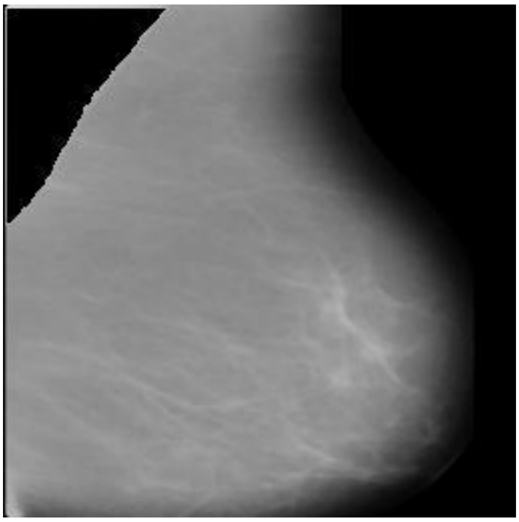


Figure 4.12. Normal image case muscle removed

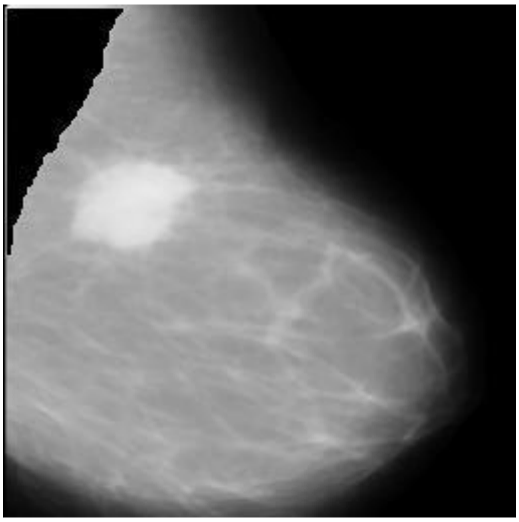


Figure 4.13. Malignant image case muscle removed

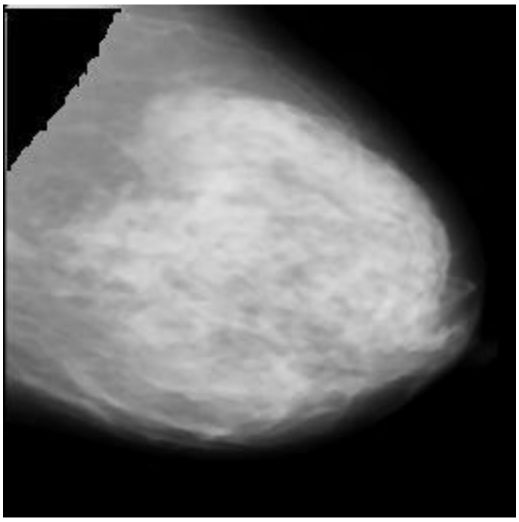


Figure 4.14. Benign image case muscle removed

4.2.4. Enhancement of mammogram images

Image enhancement desires to improve the perception of the available information in the images. Enhancement tools prepare the images for segmentation technique as clustering. Methods as low pass filter, histogram modification techniques, and contrast enhancement were used. Gaussian filter, unsharp filter and then linear histogram modification were applied on the images.

Gaussian Filter: It is a low pass filter which used to smooth digital images and enhance them. For removing noises and details and blur the image; this filter uses a special smoothing operator [40]. Gaussian filter has impulse response of a gaussian function. It removes high frequency components from the image without affecting the important data [41]. The main formula of gaussian filter is given by

$$G(x,y) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x^2+y^2)}{2\sigma^2}} \quad [40] \tag{4.2}$$

Where σ is the standard deviation of the distribution, x and y are pixel coordinate values and G is the output image. The values of the of gaussian filter's template elements depend on their distance from the template center.

For enhancing high frequency components and edges; unsharp filter is used. It achieves this by subtracting smooth copy of an image from the real one. This filter is known as a simple and effective in using to improve the quality of mammography digital images; it is widely used for conservative high frequency components and enhances the edges. This filter highlights the fine and blurred details. The main formula of unsharp filter is given by

$$g(x,y) = f(x,y) - fsmooth(x,y), \text{ [42]} \qquad (4.3)$$

Where f is the original image, x and y are the pixels coordinates values and g is the output image.

Histogram Modification: The pixel intensity values in digital image can be displayed using histogram representation of the image. The pixels intensity of mammogram images may be modified using some tools. It is a graph which display the number of pixels that have the same value of intensity. The mammogram digital images used in this research are 256*256 grayscales, the histogram of these images shows their pixels intensity distribution. Figure 4.15. display histogram representation of a mammogram image.

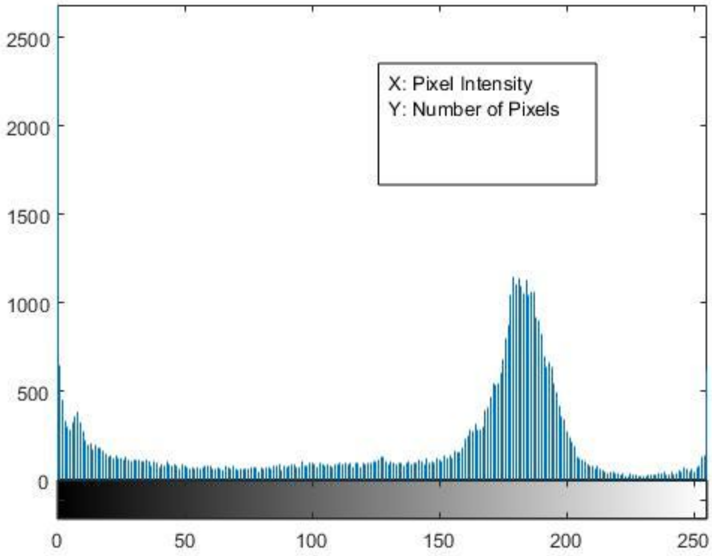


Figure 4.15. Histogram representation of a mammogram image

To modify the histogram of an image, a linear modification may be applied. It covers all the ranges from 0 to 255 of the gray images. The equation of linear modification is given by

$$O_{j,i} = \frac{I_{j,i}}{I_{\max - \min}} * (O_{\max} - O_{\min}) + O_{\min} \text{ , [43]} \tag{4.4}$$

Where I is the input image, I maximum and I minimum are the highest and lowest intensities values of input image respectively. O is the output image; O maximum and O minimum are the highest and lowest intensities of the output image respectively. j and i are the pixel coordinates value in the input image.

4.3. Result of Overall Pre-Processing of Mammogram Image

The preprocessing result of mammogram images are shown below for three type of mammogram images normal, malignant and benign in Figure 4.16., Figure 4.17. and Figure 4.18. respectively. This stage affects positively on the images; the memory spaces were reduced, and the images were enhanced. After preprocessing step, the images were given as input to the clustering algorithms.

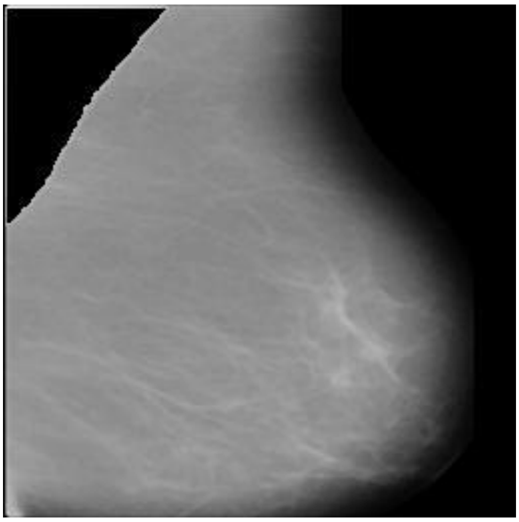


Figure 4.16. Normal image case

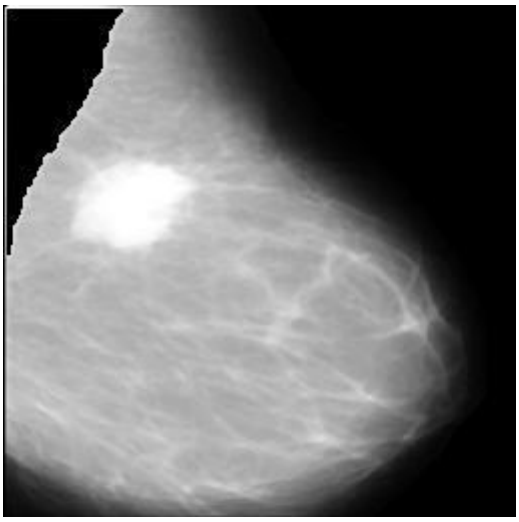


Figure 4.17. Malignant image case

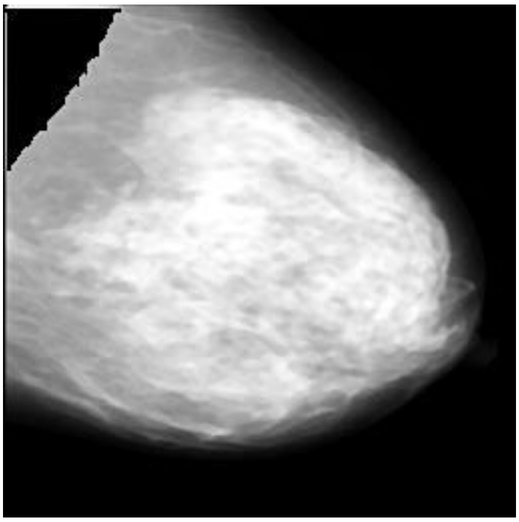


Figure 4.18. Benign image case

In this chapter; three types of mammogram images were preprocessed using digital images processing tools. Techniques as median filter, removing of lines and labels, removing of muscle and enchantments tools were used. These techniques improved the images and made the region of interests appear very clearly. The quality of images after all these phases improved which it was clear from the region of interests. The preprocessing of images in DM is a main contribute as it is preparing the data for coming phase. The output of these preprocessing methods is given as input to clustering algorithms.

