Sudan University of Science and Technology

College of Graduate Studies

Enhancing the Mammogram image classification using Mutual Information Feature Selection

تحسين تصنيف صور الماموجرام باستخدام خوارزمية ألمعلومات المتبادلة لأختيار الصفات

A Thesis Submitted in Partial Fulfillment of the requirements for the degree of M.Sc.in Information Technology

BY:

Njwan Salim Musa Mohammed

Supervised by:

Dr. Wafaa Faisal Mukhtar

February 2021

الأيـــــــــــــة

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

قال تعالى:

"ن وَالْقَلَمِ وَمَا يَسْطُرُونَ"سورة القلم(1)

صدق الله العظيم

Dedications

This study is dedicated to my parents

My late mother MARIAM, and my father, SLEAM

For their endless love, support and encouragement

Acknowledgments

First and foremost, I have to thank **my parents** for their love and support throughout my life. Thank you both for giving me strength to reach for the stars and chase my dreams.

Thank you my husband, **Munir Ibrahim** who has encouraged me all the way and whose encouragement has made sure that I give it all it takes to finish that which I have started.

Thank you **My brothers and my sisters** for your understanding and encouragement in my many, many moments of crisis.

I would like to sincerely thank my supervisor, **Dr. Wafaa Faisal Mukhtar** for her guidance and support throughout this study, and especially for her confidence in me.

I would also like to thank **DR. Ali Ahmed Alphaki** for his efforts.....

Thank you all **my friends** Your friendship makes my life a wonderful experience.

Thank you, Lord, for always being there for me.

This thesis is only a beginning of my journey.

Abstract

The Breast Cancer is one of the main causes of death for women all over the world. With early and accurate diagnosis of the breast cancer the cure rate rises from 56% to more than 86%. The objective of this study is to enhance the classification accuracy of the mammograms images based on feature selection method to detect if the input image is normal or affected by the diseases. The accuracy of most of the classification methods depend on important features extracted from the mammogram images and the classifier itself. This study propose a classification method based on K-Nearest Neighbor (KNN) and Support vector machine (SVM) using important features selected from data set of features extracted from the mammogram images The MIAS Mini data set (Mammographic Image Analysis Society) include 209 normal images, 23 images of CIRC (Circumscribed masses), 19 images of SPIC (Speculated masses),19original images of MISC(ill-defined masses), 23 images of CALC(Calcification)) based on mutual information (MI) feature selection method.in this study the classification process includes five basic steps; beginning with the mammogram Image collection, image processing, features extraction , classification and testing and evaluation ;firstly by using all features, secondly by using more important features based on mutual information (MI) features selection method, the last step is testing and evaluation. This study used set of thirteen features, extracted from mammogram images that taken from MAIS database, then it applies K-nearest neighbors (KNN) and Support vector machine (SVM) based classification method. In this study, the dataset splited into two parts, namely: training and testing. After the construction of the classifier based on training data, the proposed model using the test data to measure the accuracy. The best accuracy obtained was 83% by KNN algorithm when using percentage of 85% and 15% for training and testing by using the most important features for the five Sub-Features best features. Using other feature selection method may results in more accuracy of the classifier.

المستخلص

يعد سرطان الثدي أحد ألاسباب الرئيسية لوفاة النساء في جميع أنحاء العالم. مع التشخيص المبكر والدقيق لسرطان الثدي ، يرتفع معدل الشفاء من 56 ٪ إلى أكثر من 86 ٪. الهدف من هذه الدراسة هو تحسين دقة تصنيف صور الماموجرام بناءً على طريقة اختيار الميزة لاكتشاف ما إذا كانت صورة الإدخال طبيعية أو متأثرة بالأمراض. تعتمد دقة معظم طرق التصنيف على الميزات الهامة المستخرجة من صور الماموجرام والمصنف نفسه. تقترح هذه الدراسة طريقة التصنيف بناءً على K-Nearest Neighbour (KNN) و Support vector machine(SVM) وأهم الميزات المحددة من مجموعة البيانات من الميزات المستخرجة من صور الماموجرام لسرطان الثدي MIAS Mini data set (Mammographic Image Analysis Society) يشمل 209 صورة عادية ، 23 صورة لـ CIRC (Circumscribed masses) ، 19 صورة لـ SPIC (Speculated masses) ، 19 صورة أصلية لـ MISC (ill-defined masses)) ، 23 صورة من CALC (Calcification) بناءً على طريقة اختيار الميزات mutual information (MI). في هذه الدراسة ، تتضمن عملية التصنيف خمس خطوات أساسية ؛ بدايةً من جمع صور الماموجرام لسرطان للثدي (mammogram image)، معالجة الصور ، استخراج الميزات ، التصنيف ، الاختبار والتقيم ؛ أولاً باستخدام جميع الميزات ، وثانياً باستخدام ميزات أكثر أهمية تعتمد على طريقة اختيار الميزات mutual information (MI) ، والخطوة الأخيرة هي الاختبار والتقييم. استخدمت هذه الدراسة مجموعة من الميزات الثلاثة عشر المستخرجة من صور الماموجرام المأخوذة من قاعدة بيانات MAIS، ثم طبقت طريقة تصنيف (KNN) ,(SVM). في هذه الدراسة ، انقسمت مجموعة البيانات إلى قسمين هما: التدريب والاختبار. بعد بناء المصنف على أساس بيانات التدريب ، فإن النموذج المقترح يستخدم بيانات الاختبار لقياس الدقة. كانت أفضل دقة تم الحصول عليها هي 83٪ بواسطة خوارزمية KNN عند استخدام نسبة 85٪ و 15٪ للتدريب والاختبار باستخدام خمسة ميزات فرعية. قد يؤدي استخدام طريقة تحديد الميزات أخرى إلى زيادة دقة التصنيف.

Table of Contents

List of tables

List of figures

Chapter One
Introduction

1.1 Data mining and Breast cancer

Data mining is the automated process of discovering interesting (non-trivial, previously unknown, insightful and potentially useful) information or patterns, as well as descriptive, understandable, and predictive models from (large-scale) data. The goals of data mining are search consistent patterns and/or systemic relationships between data, validate the findings by applying the detected patterns to new subsets of data and predict new findings on new datasets. (Jiawe Hana, Jian Pei , Micheline KAMBER, 2011)

Data Mining is all about the analysis of large amount of data usually found in data repositories in many organizations. Its application is growing in leaps and bounds and has touched every aspect of human life ranging from science, engineering to business applications. Data mining can handle different kinds of data ranging from ordinary text and numeric data to image and voice data. It's a multidisciplinary field that has applied techniques from other fields especially statistics, database management, machine learning and artificial intelligence.

With the aid of improved technology in recent years, large volumes of data are usually accumulated by many organizations and such data are usually left to waste in various data repositories. With the help of data mining such data can now be mined using different mining methods such as clustering, classification, association and outlier detection method in order to unravel hidden information that can help in improved decision making process. (Obuandike Georgina N, Audu Isah, John Alhasan, 2015).

The application area of data mining includes many disciplines such as clustering, prediction, marketing, e-commerce, medicine, intelligence, medical diagnosis and web mining. Medical diagnosis for the early detection of diseases using clinical test results. Image analysis for diagnostic purpose is another field of investigation. In Web Mining applications are intended for the analysis of so-called click streams, the sequences of page visited and the choices made by a web surfer. Which used for: the analysis of e-commerce sites, in caching the most popular pages or in evaluating the effectiveness of an e-learning training course. (KUMARI, Praveen, et, 2016).

The human body comprises of millions of cells each with its own unique function. When there is unregulated growth of any of these cells it is termed as cancer. In this, cells divide and grow uncontrollably, forming an abnormal mass of tissue called as tumor. Tumor cells grow and invade digestive, nervous and circulatory systems disrupting the bodies' normal functioning. Cancer is classified by the type of cell that is affected, more than 200 types of cancers are known. Thousands of Women fall victim of Breast Cancer every year. Recent years have seen an intense improvement in survival rates for women with breast cancer, which can be mainly attributed to an extensive screening and enhanced treatment. The recent advances in data collection and storage techniques have made it possible for various medical companies and hospitals to formally, data mining is the process of running powerful algorithms on data to extract useful information. The uses and potentials of these methodologies have found its scope in medical data. Predicting outcome of a disease is a challenging task. Data mining techniques teni1zds to simplify the prediction segment. Automated tools have made it possible to collect large volumes of medical data, which are made available to the medical research groups. The results being an increasing popularity of data mining techniques to identify patterns and relationship among large number of variables, which make it possible to predict the outcome of the disease using pre-existential datasets. (SUMBALY, Ronak, 2014)

1.2 Problem Statement

Breast cancer is one of the leading causes of death for women all over the world. With early and accurate diagnosis of breast cancer, the cure rate rises from 56% to more than 86%. Data mining is the process of running robust data algorithms to extract useful information. The uses and capabilities of these methodologies have found their scope in medical data. Predicting disease outcomes is a difficult task. Machine tools have made it possible to collect large amounts of medical data, which are provided to medical research groups. The result is the increasing popularity of data mining techniques for determining patterns and the relationship between a large number of variables, which makes it possible to predict the outcome of a disease using pre-existing data sets. But choosing the appropriate algorithm with possible features combination is always the great issue to increase the accuracy of the rating.

1.3 Research objectives

- Applying features selection algorithm.
- Selecting the most appropriate classification algorithms.
- Building a model to classify the MIAS Mini data set (Mammographic Image Analysis Society).

1.4 Research Methodology

The classification process involves five major steps namely image collection , preprocessing, feature extraction, classification and evaluation. Image collection step involves selection of images ranging from computed tomography (CT), magnetic resonance images (MRI) to x-ray etc. Pre-processing is a course of actions that is executed on raw data in order to achieve the best recital for ones datasets. feature extraction is a process to analyze objects and images to extract the most prominent features that are correspondence of various classes of objects. For classification, different applications of data mining techniques are used to predict class (group) membership for data instances.

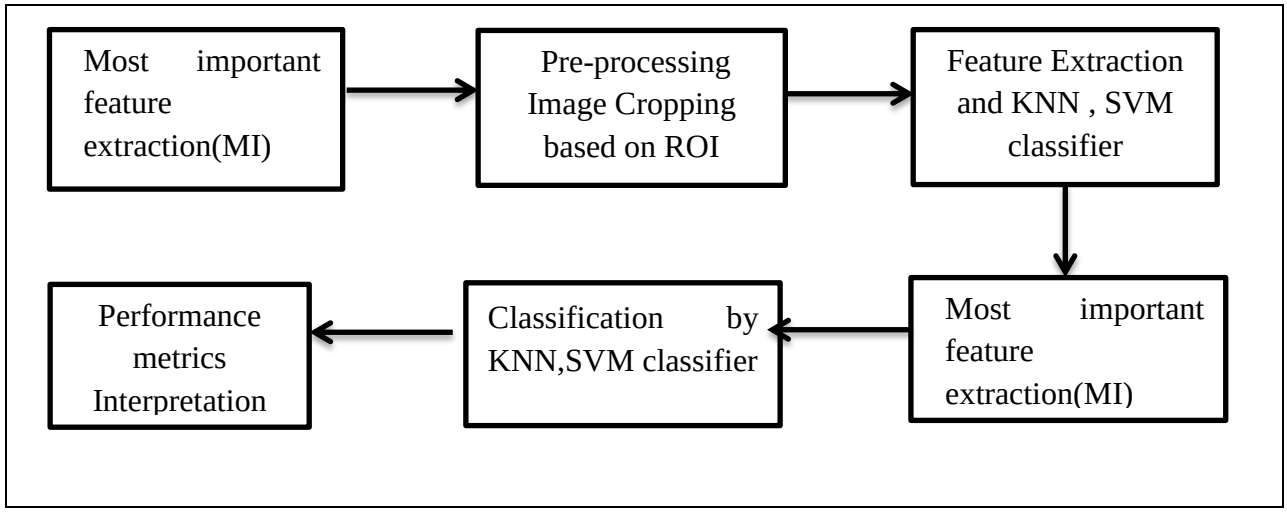


Figure (1-1) Medical Image Classification Process

1.5 Research Scope

This study covers the offline classification and considers the Mammogram images that taken from MIAS Data set. The evaluated measures that will used in this study is the Confusion Matrix (true positive, true negative, False Positive and False negative) to determine and examine the accuracy of the classifier that is used during the study.

1.6 Thesis Organization

Chapter one provided a general definition about data mining and its functionality also describe the problem statement of the study and objective, significant , expected result and the scope of the study. Chapter two is a Literature review and decision tree classification method, medical image classification and the evaluation measure. Chapter three describe the research methodology, the five phases of the study and materials that used in the study. Chapter four describes and discusses the classification results and lastly the thesis will be concluded and recommendations will be given in chapter five.

Chapter Two
Literature Review

2.1 Introduction

This chapter describes the latest studies in the classification of mammograms images of breast cancer using different techniques. He cited relevant sources of information and publications. The first part presents images classification .The second part presents the data extraction techniques especially those implemented in the educational section and the various datasets that have been used as well as the results obtained. The Third part presents the related work that applies the system of classification of mammograms images. The final summary of the literature review is mentioned in the educational section.

2.2 Images Classification

In our everyday life, classification helps us in taking decisions. The need for classification arises whenever an object is placed in a specific group or class depending upon the attributes corresponding to that object. Most of the industrial problems are classification problems. Scientists have devised advanced classification techniques for improving classification accuracy .Every single day numerous images are produced, which creates the necessity to classify them so that accessibility is easier and faster. The information processing which is done during the classification helps in image categorization into various groups (Siddhartha Sankar Nath et al, 2014).

Image classification refers to the task of extracting class's information from a multiband raster image. The resulting raster from image classification can be used to create thematic maps. Depending on the interaction between the analyst and the computer during classification, there are two types of classification: supervised and unsupervised.

Supervised Classification: It is the process of identification of classes within a remote sensing data with inputs from and as directed by the user in the form of training data.

Unsupervised Classification: It is the process of automatic identification of natural groups or structures within a remote sensing data.

Both the classification approaches differ in the way the classification is performed. In the case of supervised classification, specific land cover types are delineated based on statistical characterization of data drawn from known examples in the image (known as training sites). Both these methods can be combined together to come up with a 'hybrid' approach of image classification. In the hybrid classification, firstly, an unsupervised classification is performed, then the result is interpreted using ground referenced information and, finally, original image is reclassified using a supervised classification with the aid of statistics of unsupervised classification as training knowledge. This method uses unsupervised classification in combination with ground referenced information as a comprehensive training procedure and, therefore, provides more objective and reliable results.

2.3 Classification Methods

Classification is the process of finding a set of models that describe and differentiate data classes and concept , also is a group of records, any record containing member group of attributes and one of those attributes the words of class, the objective is to eventually be the records would be allocated to accurately class if possible. Divides the data set to two sections that are training set is used to build the model and test set which is used to determine the accuracy of model. Data classification is a two-step process Learning step, where a classification model is constructed, and Classification step, where the model is used to predict class labels for given data (Chhabra, G. Kaur and A., 2014).

2.3.1 Naive Bayésien method

The Naive Bayesian classifiers are statistical classifiers. They can predict class membership probabilities such as the probability that a given tuple belongs to a particular class. Naive Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. (Jiawe Hana, Jian Pei , Micheline KAMBER, 2011).