CHAPTER 5.
CLUSTERING AND CLASSIFICATION TECHNIQUES

In this chapter; DM techniques were used for analyzing mammogram images. Two DM techniques; clustering and classification were used. Clustering algorithms were applied to segment the mammogram digital images. Then classification of BC data was achieved. The first part of the chapter discussed the clustering techniques and the second part discussed the classification techniques.

5.1. Clustering Technique

5.1.1. Introduction

This part of the chapter discussed the clustering algorithms. Two types of clustering techniques algorithms; KM and FCM were applied on the preprocessed images. The mammogram images were filtered and enhanced to be ready for clustering techniques. Segmentation is the other name for clustering in some different application [44,45]. Segmentation is used mainly for dividing objects into dissimilar groups. This technique means to put the image's pixels into different groups. For dividing of images based on their properties; color, pixel intensity, contrast and brightness; segmentation by means of clustering technique is preferred. Clustering play a very important major in the grouping the competent of images into different groups. In clustering process; distance function is used. In medical images; segmentation is used for locating different abnormalities which may include location of tumor and lesions [46]. The result of clustering techniques are different clusters or groups of data. The component of each cluster are objects similar between them. In the following parts; the used techniques of clustering were examined then the result of each used algorithm was shown for the three type of input images. Both clustering techniques show their proficiency; mainly two functions of clustering techniques were focused; run time and space memory which an algorithm needs as result storage.

5.1.2. KM clustering

For dividing a set of objects intro distinct group, a clustering algorithm is used. The clustering algorithm is processed using distance functions for measuring the distance across the objects. Clustering techniques were applied in many different applications; image processing, information analysis and in medical application for clustering medical images. Clustering is considered as special topic in classification technique. KM clustering is an unsupervised learning algorithm [47]. This technique uses a simple classification way by demanding on a pre-given clustering number. This technique is processed by minimizing this function,

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \| x_i^{(j)} - c_j \|^2 \tag{5.1}$$

Where $\| x_i^{(j)} - c_j \|^2$ is the distance function and it is used for finding the distance between the data $x_i^{(j)}$ and each cluster point c_j. This point is used as indicator of the distance of the cluster centers to the n data points [48]. According to [48]; the main outline/steps of KM are given as,

1. Number of clusters must be defined.
2. The K group centers are randomly chosen.
3. The centers of clusters are calculated.
4. The distances between each cluster and the distinct pixels are calculated.

5. Check the calculated distance if they are close to the center. If this is the case; moving to that cluster should be done. Apart from that, moving to the following group should be achieved.
6. Re-evaluate the cluster midpoint.
7. Until the midpoint doesn't move, repeat the process.

5.1.3. KM clustering results

Segmentation of mammogram images were carried out in this research for finding the abnormal areas in the breast image using KM clustering technique. The segmentation by means of clustering process was applied to find the cancer affected region and to extract it from the other part of the image. The segmentation process is applied to mammogram in order to extract the cancer region based on the intensity of the pixels. The KM clustering is always used in the intensity-based segmentation. KM clustering was applied to normal mammogram images, malignant images and benign images. The k value which is the cluster number value was given to the implemented algorithm, for KM algorithm the number of clusters were five. This means 5 clusters were clustered from each mammogram image. The cancer affected region is mostly the fifth cluster.

The result gathered by KM clustering algorithm for all type of mammogram images were discussed in detail below with the help of cluster images generated during the clustering process and the experimental values are also tabulated as well. The following figure shows the result of KM clustering applied to a normal mammogram image. Five clusters were segmented from the input mammogram digital image. It appears that the last cluster is empty which show that the clustered normal image is different from malignant and benign clustered image.

Figure 5.1. shows the result of a normal image clustered by using of KM clustering technique, the original image which was used as input followed by the result of the KM algorithm. The five clusters of the normal image were segmented using KM and are shown respectively in cluster No: 1, No: 2, No: 3, No: 4 and No: 5. It is clearly shown that there is no abnormal area in the image and the fifth cluster is empty for normal image input.

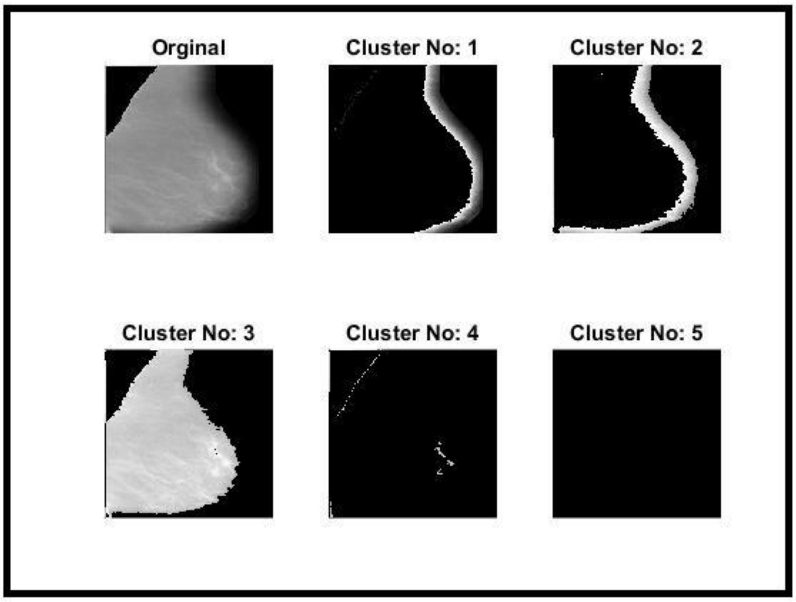


Figure 5.1. Normal mammogram image clustered by KM

Figure 5.2. reflects the result of KM clustering applied to a malignant mammogram image, the original image which was used as input followed by the result of the KM algorithm. It is clear that the cancer affected area was in the last cluster. The five clusters of the malignant image were segmented using KM and are shown respectively in cluster No: 1, No: 2, No: 3, No: 4 and No: 5. The intensity of the abnormal part of mammogram image is clearly different from other part of the image. The cluster No:5 have the abnormal part of mammogram image.

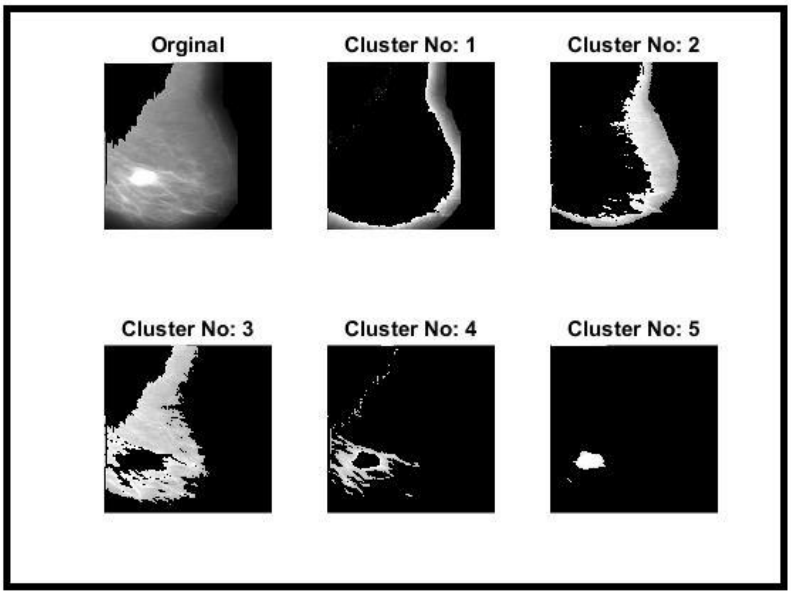


Figure 5.2. Malignant mammogram image clustered by KM

Figure 5.3. reflects the result of KM clustering applied to a benign mammogram image, the original image which was used as input followed by the result of the KM algorithm. It is clear that the cancer affected area was the last cluster. The five clusters of the malignant image were segmented using KM and are shown respectively in cluster No: 1, No: 2, No: 3, No: 4 and No: 5. The intensity of the abnormal part of mammogram image is clearly different from other part of the image. The cluster No:5 has the abnormal part of mammogram image.

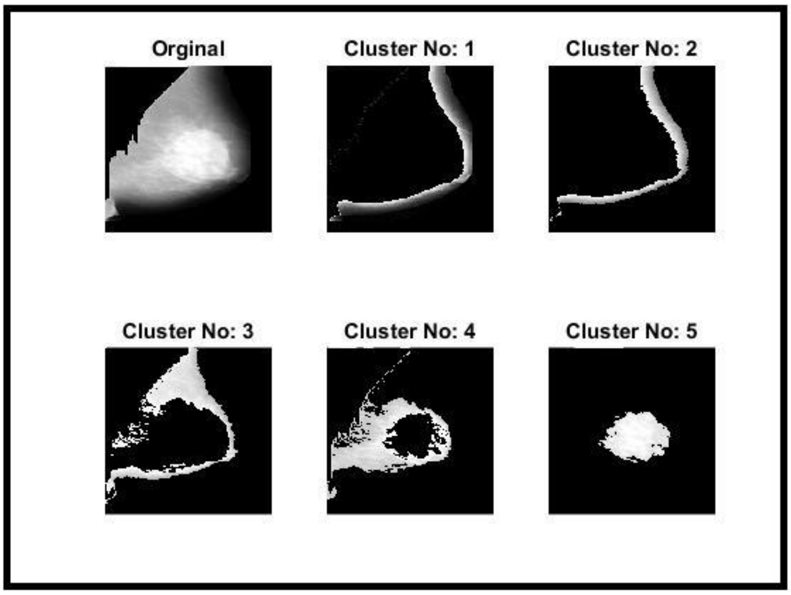


Figure 5.3. Benign mammogram image clustered by KM

For more understanding the results of KM clustering applied to the three types mammogram images; few tables were prepared to summarize all the result gathered by KM. The run time and memory space for each type of image were mentioned as well.

Table 5.1. shows the result of KM applied to normal mammogram images. Tables 5.2., 5.3. and 5.4. shows the result of KM applied to malignant mammogram images. Tables 5.5., 5.6. and 5.7. shows the results of KM applied to benign mammogram images. For these tables; the used terms are C which mean cluster, W is the number of white pixels and B is the number of black pixels.

Table 5.1. The result of KM applied to normal mammogram images

Image No	**C1**		**C2**		**C3**		**C4**		**C5**		**Run Time**	**Memory Space**
	W	**B**	**W**	**B**	**W**	**B**	**W**	**B**	**W**	**B**	**Millisecond**	**KB**
1	8426	57110	7469	58067	35538	29998	546	64990	65	6547	1270	22
2	8846	56690	24841	40695	8382	57154	130	65406	64	65472	1429	27
3	9478	56058	9168	56368	21681	43855	450	65086	71	65465	1401	24
4	9185	56351	7483	58053	28618	36918	89	65447	21	65515	1624	25
5	9636	55900	5557	59979	27808	37728	8419	57117	75	65461	1357	35
6	12471	53065	21731	43805	21513	44023	85	65451	29	65507	1353	37
7	53362	12174	8109	57427	31470	34066	336	65200	36	65500	1423	23
8	10515	55021	7293	58243	26404	39132	9595	55941	23	65513	1432	33
9	14492	51044	8755	56781	26945	38591	141	65395	19	65517	1464	22
10	13970	51566	7586	57950	35616	29920	149	65387	31	65505	1338	23

Table 5.2. The result of KM applied to malignant mammogram images

Image No	C1		C2		C3		C4		C5		Run Time	Memory Space
	W	B	W	B	W	B	W	B	W	B	Millisecond	KB
1	9130	56406	11822	53714	18519	47017	3625	61884	1100	64436	1405	32
2	6645	58891	3058	62478	3696	61840	19487	46049	6712	58824	1530	33
3	8404	57132	4949	60587	6483	59053	10508	55028	3430	62106	1352	30
4	9851	55685	4258	61278	6987	58549	11605	53931	6975	58561	1289	37
5	11548	53988	4987	60549	7366	58170	13728	51808	7147	58389	1386	35
6	7206	58330	4973	60563	5719	59817	6870	58666	4171	61365	1341	29
7	8180	57356	9183	56353	5878	59658	10898	54638	4600	60936	1398	32
8	11373	54163	5309	60227	5528	60008	11469	54067	4998	60538	1378	32
9	10835	54701	4737	60799	6658	58878	23530	42006	7960	57576	1403	45
10	11455	54081	5476	60060	6311	59225	25997	39539	8770	56766	1334	43
11	12424	53112	4276	61260	4570	60966	21623	43913	9608	55928	1337	37
12	8283	57253	4930	60606	10935	54601	8369	57167	5700	59836	1403	36
13	7600	57936	5699	59837	7014	58522	12160	53430	4318	61218	1313	35
14	9487	56049	5807	59729	5741	59795	11045	54491	7167	58369	1313	38

Table 5.3. The result of KM applied to malignant mammogram images

Image No	C1		C2		C3		C4		C5		Run Time	Memory Space
	W	B	W	B	W	B	W	B	W	B	Millisecond	KB
15	10807	54729	5150	60431	7715	57821	23864	41672	6840	58696	1320	40
16	13816	51720	6086	59468	9799	55737	14251	51285	3387	62149	1303	34
17	9670	55866	5331	60205	6542	58994	14844	50692	10005	55531	1265	35
18	18952	46584	10294	55242	8411	57125	10038	55498	2445	63091	1259	34
19	10892	54644	3638	61898	4800	60736	16491	49045	58012	7524	1337	35
20	8539	56943	4563	61000	10880	54656	9723	55813	1998	63538	1312	33
21	9152	56384	4582	60954	12306	53230	9679	55857	3896	61640	1354	39
22	11576	53960	5879	59657	12650	52886	8128	57408	12601	52935	1395	35
23	11040	54496	5817	59719	17055	48481	9980	55556	4192	61344	1512	42
24	10113	55423	4342	61194	8396	57140	13267	52269	3770	61766	1408	34
25	7266	58270	6260	59276	9510	56026	12522	53014	6934	58602	1304	39
26	8340	57196	5255	60281	5844	59692	7019	58517	11957	53561	1375	30
27	8139	57397	3313	62223	5426	60110	16352	49184	5448	60088	1309	38
28	10312	55224	6023	59513	14478	51058	4884	60652	3127	62409	1385	30

Table 5.4. The result of KM applied to malignant mammogram images

Image No	C1		C2		C3		C4		C5		Run Time	Memory Space
	W	B	W	B	W	B	W	B	W	B	Millisecond	KB
29	9577	55959	4725	60811	8267	57269	8120	57416	4146	61390	1392	30
30	8070	57466	5340	60196	6478	59028	9649	55857	5024	60512	1392	33
31	10072	55464	5155	60381	7252	58284	27626	37910	10777	54759	1308	45
32	10638	54898	5408	60128	5807	59729	22360	43176	6047	59489	1312	36
33	15830	49706	13364	52172	9802	55734	7513	58023	3726	61810	1235	38
34	10957	54579	5725	59811	9400	56136	13923	51613	3817	61719	1346	40
35	9977	55559	5715	59785	12054	53482	9677	55895	5393	60143	1290	32

Table 5.5. The result of KM applied to benign mammogram images

Image No	C1		C2		C3		C4		C5		Run Time	Memory Space
	W	B	W	B	W	B	W	B	W	B	Millisecond	KB
1	9865	55671	4050	61486	4789	60747	14100	51436	10563	54973	1310	39
2	12289	53238	4108	61428	5277	60259	13776	51760	16581	48955	1306	38
3	8730	56806	5661	59875	8396	57140	15763	49773	8761	56775	1325	43
4	9155	56381	6400	59136	10853	54683	9061	56475	4732	60804	1654	31
5	8227	57309	3944	61592	11988	53548	13871	51665	11635	53901	1327	45
6	6972	58564	3423	62113	4297	61239	18850	46686	17267	48269	1263	45
7	6620	58916	3362	62174	16639	48897	9310	56226	2358	63178	1331	32
8	4582	60954	1261	63375	2376	63160	36231	29305	18822	46714	1345	33
9	11249	54242	10834	54702	11758	53778	6505	59031	5015	60521	1353	36
10	14365	51171	4975	60561	20354	45182	12807	52729	3732	61804	1569	41
11	12761	52775	4057	61479	7458	58051	17733	47803	6667	58869	1415	39
12	10827	54709	5389	60147	9428	56108	10396	55140	5705	59831	1336	33
13	7194	58342	5756	59780	6835	58707	9056	56480	4840	60696	1299	32
14	13358	52178	4358	61178	5398	60138	14477	51059	3354	62182	1286	33

Table 5.6. The result of KM applied to benign mammogram images

Image No	C1		C2		C3		C4		C5		RT	Memory Space
	W	B	W	B	W	B	W	B	W	B	MS	KB
15	16918	48618	6367	59169	6629	58907	10279	55257	5977	59559	1342	36
16	8205	57331	7777	57759	6387	59149	9645	55891	5181	60355	1436	31
17	12201	53335	4256	61280	7037	58499	12597	52939	13012	52524	1365	37
18	9338	56198	5458	60078	8575	56961	12232	53304	9890	55646	1392	39
19	12181	53355	4293	61243	6867	58669	16097	49439	7970	57566	1584	39
20	11972	53564	4676	60860	6969	58567	16614	48922	8228	57308	1671	38
18	9338	56198	5458	60078	8575	56961	12232	53304	9890	55646	1392	39
21	10943	54593	4963	60573	5363	60173	7708	57828	5662	59874	1296	29
22	12616	52920	9234	56302	5503	60033	6860	58676	4718	60818	1502	32
23	8363	57173	3994	61542	4448	61088	17831	47705	7424	58112	1267	30
24	5686	59850	2785	62751	3478	62058	20147	45389	6299	59237	1311	37
25	8200	57336	5040	60496	8892	56707	2093	42443	7201	58335	1409	41
26	13493	52043	7473	58063	13946	511590	8673	56883	4412	61124	1257	32

Table 5.7. The result of KM applied to benign mammogram images

Image No	**C1**		**C2**		**C3**		**C4**		**C5**		**Run Time**	**Memory Space**
	W	**B**	**W**	**B**	**W**	**B**	**W**	**B**	**W**	**B**	**Millisecond**	**KB**
27	9096	56440	6731	58805	8204	57332	12222	53314	5048	60488	1278	32
28	11331	54205	7768	57768	22465	43071	6115	59421	3112	62424	1331	33
29	9378	56158	4804	60732	10024	55512	15224	50312	7880	57656	1318	42
30	10727	54809	7081	58455	8996	56540	11970	53566	9783	55753	1343	41
31	7484	58052	7555	57981	11014	54522	9565	55971	4526	60974	1391	34
32	8733	56803	4654	60882	7093	58443	9222	56314	4941	60595	1394	33
33	8455	57081	11712	53824	15305	50231	10693	54843	3541	61995	1314	35
34	8038	57453	4203	61333	5897	59639	13354	52182	6884	58692	1284	38
35	13783	51753	5141	60395	13331	52205	13479	52057	5528	60008	1360	40

The result of KM applied to all type of mammogram images where summarized in the tables (5.1. - 5.7.). Table 5.8. reveal the average results of KM clustering applied to the three types of images.