2.3.2 K -nearest neighbor classifier

K-Nearest neighbor classifier is based on learning by analogy. The training samples are described by n-dimensional numeric attributes. Each sample represents a point in an n-dimensional space. In this way, all of the training samples are stored in an n-dimensional pattern space. When given an unknown sample, a k -nearest neigh b or classier searches the pattern space for the k training samples that are closest to the unknown sample.

2.3.3 Artificial Neural Network (ANN)

An artificial neural network consists of an interconnected group of artificial neurons and is trained to perform a particular function by adjusting the values of the connections (weights and biases) between the neurons of different layers. Neural network is defined by the interconnection pattern between different layers of neurons, the learning process and the activation function of the neurons. Input and target pairs are needed to train a neural network. The weights are adjusted, based on a comparison of the output and the target, until the network output matches the target. The number of features extracted from transforms decides the number of neurons in the input layer of neural network. The number of neurons in the output layer is made equal to the elements of target vector (R. Kumar, B. Singh, D. Shahani, A. Chandra, and K. Al-Haddad,, 2015).

2.3.4 Support Vector Machine (SVM)

Support Vector Machine" (SVM) is a supervised machine learning algorithm invented by vapnik in 1960,,s (S. Sharma, J. Agrawal, S. Agarwal, and S. Sharma,, 2013) The support vector machine is a training algorithm for classification rule from the data set which trains the classifier; it is then used to predict the class of the new sample. SVM is based on the concept of decision planes that define decision boundary and point to form the decision boundary between the classes called support vector threat as parameter, and structure risk minimization principle to prevent over fitting. There are 2 key implementations of SVM technique: mathematical programming and kernel function (S. Sharma, J. Agrawal, and S. Sharma,, 2013).

2.4 Related work

(Mohsen, H., El-Dahshan, E. S. A., El-Horbaty, E. S. M., & Salem, 2018) Used Deep Neural Network classifier which is one of the DL architectures. to classifying a dataset of 66 brain MRIs

into 4 classes which are; normal, glioblastoma, sarcoma and metastatic bronchogenic carcinoma tumors. The classifier was combined with the discrete wavelet transform (DWT) the powerful feature extraction tool and principal components analysis (PCA) and the evaluation of the performance was quite good over all the performance measures.

(M Manoj krishna, M Neelima, M Harshali and M Venu Gopala Rao, 2018) studied the image classification using deep learning. Used Alex Net architecture with convolutional neural networks for that purpose. Four test images were selected from the Image Net database for the classification purpose. The images were cropped for various portion areas and conducted experiments. The results show the effectiveness of deep learning based image classification using Alex Net.

Neural networks have been used by (Sertan Kaymaka , Abdulkader Helwana, Dilber Uzun, 2017) , (Saira Charan, Muhammad Jaleed Khan, Khurram Khurshid, 2018) and (Sandhya G 1, D Vasumathi2, G T Raju3, 2015) on three different databases of breast cancer (MIAS data set, DDSM data set, Obtained data set from Near East University Hospital by using classification accuracy and confusion matrix. Sandhya, D Vasumathi and G T Raju used Gray Level co-occurrence matrix (GLCM) as a feature extraction method. Experimental results are shown when applying neural networks with DDSM data set which was more accurate than the other data sets that used.

(Animesh Hazra, Subrata Kumar Mandal, Amit Gupta, 2016) compare different classifiers Naïve Bayes (TAN, BAN ,BBN) Support Vector Machine and Ensemble, The authors used the Wisconsin Diagnosis Breast Cancer(WDBC) data set by using classification accuracy , PCA is used for feature extraction and mapped the data into a lower dimensional space (here five dimensional space have been taken). And also combine classifiers to get better accuracy. the experimental results show that Naïve Bayes is the best classification technique having least time complexity and it gives better classification accuracy with only five dominant features after introduction of the binning concept ,TAN also produced the best performance with respect to rating and accuracy compare with BAN and BBN.

Breast cancer images in DICOM (Digital Imaging and Communications in Medicine) for 250 patients at the Cancer Institute, Adyar, Chennai, Tamilnadu, India are taken by

(B.Padmapriya , T.Velmurugan, 2016) for classification using J48 Algorithm, Classification And Regression Tree (Cart) and Alternating Decision Tree(Adtree) and The accuracy of taken algorithms is measured by various measures like specificity, sensitivity and kappa statistics (Errors). The classifiers J48 have accuracy of 98.1 %, ADTree have 97.7% and highest accuracy value 98.5 % is found in CART. also as Ahamed Lebbe Sayeth Saabith, Elankovan Sundararajan, Azuraliza Abu Bakar used J48,MLP and Rough set classifiers for c data Set classification that are taked from UCI machine learning repository . The results of the experiment showed higher accurate (79.97%) J48 algorithm

(Shofwatul 'Uyun, Lina Choridah, 2018) and (Dr. R. J. Ramteke, Khachane Monali, 2012) proposed system used the KNN classifier compared with kernel based SVM classifier (Linear and RBF), decision tree, and Bayesian naive. esearchers Shofwatul 'Uyun, Lina Choridah to focused on the feature selection from primary data in the form of mammogram image produced by digital mammography imaging technology .The best classification results based on the five features (slice, integrated density, Area fraction, gray capital value, center of mass) are generated by the decision tree algorithm with accuracy, sensitivity, specificity, FPR and TPR of 93.18%; 87.5%; 3.89%; 6.33% and 92.11%. but Dr. R. J. Ramteke, Khachane Monali6 proposed system used the KNN classifier compared with kernel based SVM classifier (Linear and RBF) for classifying image on real data used in this work collected from CT scan centers. The data contain normal as well as abnormal CT scan brain images. using confusion matrix computed result shows that KNN obtain 80% classification rate which is more than SVM classification rate.

The objective of this paper is to build a CAD model to discriminate between cancers, benign, and healthy parenchyma. for experimental purpose, Digital Database for Screening Mammography (DDSM) is used. neural networks classifier to the classification is used. The experimental results are obtained from DDSM data set for different types. used both Receiver Operating Characteristics (ROC) and Confusing matrix to measure the performance of different classifiers. obtained overall classification accuracy of 89%, with 88.6% sensitivity and 83.3% specificity (Sandhya G 1, D Vasumathi2, G T Raju3, 2015). Table (2-1) shows a summary for the related work

Table (2-1): summary of related work

Author	Dataset	Classification techniques	Feature selection method	Accuracy
(Xiaoming Liu,et al., 2012) (Dr. R. J. Ramteke, Khachane Monali, 2012)	real data from 51 CT scan centers.	K-Nearest Neighbour		80%
		Support Vector Machine		67% with linear kernel
				69% With RBF
(Ahamed Lebbe Sayeth Saabith, 2014)	UCI data Set with 286 records .	J48		79.97%
		MLP		75.35%
		Rough set		71.36%
(Sandhya G 1, D Vasumathi2, G T Raju3, 2015)	DDSM data set 410 image	Neural Network		89%
(Animesh Hazra, Subrata Kumar Mandal, Amit Gupta, 2016)	WDBC Dataset 569 instances 32 attributes	Naïve Bayes		97.3978%.
		Support Vector Machine		95.5390%
		Ensemble		95.9108%
B.Padmapriya, T.Velmurugan (2016)	Breast cancer images in	Alternating Decision Tree		97.70

	DICOM for 250 patients	(Adtree)		
		J48 Algorithm		98.10%
		Classification And Regression Tree (Cart)		98.50%
(Sertan Kaymaka , Abdulkader Helwana, Dilber Uzun, 2017)	Obtained data set from Near East University Hospital 176 images 122 images abnormal 64 images normal	Neural Network (BPPN) and (RBFN		59.0% with (BPPN 70.4% with (RBFN)
(Saira Charan, Muhammad Jaleed Khan, Khurram Khurshid, 2018)	Mammographic Image Analysis Society(MIAS) dataset 322 images 133 abnormal images 189 image normal	Neural Network		65%
(Shofwatul 'Uyun, Lina Choridah, 2018)	primary data in the form of mammogram image 79 benign lesions 38 malignant lesions	Decision tree		82,05%
		k-nearest neighbors		76,29%
		Naïve Bayes		60,83%
		Alternating Decision Tree		97.70%