Table 5.8. Average result of KM clustering

Image Type	**C1**		**C2**		**C3**		**C4**		**C5**		**Average Run Time Millisecond**	**Average Memory KB**
	W	**B**	**W**	**B**	**W**	**B**	**W**	**B**	**W**	**B**		
Normal	15038	50498	10799	54737	26398	39139	1994	65342	43	65493	1410	27
Malignant	10176	55359	5753	59782	8408	57128	13167	52370	7320	58216	1351	36
Benign	8632	56898	4883	60653	8486	37050	15247	50262	10637	54895	1370	38

5.1.4. FCM clustering

FCM algorithm is a non-supervised technique which has been very capacious in different ML applications. This technique allows the data to be in the contents of more than one cluster and it uses a function of memberships for this aim. The degree of belonging for different clusters is described by the membership values. It was derived from Ruspini fuzzy clustering theory proposed in 1990 [49]. The function of FCM is given by

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^{m} \| x_i - c_j \| \ , \quad 1 \leq m < \infty \tag{5.2}$$

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\| x_i - c_j \|}{\| x_i - c_k \|} \right)^{\frac{2}{m-1}}} \tag{5.3}$$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} . x_i}{\sum_{i=1}^{N} u_{ij}^{m}} \tag{5.4}$$

The m is a real number and the degree of membership of x_i in the cluster j is given by u_{ij}. The d-dimensional measured data; x_i is the ith and the d dimension center of the cluster is given by c_j. The similarity between any measured data and the center is found by ||*|| which is a norm expressing. The FCM function is given by the above equations. This membership function u_{ij} is to be updated with cluster center c_j.

The $\backepsilon$ is a termination criterion and it has the value between 0 and 1. The iteration stops when $\max_{ij}\left\{|u_{ij}^{(k+1)} - u_{ij}^{(k)}|\right\} < \backepsilon$. K is the iteration step. The saddle point J_m is used for converges procedures. According to [50] the steps of the algorithm are given by,

1. Initialize the U= [u_{ij}] matrix, U $^{(0)}$.
2. Calculate the centers vectors C$^{(k)}$= [c_j] with U$^{(k)}$ at k-step:

   i. $$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} . x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$
3. Update the U$^{(k)}$, U$^{(k+1)}$ is preformed

   i. $$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\| x_i - c_j \|}{\| x_i - c_k \|} \right)^{\frac{2}{m-1}}}$$
4. Check if || U$^{(k+1)}$ - U$^{(k)}$||< $\backepsilon$, then stop. A part of that return to step 2.

5.1.5. FCM clustering results

This part discussed the result of segmentation the mammogram digital images using FCM. The technique is known as soft segmentation method. The three type of mammogram digital images; normal, malignant and benign were analyzed in this part. The algorithm was implemented in MATLAB for clustering the mammogram images and finding the abnormal part in the image alone. The following figures show the result of FCM applied to mammogram images. The whole results for all images were analyzed and summarized in tables as well.

The result of FCM applied to a normal image is shown in Figure 5.4.; normally 4 numbers of clusters were segmented from the original image. There was no different region with different intensity, so the image does not have abnormal area or affected area of cancer. The number of white pixels is decreasing after the second cluster. Four number of clusters were segmented from the normal images are shown in cluster No: 1, No: 2, No: 3 and No: 4.

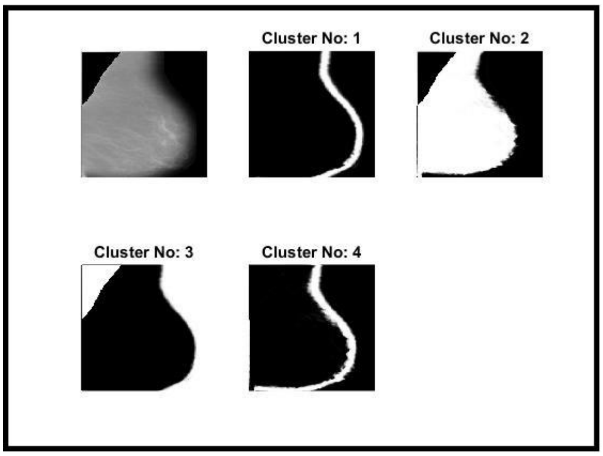


Figure 5.4. Normal mammogram image clustered by FCM

The result of FCM applied to a malignant image is shown in Figure 5.5., again 4 numbers of clusters were segmented from this image. The FCM segmented the image to just four clusters as it shown that the abnormal region of the image appears in the first cluster. The segmented clusters are shown respectively in cluster No: 1, No: 2, No: 3 and No: 4. The cluster No: 1 has the abnormal region which it appears clearly with different intensity.

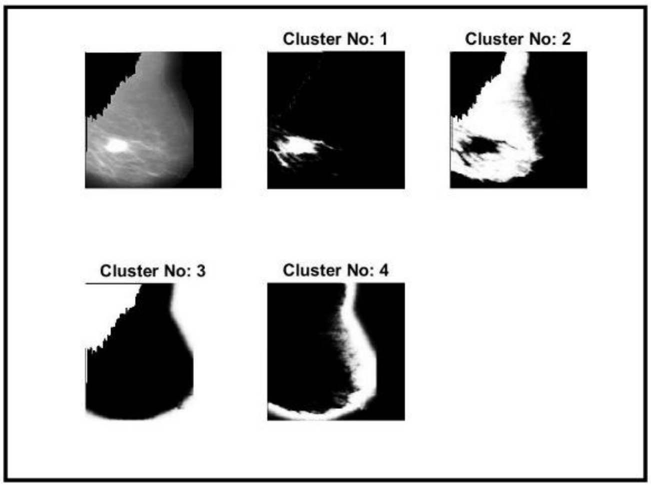


Figure 5.5. Malignant mammogram image clustered by FCM

The result of FCM applied to a benign image is shown in Figure 5.6.; as malignant type 4 numbers of clusters were segmented from this image. The affected part in the image appeared in cluster No: 3. Again FCM segmented the image with four number of clusters are shown in cluster No: 1, No: 2, No: 3 and No: 4 respectively. The affected part appears clearly with different intensity.

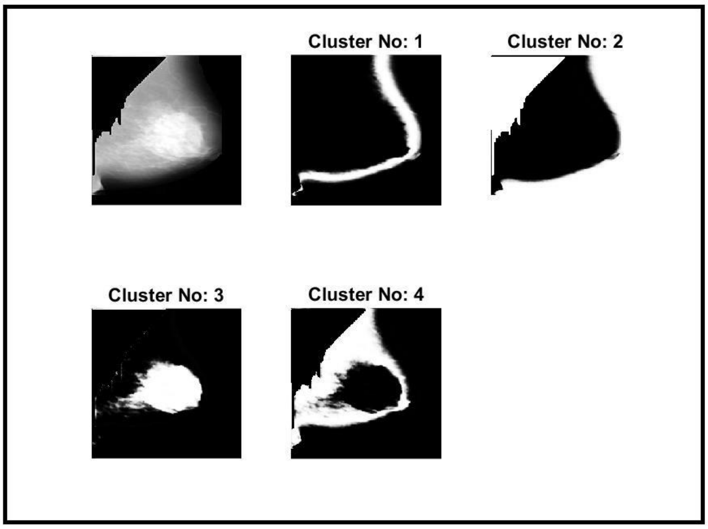


Figure 5.6. Benign mammogram image clustered by FCM

The experimental result of FCM applied to normal, malignant and benign mammogram type images were summarized in the following tables. Table 5.9. shows the result of FCM applied to normal mammogram images. Tables 5.10., 5.11. and 5.12. shows the result of FCM applied to malignant mammogram images. Tables 5.13., 5.14. and 5.15. shows the results of FCM applied to benign mammogram images. In these tables; the used terms are; C which means cluster, W is the number of white pixels and B is the number of black pixels.

Table 5.9. The result of FCM applied to normal mammogram images

Image No	C1		C2		C3		C4		Run Time Millisecond	Memory KB
	W	B	W	B	W	B	W	B		
1	4508	61028	6210	59326	35739	29797	19037	46499	7199	18
2	16235	9301	13145	52391	4173	61363	31890	33646	5951	23
3	4539	60997	6512	59024	18844	46692	35600	29936	3727	15
4	4091	61445	4210	61326	26730	38806	30480	35056	5480	16
5	4658	60878	25790	39746	5263	60273	29794	35742	4174	17
6	32860	32676	6883	58653	6518	59018	19219	46317	5288	17
7	21207	44329	5738	59798	6513	59023	5755	33505	5204	17
8	18238	47298	17178	48358	24274	41262	5755	59481	2750	21
9	27895	37641	25117	40419	6377	59159	605	59431	5429	16
10	5654	59882	31437	34099	20147	45389	8235	57301	5594	20

Table 5.10. The result of FCM applied to malignant mammogram images

Image No	C1		C2		C3		C4		Run Time	Memory Space
	W	B	W	B	W	B	W	B	Millisecond	KB
1	11273	54263	23714	41822	28085	37451	2394	63142	5148	19
2	33724	31812	8918	56618	4495	61041	18322	47214	3285	21
3	6835	58701	7086	58450	12148	5338	39395	26141	3839	18
4	5592	59944	35719	29817	10629	54907	13550	51986	4882	20
5	6589	58947	11608	53928	32617	32919	14660	50876	4967	18
6	42582	22954	6982	58554	6786	58750	91333	56403	3070	17
7	33958	31578	13157	52379	7819	57645	10444	55092	4266	19
8	37884	27652	7598	57938	12578	52958	7412	58124	2550	19
9	6534	59002	15555	49981	24856	40680	18571	47019	5122	28
10	14457	51079	21292	44244	7434	58102	22257	43279	5106	23
11	26842	38694	5967	59569	23590	41964	9073	56463	5200	19
12	10450	55086	14023	51513	5895	59641	35106	30430	4360	19
13	8270	57266	7907	57629	36158	29378	13104	52432	5339	21
14	12476	53060	8892	56644	36698	28838	7425	58111	4369	20

Table 5.11. The result of FCM applied to malignant mammogram images

Image No	C1		C2		C3		C4		Run Time	Memory Space
	W	B	W	B	W	B	W	B	Millisecond	KB
15	7135	58401	19537	45999	14991	50545	23745	41791	3588	25
16	12271	53265	15245	50291	31085	34451	6862	58674	5073	21
17	7796	57740	29524	36012	16868	48668	11272	54264	4817	18
18	11042	54494	13531	52005	29410	36126	11467	54069	4395	22
19	6853	58683	34434	31102	19717	45819	4481	61055	5219	20
20	5718	59755	38237	2799	4923	60613	16532	60613	5958	19
21	5705	59831	8069	47467	35266	30270	6444	59092	2840	19
22	24356	41180	16679	48857	16846	48690	7590	57646	2603	19
23	6909	58627	21881	43655	9342	56194	27356	38180	3578	23
24	5443	60093	37027	28509	9229	56307	13781	51755	3894	19
25	7984	57552	13975	51561	31842	33694	11659	53877	3079	21
26	35453	30083	6152	59384	8103	57433	15783	49753	3003	18
27	4567	60969	10572	54964	36522	29014	13800	51736	3581	23
28	17536	48000	7594	57942	34921	30615	5448	60088	2785	17

Table 5.12. The result of FCM applied to malignant mammogram images

Image No	C1		C2		C3		C4		Run Time	Memory Space
	W	B	W	B	W	B	W	B	Millisecond	KB
29	12114	53422	39333	26203	6154	59382	7879	57657	3518	21
30	6963	58573	38984	26552	7740	57796	11792	53744	4732	20
31	17628	47908	21491	44045	7245	58291	19073	46463	3922	26
32	25497	40039	6512	59024	7411	58125	26060	39476	4686	17
33	10833	54703	31648	33888	8188	57348	1469	50767	3903	20
34	10018	55518	14791	50745	32984	32552	7631	57905	3999	21
35	6994	58542	31344	34192	14767	50769	12370	53166	3013	17

Table 5.13. The result of FCM applied to benign mammogram images

Image No	C1		C2		C3		C4		Run Time	Memory Space
	W	B	W	B	W	B	W	B	Millisecond	KB
1	5615	59921	8173	57363	32983	32598	18747	46789	5071	18
2	8359	57177	5281	60255	27284	38252	24563	40973	4021	19
3	11374	54162	26361	39175	7504	58032	20220	45316	3942	21
4	13839	51697	11024	54512	8284	57252	32333	33203	3085	16
5	14988	50548	19247	46289	4769	60676	26464	39072	3169	23
6	10468	55068	25596	39940	5271	60265	24126	41410	2383	20
7	18263	47273	32622	32914	4435	61101	10178	55358	1947	19
8	25587	39949	25692	39844	11047	54489	3141	62395	2122	22
9	10798	54738	9995	55541	28392	37144	16271	49265	4154	19
10	24320	41216	4344	61192	12224	53312	26416	40922	3337	22
11	18266	47270	31384	34152	11150	54386	4674	60862	5119	23
12	34397	31139	8020	57516	7017	58519	16036	49500	3838	18
13	11584	53952	7899	57637	38970	26566	6992	58544	5327	19
14	5996	59540	37854	27682	5054	60482	16576	48960	3335	19

Table 5.14. The result of FCM applied to benign mammogram images

Image No	C1		C2		C3		C4		Run Time	Memory Space
	W	B	W	B	W	B	W	B	Millisecond	KB
15	8606	56930	12354	53178	34510	31026	9996	55540	4223	21
16	36310	29226	10292	55244	9422	56114	9450	56086	5043	18
17	27817	36819	5819	59717	10318	55218	20621	44915	3535	17
18	14185	51351	14225	51311	29524	36012	7522	58014	2966	21
19	32028	33508	8702	56834	5360	60176	19402	46134	3124	18
20	9356	56180	19959	45577	30814	34722	5351	60185	5207	20
21	41087	54449	6593	58943	8001	57535	9796	55740	3163	17
22	11088	54448	8339	57197	38184	27352	7875	57661	3224	19
23	32603	32933	12454	53082	6034	29502	1435	51184	5292	18
24	4072	61464	34282	31254	16973	48563	10157	55379	3862	23
25	10471	55065	22088	43448	27188	38348	5738	59798	3973	21
26	10265	55271	7214	58322	30149	35387	17833	47703	4670	17
27	8827	56709	11563	53973	32806	32730	12264	53272	4288	19
28	7593	57943	6123	59413	25749	39787	26010	39526	4023	19

Table 5.15. The result of FCM applied to benign mammogram images

Image No	C1		C2		C3		C4		Run Time	Memory Space
	W	B	W	B	W	B	W	B	Millisecond	KB
29	14592	50944	14402	51134	30082	35454	6373	59163	3981	21
30	12414	53122	9115	56421	27460	38076	16488	49048	3451	21
31	10870	54666	8600	56936	29852	35684	16127	49404	6342	17
32	39244	26292	11047	54489	5886	59650	9310	56226	3964	18
33	13763	51800	20762	44810	5395	60141	25991	39945	3093	21
34	15111	50425	5631	59905	35940	29596	8805	56731	4563	19
35	17540	47996	12296	53240	6007	59529	29657	35849	2842	20

The result of FCM applied to all type of mammogram images where summarized in the tables (5.10. - 5.15.). Table 5.16. presents the average results of FCM clustering applied to the three types of images.

Table 5.16. Average results of FCM clustering

Image Type	**C1**		**C2**		**C3**		**C4**		**Average Run Time Millisecond**	**Average Memory KB**
	W	**B**	**W**	**B**	**W**	**B**	**W**	**B**		
Normal	13989	51584	14222	51314	15458	50078	21815	43727	5079	18
Malignant	14465	51069	18714	46822	18095	47438	14189	51680	4105	20
Benign	16931	48605	14724	50813	18573	46965	15259	50289	3770	20

5.2. Classification Technique

5.2.1. Introduction

Classification technique can be used to predict the performance of clustering techniques with the help of number of pixels in the produced results by clustering algorithms [51]. This part discussed classification algorithms applied to cancer data. For evaluating the performance clustering techniques; classification algorithms were used. The mammogram images were clustered to more than three images and the most important one is the cluster of abnormal regions. The cancer data here are mainly five attributes. The number of white pixels of abnormal part in the image is considered as the most important attribute. These attributes are; abnormality class, position coordinates and the character of the background tissue. Three classification algorithms were used in this research; ANN, SVM and KNN. The cancer data of three type of mammogram images; normal, malignant and benign were saved in a one Excel sheet and analyzed in the same time. The next part of this section summarizes the used classification algorithms and the last part describes the study that evaluates the results of these classification algorithms.

5.2.2. Classification algorithms

ANN; or artificial neural network is considered as the most popular approach to machine learning. ANN is a model of reasoning which was simulated and designed first based on the human brain. It is the most commonly used techniques due to their nonlinear nature. ANN is a modeling technique that inspired from the human brains structure. It has some layers where the data are processed through these layers. The layers have connections between them, and these connections are managed by weights. The ANN learn and process results by changing the given values to the weights. The network generates better and better predictions through each iteration. However, as the number of iterations reaches a certain level, the network begins to memorize the input data and over fitting occurs. So, the number of iterations must be neither too small so that the network can learn the patterns, nor too large so that over fitting does not occur.

SVM; Support Vector Machine algorithms was introduced in 1963 by Vapnik in [52]. The technique achieves its mission by using of a structural risk minimization principle. By this principle, minimizing of upper bound generalization error. SVM aim to find a hyperplane, or a decision surface, that divides a data set into two classes [53]. Data points that determine the hyperplane are called "support vectors" [54, 55]. In SVM, a margin is the distance between two classes. The classes depend on the dimension of the hyperplane which is related to the number of input features. SVM method finds a hyperplane in a n dimensional space that has the maximum margin.