		(Adtree)		
(Bazila Banu , and Ponniah Thirumalaikolundusubramanian, 2018)	WDBC data set 569 instances 32 attributes.	Tree Augmented Naive Bayes		94.11%
		Boosted Augmented Naive Bayes		91.7%
		Bayes Belief Network		91.7
(Mohsen, H., El-Dahshan, E. S. A., El-Horbaty, E. S. M., & Salem, 2018)	real dataset 66 human brain MRIs 22normal images 44 abnormal images	Deep Neural Network (DNN).		0.97

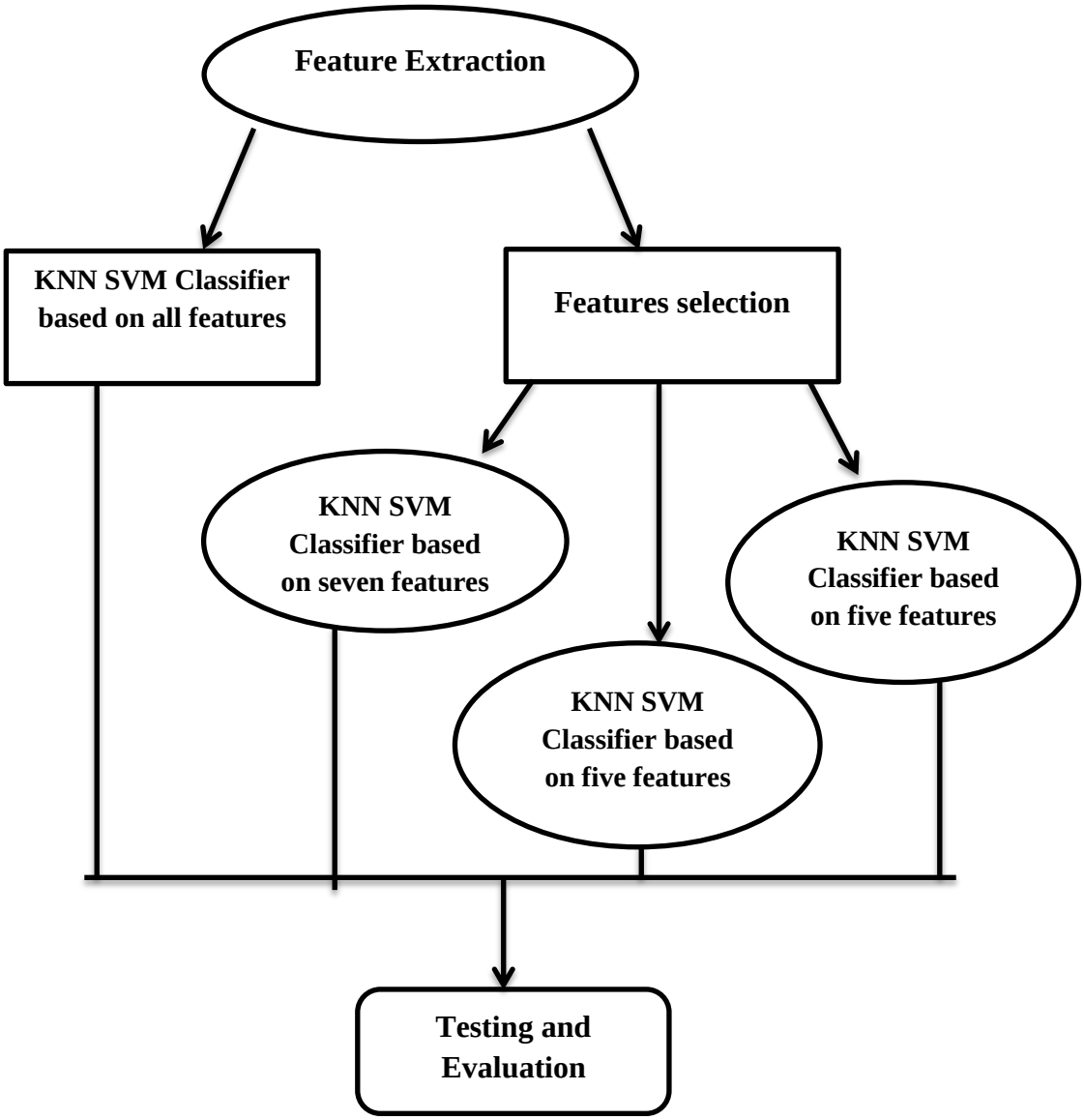
2.5 Summary

The reviewed literature so far concludes that Breast cancer is one of the most dangerous diseases women face. The computerized breast cancer classification system is the latest widely used method which relies mainly on the accuracy of the classification. In this chapter the most important classification techniques used to classify mammograms images of breast cancer are presented. Moreover, Table (2-1) summarized the techniques discussed in the related work. Increased ranking accuracy has been observed when using feature selection techniques. This study will adopt a methodology according to CAD system based on data mining techniques which will be explained broadly in next chapter

Chapter Three
Research Methodology

3.1 Introduction

The research phases include five phases starting in collect images, preprocessing, features extracting, classification and end with testing and evaluating. In this chapter, each stage will be explained in detail.



Figure(3-1): Research phases

3.1.1 Phase (1) Mammogram images collection

Dataset used in this study downloaded from the MIAS Mini data set (Mammographic Image Analysis Society) The MIAS are UK based research groups that work on breast cancer.

3.1.1.1 Dataset

All images are digitized at the resolution of 1024 × 1024 pixels and 8 bit accuracy (gray level). The testing images include 209 normal images, 23 images of CIRC (Circumscribed masses), 19 images of SPIC (Speculated masses),19 original images of MISC(ill-defined masses), 23 images of CALC(Calcification).

Table (3-1): The MIAS Database Details

Column No	**Description**	**Details**
1st	MIAS database reference number	
2nd	Character of background tissue	F – Fatty G - Fatty-glandular D - Dense-glandular
3rd	Class of abnormality present	CALC - Calcification CIRC–Well defined/circumscribed SPIC - Speculated masses MISC - Other, ill-defined masses ARCH - Architectural distortion ASYM - Asymmetry NORM – Normal
4th	Severity of abnormality	B – Benign M – Malignant
5th, 6th	X ,y image-coordinates of center of abnormality	
7th	Approximate radius (in pixels) of a circle enclosing the abnormality	

3.1.2 Phase (2) Preprocessing (ROI) image cropping

A preprocessing phase of the images is necessary to improve the quality of the images and make the feature extraction phase more reliable. Mammograms are medical images that are difficult to interpret, thus a preprocessing phase is needed to improve the image quality and make the classification results accurate.The extracted portion of ROI of mammography with benign and malignant.

Extracting the Region of Interest (ROI) using the function from [x] position to [y] position and [radius] depend of the MIAS dataset using the following function in matlab:

```
function [ m1 ] = moment1( A )
%this function compute the first moment for 2-D array
m1=mean2(A);
end
```

Figure (3-2): Extracting the Region of Interest (ROI) matlab code.

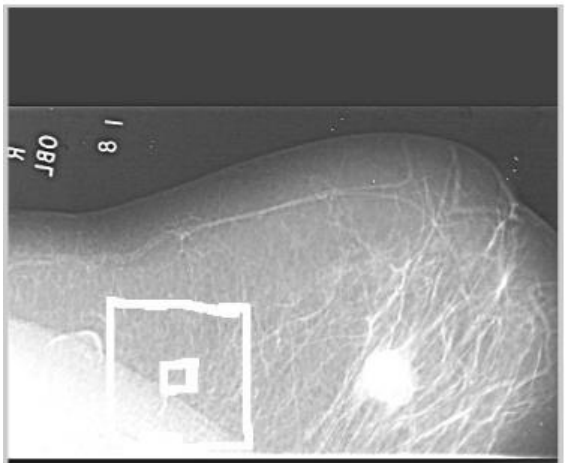


Figure (3-3): the image before Detecting the Region of Interest

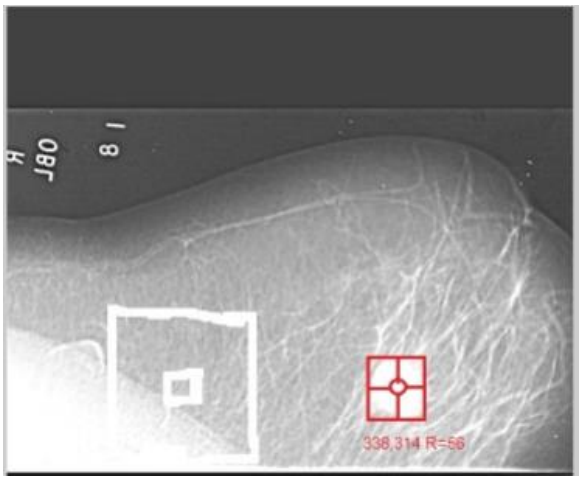


Figure 3-4): the image After Detecting the Region of Interest

3.1.3 Phase (3) Features Extraction

Feature extraction is a process to analyze objects and images to extract the most prominent features that are correspondence of various classes of objects. It is lead to that improving feature extraction process will be likely improving performance of a described classification algorithm.

Extract "Mean, Standard Deviation, Skewness, Smoothness ,Kurtosis, Entropy, Graythresh, Contrast, Homogenerty, Conrelohan, Energy, Max, Min" features from each mammogram image after cropping.

The following code sections illustrate the part of features extraction by process in this study:

3.1.3.1 Mean

The Mean is a measure of the average intensity of the neighboring pixels of an image.

$$\textbf{Mean} = \sum_{i=0}^{l-1} z_i * p(z_{i)}$$

Equation (3.1): Mean

```
function [ m1 ] = moment1( A )
%this function compute the first moment for 2-D array
m1=mean2(A);
end
```

Figure (3-5): Mean

3.1.3.2 Standard Deviation

The Standard Deviation is a measure of how spread out numbers is

$$\mathbf{Std} = \sum_{i=0}^{l-1} (z_i - m)^2 * p(z_i)$$

Equation (3. 2): Standard Deviation

```
function [ stda ] = moment2( A )
%UNTITLED2 Summary of this function goes here
m=size(A,1);
n=size(A,2);
N=m*n;
E=moment1(A);
stda=sqrt(sum((sum((A-E).^2)))/N);
end
```

Figure 3-6): Standard Deviation

3.1.3.3 Skewness

The Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point. The Skewness for a normal distribution is zero, and any symmetric data should have Skewness near zero. Negative values for the Skewness indicate data that are skewed left and positive values for the Skewness indicate data that are skewed right.

$$\mathbf{Skewness} = \sum_{i=0}^{l-1} (z_i - m)^2 * p(z_i)$$

Equation (3. 3): Skewness

```
function [ skew ] = moment3( A )
%UNTITLED3 Summary of this function goes here
m=size(A,1);
n=size(A,2);
N=m*n;
E=moment1(A);
skew=(sum((sum((A-E).^2)))/N).^(1/3);
end
```

Figure(3-7): Skewness

3.1.3.4 Smoothness

Measures the relative intensity variations in a region

$$\textbf{Smoothness} = 1 - \frac{1}{(1 + \sigma^2)}$$

Equation (3. 4): Smoothness

```
function [ smooth ] = smoothness( A )
%UNTITLED4 Summary of this function goes here
m=size(A,1);
n=size(A,2);
N=m*n;
v=moment2(A);
smooth=(1-(1/(1+v)));
end
```

Figure 3-8): Smoothness

3.1.3.5 Kurtosis

The Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. That is, data sets with high kurtosis tend to have a distinct peak near the mean.

$$\textbf{Kurtosis} = \sum_{i=0}^{l-1} (z_i - m)^4 * p(z_i)$$

Equation (3. 5): Kurtosis

```
function [ kurtos ] = kurtosis( A )
%UNTITLED2 Summary of this function goes here
m=size(A,1);
n=size(A,2);
N=m*n;
E=moment1(A);
kurtos=(sum((sum((A-E).^4)))/N).^(1/4);
end
```

Figure(3-9): Kurtosis

3.1.3.6 Contrast

The Contrast is the difference in luminance and/or color that makes an object (or its representation in an image or display) distinguishable. In visual perception of the real world, contrast is determined by the difference in the color and brightness of the object and other objects within the same field of view.

$$\textbf{Contrast} = \sum_{i=0}^{l-1} \sqrt{(z_i - m)^2 * p(z_i)}$$

Equation (3. 6): Contrast

```
function [ cont ] = contrast( A )
%UNTITLED3 Summary of this function goes here
m=size(A,1);
n=size(A,2);
N=m*n;
E=moment1(A);
v=moment2(A);
cont=sqrt(v);
end
```

Figure(3-10): Contrast

3.1.3.7 Entropy

Statistical measure of randomness that can be used to characterize the texture of the input image.

$$\mathbf{Entropy} = -\sum_{k=0}^{i=1} pik(\log_2 pik)$$

Equation (3. 7): Entropy

3.1.3.8 Graythresh

Measure compute a normalized intensity value that lies in the range [0, 1].The graythresh function uses Otsu's method, which chooses the threshold to minimize the intraclass variance of the black and white pixels.

$$\mathbf{Level} = \text{graythresh(I)}$$

Equation (3. 8): Graythresh

3.1.3.9 Homogeneity Inverse Difference Moment (IDM)

IDM is also influenced by the homogeneity of the image. Because of the weighting factor (1+ (i−j) 2) −1 IDM will get small contributions from inhomogeneous areas (i 6= j).

$$\mathbf{IDM} = \sum_{I=0}^{G-1} \sum_{I=0}^{G-1} 1/(1 + (i + j)p(i,j)$$

Equation (3. 9): IDM

3.1.3.10 Correlation

Is measure of gray level linear dependence between the pixels at the specified positions relative to each other.

$$\mathbf{Correlation} = \sum_{I=0}^{G-1} \sum_{I=0}^{G-1} \{i * j\} * p(i,j)/\{\delta * \delta y\}$$

Equation (3. 10): Correlation

3.1.3.11 Energy

Return the sum of squared element in the Gray Level Co-occurrence Matrix (GLCM) the renge of energy is [0, 1].

$$\textbf{Energy} = \sum_{i\,j} p^{(i\,j)}$$

Equation (3. 11): Energy

```
f1=moment1(roi_of_img);  % Mean
f2=moment2(roi_of_img);  % STD
f3=moment3(roi_of_img);  % Skew
f4=smoothness(roi_of_img);
f5=kurtosis(roi_of_img);
f6=entropy(roi_of_img);
f7=graythresh(roi_of_img);
glcm=graycomatrix(roi_of_img);
F=graycoprops(glcm,{'Contrast','Homogeneity','Correlation','Energy'});
f8=F.Contrast;
f9=F.Homogeneity;
f10=F.Correlation;
f11=F.Energy;
f12=max(max(roi_of_img));
f13=min(min(roi_of_img));
```

Figure (3-11): entropy, graythresh, homogeneity, correlation, energy, max, min.