KNN; K Nearest Neighbor is a very popular algorithm from classification techniques algorithms. It was first discovered in 1950s and became popular in 1960s [56]. This algorithm is generally used in the application like pattern and facial recognition [57]. KNN work to assigns label to data point from its neighbors considering the majority of closest neighbor points. KNN algorithm needs to see the test tuple and then perform generalization to classify the data by finding the closest the stored training tuples. The two important points in KKN are: choosing the right k; how many neighbors will be chosen for the tuples and calculating the distance between test instances and its neighbors [58].

5.2.3. Evaluation of classification algorithms

Classification technique was used to asses breast cancer data and predict class label. For this mission classification algorithms were used; ANN, SVM and KNN. Evaluation of any ML algorithm is an essential part of the research. For the evaluation of classification algorithms applied in this research; different measures were used. These measures are accuracy, precision, recall and F measure [59, 60]. A comparative study for evaluation these techniques algorithms was discussed here. The used formulas were explained and the result for all three classification then were given.

There are four important terms used for calculating these formulas: A TP or true positive, is a correctly predicted positive class [59]. Similarly; a TN or true negative, is a negative class which predicted positively. A FP or false positive, is a positive class which predicted incorrectly [59]. Similarly; a FN or false negative is a negative class which predicted incorrectly. Other important terms to consider are the TPR and FPR. These two important terms are used for the proportion of the negative and positive results with respect to the all gathered results. They are given by

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{5.9}$$

and by

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP+TN}} \tag{5.10}$$

Accuracy is the segment of prediction where the model predicated right; its formula is given as

$$\mathrm{Accuracy} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \tag{5.11}$$

Pression is the proportion of positive identifications which were actually correct. It is used to calculate the proportion of the predicated positive cases. It is given as

$$\text{Pression} = \frac{\text{TP}}{\text{TP+FP}} \tag{5.12}$$

Recall is the proportion of actual positives which the model identified correctly and its given by

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{5.13}$$

F1 measure gives the harmonic mean between recall and precision. It aims to find the balance between these two measures. It should be a value between zero and one.

$$\text{F1} = 2 * \frac{1}{\frac{1}{\text{Pression}}+\frac{1}{\text{Recall}}} \tag{5.14}$$

The result of all these formulas applied to the results of three classification algorithms' result are summarized and given in this part. The result of clustering algorithm with other cancer data were given to classification algorithms and class labels were predicted. The following table reflects the various error measure for the three mammogram type images using ANN.

Table 5.17. Evaluation of ANN classification algorithm

Type of Image	TPR	FPR	Accuracy	Pression	Recall	F1 Measure
Normal	0.0	0.0	0.0	0.0	0.0	0.0
Malignant	0.94	0.0	0.97	1.0	0.94	0.95
Benign	1.0	0.05	0.97	0.94	1.0	0.96

The following table reflects the various error measure for the three mammogram type images using KNN.

Table 5.18. Evaluation of KNN classification algorithm

Type of Image	TPR	FPR	Accuracy	Pression	Recall	F1 Measure
Normal	0.0	0.0	0.0	0.0	0.0	0.0
Malignant	0.96	0.08	0.94	0.91	0.96	0.93
Benign	0.91	0.03	0.94	0.97	0.91	0.94

The following table reflects the various error measure for the three mammogram type images using SVM.

Table 5.19. Evaluation of SVM classification algorithm

Type of Image	TPR	FPR	Accuracy	Pression	Recall	F1 Measure
Normal	0.0	0.0	0.0	0.0	0.0	0.0
Malignant	0.57	0.45	0.52	0.45	0.57	0.50
Benign	0.54	0.42	0.52	0.65	0.54	0.59

This chapter discussed both clustering and classification techniques applied to mammogram digital images and cancer data. KM and FCM were used for clustering mammogram images. The result of both clustered methods award significant results; run times and memory spaces were measured and tabulated. The implementation of KM algorithm cluster the images to five clusters and the implementation of FCM cluster the image to four clusters. For more checking the performance of these implemented algorithms; classification techniques were applied to the extracted data from abnormal cluster. It is represented by the number of pixels of the abnormal region. The classification techniques predicated the class label of disease using cancer data. The highest accuracy was found using ANN classification algorithm. The next chapter revise all the results detected with brief discussion.

CHAPTER 6. DISCUSSION OF RESULTS

6.1. Introduction

This chapter briefly discuss in detail the results of clustering algorithms KM and FCM with respect to their quality, run time and memory space. Also, the results of classification algorithms ANN, SVM and KNN in terms of their accuracy were analyzed and compared in detail. These classification algorithms were applied mainly for finding the efficiency of the clustering algorithms.

6.2. Results Discussion

Table 6.1. reflects the results of clustering algorithm with respect to clustering quality in which it has average number of white color pixels, Table 6.2. reflects the results of clustering algorithm with respect to time complexity and Table 6.3 reflects the average space results used to store the results produced by the clustering algorithms. The pictorial representation of Tables 6.1., 6.2. and 6.3. is shown in the Figures 6.1., 6.2. and in Figure 6.3.. For more checking of results, this work used classification techniques; ANN, SVM and KNN. These algorithms were applied for finding the efficiency of clustering algorithms in terms of accuracy.

Table 6.1. shows average number of pixels in the affected cluster of normal, benign, and malignant images generated by clustering algorithms. It is evident from Table 6.1. that KM algorithm has low average (quality) when compared with FCM. Also, the clustering algorithm FCM gets results in four cluster. But KM algorithm gets result in five clusters.

Table 6.1. Average number of white pixels

Image Type	Average Number of White Pixels in the Affected Cluster	
	KM	FCM
Normal	43	Non
Malignant	7320	12370
Benign	10637	16608

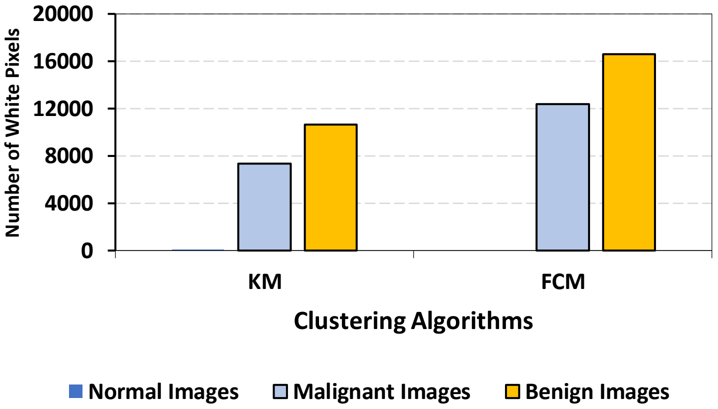


Figure 6.1. Quality based inference of Clustering Algorithms

Table 6.2. shows the processing time taken by the algorithms for the taken dataset. From Figure 6.2. it can be inferred that KM has less processing time when compared with FCM.

Table 6.2. Average run times of clustering algorithms

Image Type	Average Run Time in Millisecond	
	KM	FCM
Normal	1410	5079
Malignant	1351	4105
Benign	1370	3770

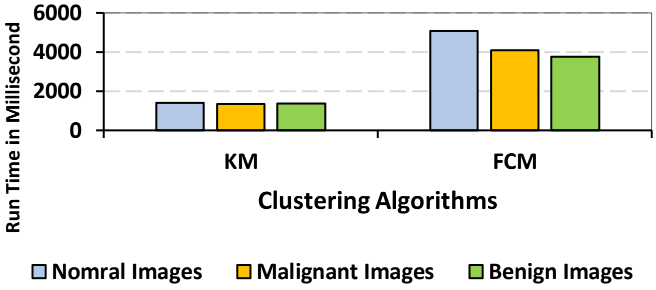


Figure 6.2. Time Complexity based inference of clustering algorithms

Table 6.3. contains the results of proposed algorithms based on the average memory used to store the clusters. From Figure 6.3., it is clear that FCM takes less memory when compared with KM.

Table 6.3. Average memory space of clustering algorithm

Image Type	Average Memory Space in Kilobyte	
	KM	**FCM**
Normal	27	18
Malignant	36	20
Benign	38	20

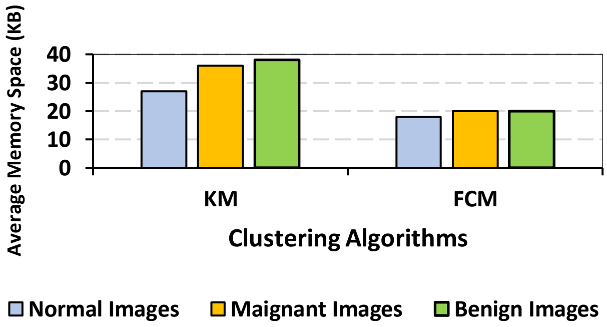


Figure 6.3. Space Complexity based inference of clustering algorithms

Table 6.4. tabulates the various parameters recorded for the tested classification algorithms. ANN classification technique performs well than others as the accuracy rate is high for all parameters while SVM classification technique performance is less compared with other techniques. Figure 6.4. is the pictorial representation of Table 6.4. From this figure, it is evident that the performance of ANN classification technique is better than the other algorithms in terms of its high-quality results.