The following snapshot is a part of numeric values of the thirteen features extracted from ROI of each image used in this study.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Mean	Standard Deviation	Skewness	Smoothness	Kurtosis	Entropy	Graythresh	Contrast	Homogenerty	Conrelohan	Energy	Max	Min
2	94	11	5	1	3	6	0	0	1	1	0	221	0
3	134	10	5	1	3	7	1	0	1	1	0	213	79
4	0	0	0	0	0	0	0	0	1	0	1	0	0
5	168	4	3	1	3	5	1	0	1	1	1	187	152
6	140	7	4	1	3	5	1	0	1	1	1	168	113
7	8	6	3	1	3	3	0	0	1	1	1	72	0
8	0	0	0	0	0	0	0	0	1	0	1	0	0
9	185	7	4	1	3	6	1	0	1	1	0	213	146
10	156	6	3	1	3	5	1	0	1	1	0	191	138
11	0	0	0	0	0	0	0	0	1	0	1	0	0
12	179	5	3	1	3	5	1	0	1	1	1	198	163
13	123	7	4	1	3	6	0	0	1	1	0	164	85
14	133	8	4	1	3	6	1	0	1	1	0	191	106
15	10	5	3	1	3	3	0	0	1	1	1	43	5
16	39	8	4	1	3	6	0	0	1	1	0	188	3
17	144	4	2	1	3	4	1	0	1	0	1	170	132
18	164	12	5	1	4	6	1	0	1	1	0	216	92
19	103	8	4	1	3	6	0	0	1	1	1	138	67
20	26	8	4	1	3	6	0	0	1	1	1	61	4
21	170	7	4	1	3	5	1	0	1	1	1	193	144
22	0	0	0	0	0	0	0	0	1	0	1	0	0
23	163	11	5	1	3	7	1	0	1	1	0	232	14
24	0	0	0	0	0	0	0	0	1	0	1	1	0
25	171	12	5	1	4	7	1	0	1	1	0	227	43
26	128	7	4	1	3	5	1	0	1	1	0	161	109
27	7	4	3	1	2	2	0	0	1	0	1	30	3
28	130	7	4	1	3	5	1	0	1	1	0	205	103

Figure(3-12): The thirteen features extracted from ROI of each image

3.1.4 Phase (4) Features Selection

Data sets for analysis may contain hundreds of attributes, many of which may be irrelevant to the mining task, or redundant. Although it may be possible for a domain expert to pick out some of the useful attributes, this can be a difficult and time consuming task, especially when the behavior of the data is not well known. Leaving out relevant attributes or keeping irrelevant attributes may be detrimental, causing confusion for mining algorithm employed .Thus the dimensionality reduction reduces the data size by removing such attributes from it. The method called attribute subset selection is applied to reduce the data size The goal of attribute subset selection is to find a minimum set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes. Mining on a reduced set of attributes has an additional benefit. It reduces the number of attributes appearing in the discovered patterns, helping to make the patterns easier to understand. In this research used Mutual information to feature selection.

3.1.4.1 Mutual Information

The mutual information (MI) is a measure of the amount of information that one random feature has about another feature (Jorge R. VergaraPablo A. Estévez, 2013).

Mutual Information algorithm provides only a score value for each feature to reflect its usefulness. In order to use these feature scoring method for subset determination, additional considerations are needed to determine the size of the subset. The feature selection method (MI) only provides a scoring and associated ranking of features, using different criteria. The size of the feature set selected by using these methods has to be estimated using some additional algorithm. The MI method result in a scoring and ranking of features, according to which chosen number of features having the highest values can be selected.

After applied Mutual Information on all features to rank and select important features; obtained two excel files (seven Sub-Features) contained the important seven features, and (five Sub-Features) contained the important top five features.

The following snapshots are part of numeric values of the seven Sub-Features and five Sub-Features ranked and selected used by mutual Information (MI).

	A	B	C	D	E	F	G
1	Max	Mean	Min	Standard Deviation	Energy	Contrast	Kurtosis
2	221	94	0	11	0	0	3
3	213	134	79	10	0	0	3
4	0	0	0	0	1	0	0
5	187	168	152	4	1	0	3
6	168	140	113	7	1	0	3
7	72	8	0	6	1	0	3
8	0	0	0	0	1	0	0
9	213	185	146	7	0	0	3
10	191	156	138	6	0	0	3
11	0	0	0	0	1	0	0
12	198	179	163	5	1	0	3
13	164	123	85	7	0	0	3
14	191	133	106	8	0	0	3
15	43	10	5	5	1	0	3
16	188	39	3	8	0	0	3
17	170	144	132	4	1	0	3
18	216	164	92	12	0	0	4
19	138	103	67	8	1	0	3
20	61	26	4	8	1	0	3
21	193	170	144	7	1	0	3
22	0	0	0	0	1	0	0
23	232	163	14	11	0	0	3
24	1	0	0	0	1	0	0
25	227	171	43	12	0	0	4
26	161	128	109	7	0	0	3
27	30	7	3	4	1	0	2
28	205	130	103	7	0	0	3
29	133	60	8	10	0	0	3

Figure (3-13): The seven Sub-Features of selected by MI method

	A	B	C	D	E
1	Max	Mean	Min	Standard Deviation	Energy
2	221	94	0	11	0
3	213	134	79	10	0
4	0	0	0	0	1
5	187	168	152	4	1
6	168	140	113	7	1
7	72	8	0	6	1
8	0	0	0	0	1
9	213	185	146	7	0
10	191	156	138	6	0
11	0	0	0	0	1
12	198	179	163	5	1
13	164	123	85	7	0
14	191	133	106	8	0
15	43	10	5	5	1
16	188	39	3	8	0
17	170	144	132	4	1
18	216	164	92	12	0
19	138	103	67	8	1
20	61	26	4	8	1
21	193	170	144	7	1
22	0	0	0	0	1
23	232	163	14	11	0
24	1	0	0	0	1
25	227	171	43	12	0
26	161	128	109	7	0
27	30	7	3	4	1
28	205	130	103	7	0
29	133	60	8	10	0

Figure(3-14): The five Sub-Features of selected by MI method

3.1.5 Phase (4) Training

Data classification is a two-step process, consisting of a learning step (where a classification model is constructed) and a classification step (where the model is used to predict class labels for given data). In the first step, a classifier is built describing a predetermined set of data classes or concepts. This is the learning step (or training phase), where a classification algorithm builds the classifier by analyzing or "learning from" a training set made up of database tuples and their associated class labels.

In the second stage, the classification model constructed previously is used to classify unknown classes' data which is known as a testing

3.1.6 Phase (5) Testing and Evaluation

The test data used to determine the accuracy of our model. Usually, the test data is not a part of the training size. After passed the data set to our classifier, four components of confusion

matrix (TP, FN, TN and FP) are computed and used to calculate the classification accuracy of our model.

The training and testing phases mentioned above will based on k-nearest neighbor and algorithm as explained and below:

3.1.7 The k-nearest neighbor (KNN).

The K-Nearest Neighbor (KNN) is the simplest method of machine learning. It is a type of instance base learning in which object is classified based on the closest training example in the feature space. The KNN algorithm is sensitive to the local structure of the data set. The special case when k = 1 is called the nearest neighbor algorithm. The best choice of k depends upon the data set; larger values of k reduce the effect of noise on the classification but make boundaries between classes less distinct. KNN has some strong consistent results. KNN has several main advantages simplicity, effectiveness and Easy to understand and implement classification technique. It is effective if the training data is large. While has disadvantages can have poor run-time performance when the training set is large. It is very sensitive to irrelevant or redundant feature (S. Sharma, J. Agrawal, and S. Sharma, 2013.).

The following paragraph explains step by step on K-nearest neighbor's algorithm.

1. Determine parameter K = number of nearest neighbors

2. Calculate the distance between the query-instance and all the training samples.

3. Sort the distance and determine nearest neighbors based on the K-the minimum distance

4. Gather the category (class) of the nearest neighbors

5. Use simple majority of the category of nearest neighbors as the prediction value of the query instance.

3.1.8 Support Vector Machine (SVM)

A method for the classification of both linear and nonlinear data. In a brief, an SVM is an algorithm that works as follows. SVM transform the original training data into a higher

dimension using nonlinear mapping. Within this new dimension, it searches for the linear optimum separating hyper-plane to differentiate the tuples among the sets. With an appropriate nonlinear mapping to an adequate high dimension, data from two sets can always be separated by a hyper-plane. The SVM finds this hyper-plane with the help of support vectors ("essential" training tuples) and margins (defined by the support vectors) .An unlimited number of separating lines that could be drawn here. The target is to identify the "best" one which will have the minimum classification error on preceding unseen tuples (Animesh Hazra,Surbrata Kumar Mandal,Gubta, 2016).

3.1.9 Evaluation measurement (Confusion Matrix)

A confusion matrix contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. A confusion matrix illustrates the accuracy of the solution to a classification problem. When evaluating a classifier, there are different ways of measuring its performance. For supervised learning with two possible classes, all measures of performance are based on four numbers obtained from applying the classifier to the test set. These numbers are called true positives TP, false positives FP, true negatives TN, and false negatives FN (Elkan, C, 2012).

Table (3-2): Confusion Matrix

True Class	Predicted Class	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

True Positives (TP): These are cases in which we predicted yes (they have the disease), and they do have the disease.

True Negatives (TN): We predicted no, and they don't have the disease.

 False Positives (FP): We predicted yes, but they don't actually have the disease. (Also known as a "Type I error.")

 False Negatives (FN): We predicted no, but they actually do have the disease. (Also known as a "Type II error.")

The evaluation measure most used in practice is the accuracy rate (Acc). It evaluates the effectiveness of the classifier by its percentage of correct predictions. Equation (1) shows how Acc is computed.

$$\boldsymbol{Acc} = \frac{|\boldsymbol{TN}| + |\boldsymbol{TP}|}{|\boldsymbol{FN}| + |\boldsymbol{FP}| + |\boldsymbol{TN}| + |\boldsymbol{TP}|}$$

Equation (3. 12): Accuracy

The Recall (R) and specificity (Spe) measures evaluate the effectiveness of a classifier for each class in the binary problem. The recall, also known as sensitivity or true positive rate, is the proportion of examples belonging to the positive class which were correctly predicted as positive.

The specificity is the percentage of negative examples correctly predicted as negative R and Spe are given by equations 3 and 4, respectively

$$\boldsymbol{R} = \frac{|\boldsymbol{TP}|}{|\boldsymbol{TP}| + |\boldsymbol{FN}|}$$

Equation (3. 13): Recall

$$\boldsymbol{Spe} = \frac{|\boldsymbol{TN}|}{|\boldsymbol{FP}| + |\boldsymbol{TN}|}$$

Equation (3. 14): Specificity

```
out_lables=[TestOutputs actual_target];
tp_fn=length(find(out_lables(:,1)==out_lables(:,2)));
Accuracy1=tp_fn/s2(1)


x1=find(out_lables(:,1)==out_lables(:,2));
y1=out_lables(x1,:);
TP=length(find(y1(:,1)==1))
TN=length(find(y1(:,1)==0))
x2=find(out_lables(:,1)~=out_lables(:,2));
y2=out_lables(x2,:);
FP=length(find(y2(:,1)==1))
FN=length(find(y2(:,1)==0))
Accuracy=(TP+TN)/(TP+FN+FP+TN)
MisclassificationRate=TP / x2;
FalsePositiveRate=FP / x1;
Specificity=TN /x1;
```

Figure (3-15): Evalution measurement

3.1.10 Matlab

MATLAB (matrix laboratory) is a multi-paradigm numerical computing environment and fourth-generation programming language. A proprietary programming language developed by Math Works, MATLAB allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs written in other languages, including C, C++, Java, Fortran and Python (Grant, M., S. Boyd, and Y. Ye, CVX, 2008).

3.1.11 Summary

In this chapter the proposed methodology for this study is explained in five stages ; first stage: downloaded from the MIAS Mini data set .The MIAS are UK based research groups that work on breast cancer , second stage: preprocessing (ROI) image cropping extracting the Region of Interest (ROI) , third stage: features extraction (Extract “Mean, Standard Deviation, Skewness, Kurtosis, Contrast, Smoothness, Entropy, Graythresh, Homogenerty, Conrelohan, Energy, Max, Min”), fourth stage: features selection used mutual information(IM),fifths stage: KNN , SVM Classifier based on important features, in final evolution measurement used Confusion Matrix.

Chapter Four
Experiments Results and Discussions

4.1 Introduction

In this chapter we will discussion of the results after apply KNN and SVM using all the features (Mean, Standard Deviation, Skewness, Smoothness ,Kurtosis, Entropy, Graythresh, Contrast, Homogenerty, Conrelohan, Energy, Max, Min) , using seven Sub-Features (Max, Mean, Min, Standard Deviation , Energy , Contrast , Kurtosis) and using five sub features (Max, Mean, Min, Standard Deviation , Energy). K-nearest neighbor (KNN) and Support vector machine (SVM) algorithm with using Features selection method (MI) prove that more features is lead to reduce classification accuracy results, important features selection can give good classification accuracy results.

4.2 Implementation

The implementation of the proposed methodology used MATLAB program. The proposed methodology applied for five stages, first stage download mammogram images from the MIAS Mini Mammographic Database, second stage: mammography images are cropped based on Region of Interest (ROI),third stage apply KNN and SVM classification algorithm using all the features and calculate the accuracy of classification result, fourth stage use Feature Selection method (MI) to rank and select most important features and divide them to sub features selection (seven Sub-Features and five Sub-Features), fifth stage apply KNN and SVM classification algorithm after features selection and show the result accuracy for each sub features selection (seven Sub-Features, five Sub-Features).

4.2.1 Result of apply KNN and SVM using all the features

The dataset were divided into three groups (60-40), (70-30), and (85-15) then applied KNN and SVM classification algorithm using all the features and calculate accuracy of classification result. Table (4_ 1) show the classifications results after five training for all features , Figures (4.1_4.2) diagram shows the KNN, SVM classifier accuracy, Recall and Specificity result based on all Features.

Table (4-1): the classifications results after five training for all features.

Split dataset	SVM			KNN		
	Accuracy	**Recall**	**Specificity**	**Accuracy**	**Recall**	**Specificity**
60_40	0.57	0.4	0.70	0.59	0.5	0.7
70_30	0.57	0.3	**0.8**	0.58	**0.65**	0.52
85_15	**0.61**	**0.63**	0.6	**0.67**	0.5	**0.8**