Table 6.4. Classification algorithms results

Parameter	ANN	SVM	KNN
TPR	0.94	0.57	0.96
FPR	0	0.45	0.08
Precision	1	0.45	0.91
Recall	0.94	0.57	0.96
F1 Measure	0.95	0.50	0.93

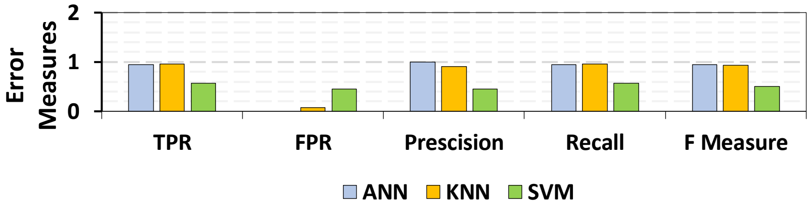


Figure 6.4. Error measures of classification algorithms

For evaluating of the accuracy of classification algorithms, it is necessary to carry out by calculating the three parameters accuracy, sensitivity and specificity. In continuation of this process, Table 6.5. shows the predicted values for breast cancer data using the three algorithms ANN, KNN and SVM algorithms. Here, ANN classification technique works fine with high accuracy yielded in less time. But, SVM algorithm has less accuracy rate.

Table 6.5. Performance comparison of classification algorithms

Parameter	Accuracy %	Sensitivity %	Specificity %
ANN	97	94	0
KNN	94	96	8
SVM	52	57	45

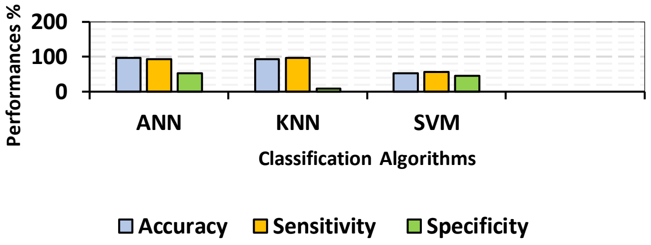


Figure 6.5. Performance comparison of classification algorithms

Figure 6.5. shows a chart illustrating the accuracy, sensitivity, specificity of produced results shown in Table 6.5. The results of the classification results are discussed in this section. The result of clustering algorithms with other data attribute were classified for label prediction of cancer probability. Three classification algorithms were used, the accuracy of ANN algorithm is 97%, SVM is 52% and KNN is 94%.

Table 6.6. shows the classification accuracy of ANN, SVM and KNN for normal, benign and malignant breast cancer dataset. The accuracy value for ANN is 97, for SVM is 52 and for KNN is 94. From table 6.6. it is obvious that ANN has the best accuracy when compared to SVM and KNN classifiers. Figure 6.6. shows a chart illustrating the average accuracy percentage for ANN, SVM and KNN with respect to the results of breast cancer dataset. It shows that the accuracy percentage is high for ANN as i.e. 97%. The accuracy of all of the classification algorithms were compared based on their accuracies. Hence, the performance of ANN is declared as the best among different classification algorithms for the given set of input images.

Table 6.6. Accuracy results of classification algorithms

Algorithm	Accuracy %
ANN	97
KNN	94
SVM	52

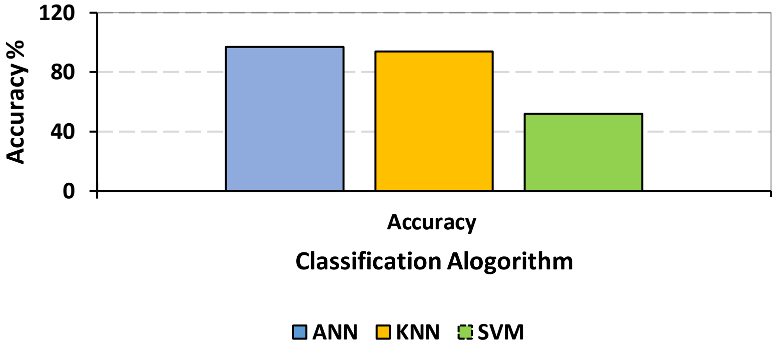


Figure 6.6. Performance results of classification algorithms

6.3. Summary

This research work evaluates the performance classification algorithms; ANN, SVM, and KNN algorithms in terms of accuracy with the using of various accuracy measures as TPR, FPR, Precision, Recall, and F-measure. In the implementation process, specific observed results were considered in the implementation of classification algorithms. The three classification algorithms were tested by calculating the error rate and accuracy. The experimental results show that ANN classifier has the highest accuracy of 97 % while SVM and KNN techniques have 94% and 52% as accuracy rates respectively. Based on gathered results, ANN was better than the other two methods techniques.

The efficiency of clustering algorithms for breast cancer images is proved by the computation of white color pixels in the resulted output of taken two algorithms. Also, run time and space used to store the results are taken for further analysis. The three parameters, namely, clustering quality, space and time complexity play an important role in predicting the performances of clustering algorithms KM and FCM. The quality of clustering algorithm constitutes it produce a clear result with less number of clusters. To prove this statement, FCM algorithm produce its clusters with four clusters and KM with five clusters.

Hence, this analysis proves that the performance of FCM was better than KM in terms of memory space. KM algorithm is better than FCM in term of time. The predicted results were analyzed, and the results showed both performance of KM and FCM. The same can be inferred from tables and graphs discussed in this chapter.

CHAPTER 7. CONCLUSION

DM is a very known term for KDD. Databases and professional systems have hidden information and with DM new knowledge can be discovered. DM involve various techniques and algorithms which can be used for analysis data. The stages of DM are exploration, pattern identification and deployment. The applications areas and fields of DM are wide and common. Applications and usage can be found in medical, space science and telecommunication.

This research analyzed mammogram images using DM techniques for BC prediction. Clustering technique' algorithms were used to segment the mammogram images and extract the affected abnormal areas. KM and FCM clustering algorithms where used to cluster the images. Both methods are commonly used in medical applications. Both of clustering algorithms gave significant results considering run time and space complexity. KM method cluster the mammogram images to five clusters whence FCM cluster the image to four clusters. The experimental result proved that the average rum time for KM clustering algorithm was lower than the average run time of FCM. The results proved also that the average memory space for KM algorithms is higher than the average memory of FCM algorithm.

For more checking the performance of clustering algorithms' results; classification technique was used. By the meaning of label prediction; different classification algorithms were used for this purpose. Classification algorithms; ANN, SVM and KNN were used to classify BC data and predict cancer probability. Classification technique can predict discrete class labels of inputs data models. The highest accuracy was found using ANN, then KNN and SVM in the last.

Finally, the future work of this research should be done with an original dataset taken from a real health center. The ability to predict cancer disease for specific region will be possible. This research can be used to implement and design of automatic diagnose systems and for predict different cancer disease; tumor, masses and micro calefaction.

REFERNCES

[1] https://www.zentut.com/data-mining/what-is-data-mining, Access Date: 25.11.2019.

[2] S. Rasheeduddin, “The Theoretical Framework of Data Mining & Its Techniques”, International Journal of Social Science & Interdisciplinary Research (IJSSIR), vol. 2, no. 1, pp. 81–85, 2013.

[3] G. PRATIYUSH and S. MANU, “Data Mining in Education: A Review on the Knowledge Discovery Perspective”, *International Journal of Data Mining & Knowledge Management Process*, vol. 4, no. 5, pp. 47-60, 2014.

[4] E. Udoka Felista and I. Chinelo Rose, “Industry Wide Applications of Data Mining”, *International Journal of Advanced Studies in Computer Science and Engineering*, vol. 2 no. 3, pp. 28-32, 2014.

[5] A. Naveen and T. Velmurugan, “A Survey on Medical Images Extraction using Parallel Algorithm in Data Mining”*, in the International Conference on Information, System and Convergence Applications,* 2015*,* pp. 86-91.

[6] http://peipa.essex.ac.uk/info/mias.html, Access Date: 25.09.2019.

[7] https://barnraisersllc.com/2018/10/01/data-mining-process-essential-steps/, Access Date: 01.01.2020.

[8] M. Fayyad, U. G. Piatetsky Shapiro and P. Smyth, “From Data Mining to Knowledge Discovery in Databases”, *AI Magazine*, vol. 17, pp. 37–54, 1996.

[9] H. Sahu, S. Shrma and S. Gondhalakar, “A Brief Overview on Data Mining Survey”, *International Journal of Computer Technology and Electronics Engineering (IJCTEE),* vol. 1, no. 3, pp. 114-121, 2011.

[10] https://nccn.com/type-of-cancer/breast-cancer/, Access Date: 10.10.2019.

[11] H. ELLIS, *Clinical Anatomy*, 11th ed. UK: Blackwell Publishing Ltd, 2006.

[12] B. T. ZIMMERMAN, *Understanding Breast Cancer Genetics*. USA: University Press of Mississippi, 2004.

[13] M. GUIDO, “Computer-Aided Detection and Classification of Masses in Digitized”, 2000.

[14] https://www.healthhype.com/characteristics-of-benign-and-malignant-tumors.html, Access Date 20.10.2019.

[15] G. Visalatchi, S. J. Gnanasoundhari, M. Balamurugan, “A Survey on Data Mining Methods and Techniques for Diabetes Mellitus”, *International Journal of Computer Science and Mobile Applications*, vol. 2, no. 2, pp. 100-105, 2014.

[16] A. Aqueel and A. Shaikh, “Data Mining Techniques to Find out Heart Diseases: An Overview”, *International Journal of Innovative Technology and Exploring Engineering*, vol. 1, no. 4, pp. 18-23, 2012.

[17] P. Venkatesan, N. R. Yamuna, “Treatment Response Classification in Randomized Clinical Trials through Decision Tree Approach”, *Indian Journal of Science and Technology*, vol. 6, no. 1, pp. 3912-3917, 2013.

[18] B. Kirthika, P. Malathi, C. L. Yashwanthi, and P. Sudharsan, “A Comparative Analysis of De-noising Techniques in Ultrasound B mode Images”, *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 3, no. 1, pp. 5136-5140, 2014.

[19] V. Anuja and R. Chitra, “Classification of Diabetes Disease Using Support Vector Machine”, *International Journal of Engineering Research and Applications*, vol. 3, no. 2, pp. 1797-1801, 2013.

[20] O. Anunciacao, C. Gomes, S. Vinga, J. Gaspar, L. Oliveira and J. Rueff , “A Data Mining Approach for Detection of High-Risk Breast Cancer Groups”, *Advances in Soft Computing*, vol.74, pp. 43-51, 2010.

[21] M. Medhat and W. Farouq, “Using Data Mining for Assessing Diagnosis of Breast Cancer”, *Proc. of International Multi conference on Computer Science*, vol. 5, pp. 11-17, 2010.

[22] K. Gandhi Rajiv, M. Karnan and S. Kannan, “Classification Rule Construction Using Particle Swarm Optimization Algorithm for Breast Cancer Datasets”, *IEEE International Conference on Signal Acquisition and Processing*, pp. 233-237, 2010.

[23] A. Soltani, A. Safavi, N. Parandeh, and M. Salehi, “Predicting Breast Cancer Survivability Using Data Mining Techniques”, 2nd IEEE Int. Conference on Software Technology and Engineering, vol. 2, pp. 222-227, 2010.

[24] M. Khan, J. Choi, H. Shin and M. Kim, “Predicting breast cancer survivability using fuzzy decision trees for personalized healthcare”, IEEE Int. Conf. on Engineering in Medicine and Biology Society, pp. 5148-5151, 2008.

[25] A. Ahirwar and R. Jadon, “Characterization of the Tumor Region Using SOM and Neuro Fuzzy techniques in Digital Mammography”, International Journal of Computer Science & Information Technology, vol. 3, no. 1, pp. 199-211, 2011.

[26] A. Shruthi, V. Vinod and A. Rampure, “Application of Fuzzy c-means and Neural Networks to Categorize Tumor Affected Breast MR Images”, International Journal of Applied Engineering Research, vol. 10, no. 64, pp. 275-281, 2015.

[27] J. Raikwal and K. Saxena, “Performance Evaluation of SVM and K-Nearest Neighbor Algorithm over Medical Data set”, International Journal of Computer Applications, vol. 50, pp. 35-39, 2012.

[28] N. Golestani, M. Etehad, “Level Set Method for Segmentation of Infrared Breast Thermo-grams”, EXCLI Journal, vol. 13, pp. 241-251, 2014.

[29] G. Sandhya, D. Vasumathi, and G. Raju, “Mammogram Image Segmentation Quality Enhancement Using Clustering Techniques”, American Journal of Engineering Research, vol. 4, no. 4, pp. 153-158, 2015.

[30] M. Antonie, R. Zaiane and A. Coman, “Application of Data Mining Techniques for Medical Image Classification”, Proceedings of the Second International Workshop on Multimedia Data Mining, pp. 94-101, 2001.

[31] M. Omran, A. Engelbrecht and A. Salman, “Particle Swarm Optimization Method for Image Clustering”, International Journal of Pattern Recognition and Artificial Intelligence, vol. 19, no. 3, pp. 297-321, 2005.

[32] K. Rajesh and S. Anand, “Analysis of SEER Dataset for Breast Cancer Diagnosis Using C4.5 Classification Algorithm”, International Journal of Advanced Research in Computer and Communication Engineering, vol. 1, no. 2, pp. 72-77, 2012.

[33] H. Yusuff, N. Mohamad, U. Ngah and A. Yahaya, “Breast Cancer Analysis Using Logistic Regression”, International Journal of Research and Reviews in Applied Sciences, vol. 10, no. 1, pp. 14-22, 2012.

[34] D. Lavanya and K. Usha, “Performance Evaluation of Decision Tree Classifiers on Medical Datasets”, International Journal of Computer Applications, vol. 26, no. 9, pp. 1-4, 2011.

[35] G. Sujatha and K. Usha, “A Survey on Effectiveness of Data Mining Techniques on Cancer Data Sets”, International Journal of Engineering Sciences Research, vol. 4, no. 01, pp. 1298-1304, 2013.

[36] S. Saheb and K. Satya, “Automatic Detection of Breast Cancer Mass in Mammograms Using Morphological Operators and Fuzzy c –Means Clustering”, Journal of Theoretical and Applied Information Technology, vol. 5, no. 6, pp. 704-709, 2009.

[37] B. Meral, “Breast Cancer Data Classification Using Svm, NB, and KNN algorithms” M.S. Thesis, Graduate School of Science Engineering and Technology, Istanbul Technical University, 2019.

[38] M. Rajalakshmi and P. Subashini, “Removal of Noise in the Chili Pepper Images Using Weighted 4-Monnected Median Filter”, International Journal of Computer Science Engineering and Information Technology Research, vol. 3, no. 3, pp. 141-150, 2013.

[39] M. P. Sukassini and T. Velmurugan, “Noise Removal Using Morphology and Median Filter Methods in Mammogram Images”, The 3rd International Conference on Small & Medium Business, Hochiminh, Vietnam, pp. 413-419, 2016.

[40] https://homepages.inf.ed.ac.uk/rbf/HIPR2/gsmooth.htm, Access Date: 25.11.2019.

[41] W. Passant, S. Amani and S. Amin, “Automated Breast Tumor Detection in Ultrasound Images Using Support Vector Machine and Ensemble Classification”, Journal of Biomedical Engineering and Biosciences, vol. 3, pp. 4-11, 2016.

[42] https://homepages.inf.ed.ac.uk/rbf/HIPR2/unsharp.htm, Access Date: 25.11.2019.

[43] A. Baba, Class Lecture, Topic: "Histogram Modification.”, Faculty of Engineering, University of Turkish Aerotactical Association, Ankara, Mar., 3, 2017.

[44] A. Lothe Savita and P. Deshmukh, “A Survey of Image Processing Techniques for Detection of Mass”, International Journal of Computer Science, vol. 2, no. 8, pp. 46-51, 2014.

[45] F. Yuhua, “Analysis on Algorithm and Application of Cluster in Data Mining”, Journal of Theoretical and Applied Information Technology, vol. 40, no. 1, pp. 416-419, 2015.

[46] D. Swagatam and K. Amit, “Automatic Image Pixel Clustering with an Improved Differential Evolution”, Applied Soft Computing, vol. 9, no. 1, pp. 226-236, 2009.

[47] T. Velmurugan and A. Dharmarajan, “Clustering Lung Cancer Data by k-Means and k-Medoids Algorithms”, International Conference on Information and Convergence Technology for Smart Society, pp. 21-24, 2015.

[48] https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html, Access Date: 25.11.2019.

[49] T. Velmurugan and T. Santhanam, “Implementation of Fuzzy C-Means Clustering Algorithm for Arbitrary Data Points”, International Conference on Systemic, Cybernetics and Informatics, pp. 68-71, 2011.

[50] https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/cmeans.html, Access Date: 25.11.2019.

[51] A. Aloraini, “Different Machine Learning Algorithms for Breast Cancer Diagnosis”, International Journal of Artificial Intelligence and Applications, vol. 3, no.6, pp. 21-30, 2012.

[52] http://web.cs.iastate.edu/~cs573x/vapnik-portraits1963.pdf, Access Date: 10.01.2020.

[53] J.E.T. Akinsola, “Supervised Machine Learning Algorithms: Classification and Comparison”, International Journal of Computer Trends and Technology (IJCTT), vol. 48, pp. 128 – 138, 2017.

[54] K. Srivastava, B. Lekha, “Data Classification using Support Vector Machine”, Journal of Theoretical and Applied Information Technology, 2010.

[55] S. Huang, N. Cai, P.P. Pacheco, S. Narrandes, Y. Wang and W. Xu, “Applications of Support Vector Machine (SVM) Learning in Cancer Genomics”, Cancer Genomics Proteomics, vol. 15, pp. 41 – 51, 2017.

[56] A. Sharma, and A. Suryawanshi, “A Novel Method for Detecting Spam Email using KNN Classification with Spearman Correlation as Distance Measure”, International Journal of Computer Applications, vol. 136, pp. 28–35, 2016.

[57] J. Han and M. Kamber, Data Mining: Concepts and Techniques, University of Illinois, San Francisco, 2006.

[58] Z. Zhang, “Introduction to machine learning: K-nearest neighbors”, Annals of Translational Medicine, vol. 4, pp. 218–218, 2016.

[59] https://developers.google.com/machine-learning/crash-course/classification/true-false-positive-negative, Access Date: 25.12.2019.

[60] https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234, Access Date: 12.12.2019.

RESUME

Mohammed Mansour was born in the city of Jenin, Palestine on 05.03.1995. He finished his Primary and Secondary school in Jenin; graduated from Jenin Secondary School for Boys in 2012 with scientific stream diploma. He started his bachelor's degree at University of Turkish Aeronautical Association, Department of Mechatronics Engineering in 2012. He graduated in 2017 and got his bachelor's degree. In 2018; he started his master's study at University of Sakarya, Department of Mechatronics Engineering.