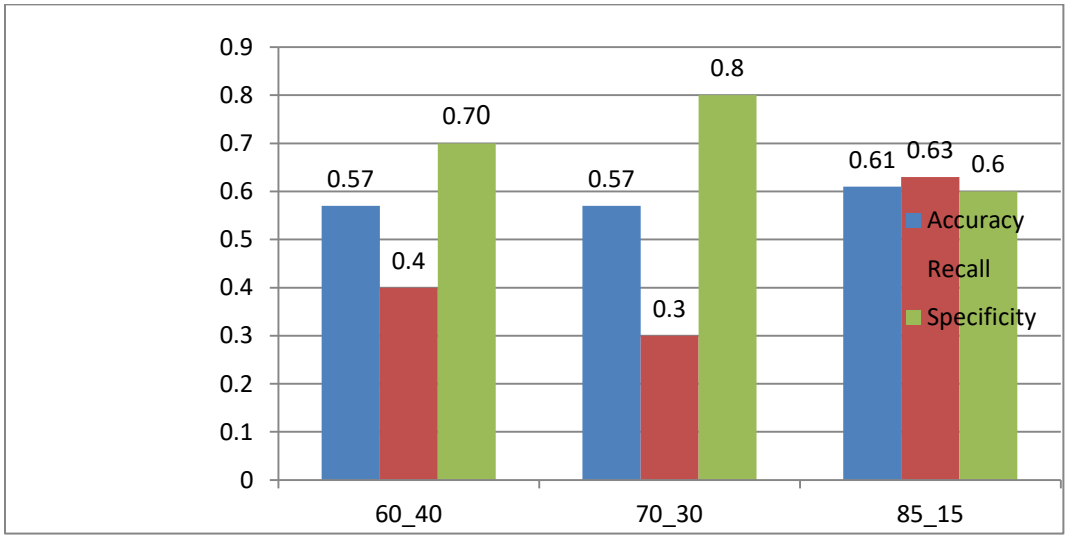


Figure (4-1): SVM Classifier result based on all Features

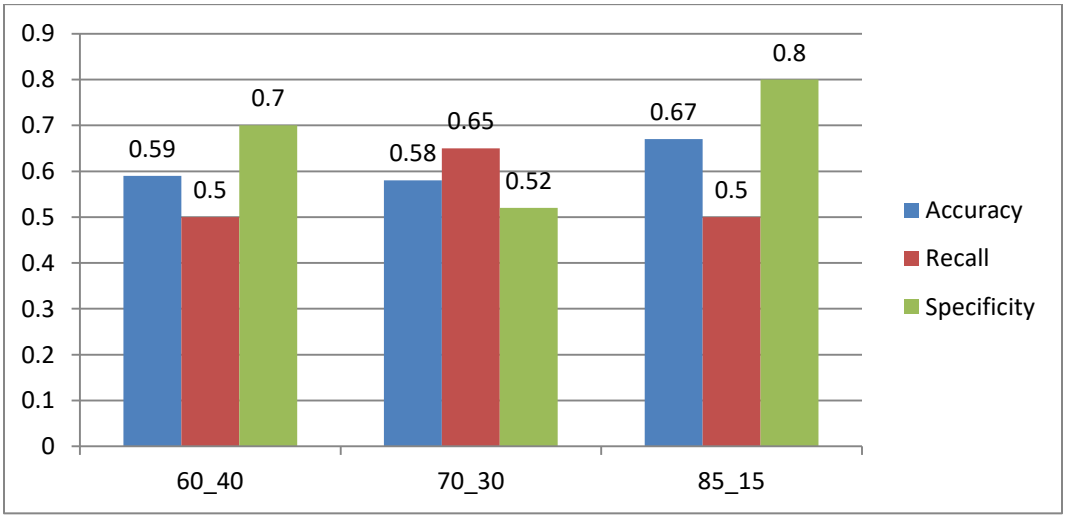


Figure (4-2): KNN Classifier result based on all Features.

4.2.2 Result of apply KNN and SVM using seven sub-features

When applying the KNN and SVM classification algorithm after selecting the important features (seven Sub-Features), the classification accuracy result is better than the result of all the features. Table (4.2) show the classification accuracy results after five training using KNN and SVM classifier based on features selected seven Sub-Features, Figure (4.3_4.4) diagram shows an accuracy, Recall and Specificity result for the KNN, SVM classifier based on seven Sub-Features

Table (4-2): KNN, SVM classifications results after five training based on seven Sub-Features

Split dataset	**SVM**			**KNN**		
	Accuracy	**Recall**	**Specificity**	**Accuracy**	**Recall**	**Specificity**
60_40	0.70	**0.6**	0.78	0.68	0.45	**0.85**
70_30	0.71	0.5	0.78	0.74	**0.75**	0.74
85_15	**0.74**	0.5	**0.81**	**0.80**	0.63	**0.85**

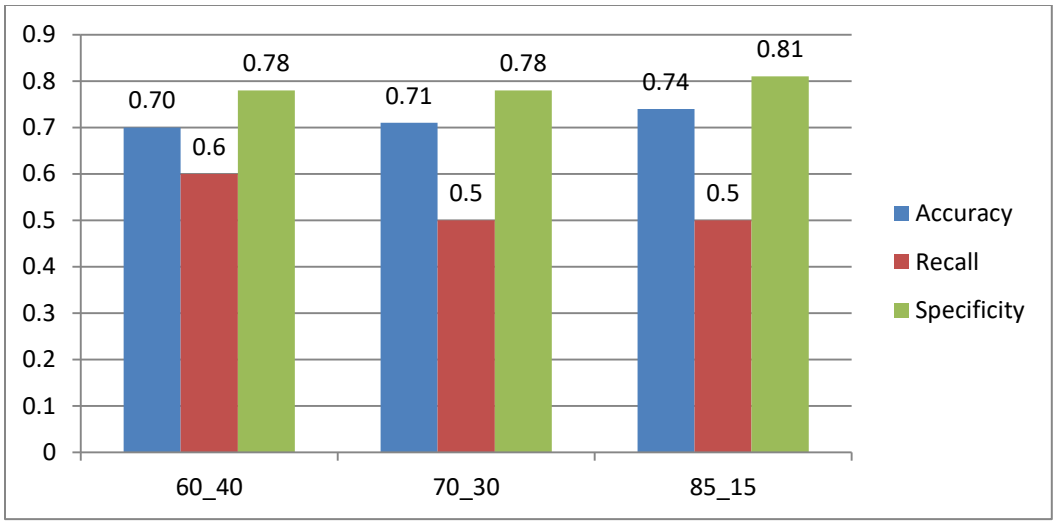


Figure (4-3): SVM Classifier result based on seven Sub-Features.

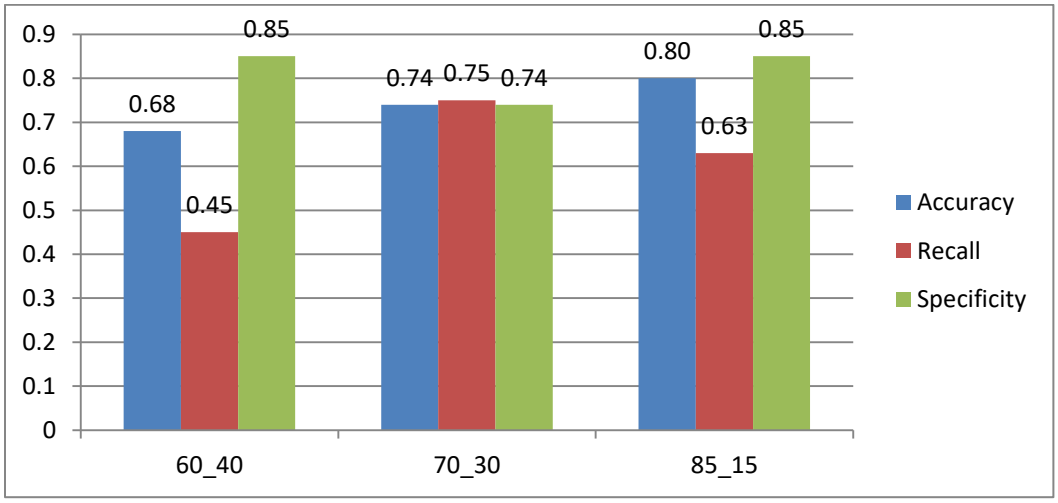


Figure (4-4): KNN Classifier result based on seven Sub-Features

4.2.3 Result of apply KNN and SVM using five Sub-Features

When applying the KNN and SVM classification algorithm after selecting the important features (five Sub-Features), the classification accuracy result is better than the result of all features and the result of selected seven Sub-Features. Tables (4.3) shows classification Results based on SVM and KNN classifier after five training and using selected five Sub-Features, Figures (4.5_4.6) diagram shows an accuracy, Recall and Specificity result for SVM and KNN classifier based five Sub-Features.

Table 4-3): KNN, SVM classifications results after five training based on five Sub-Features

Split dataset	**SVM**			**KNN**		
	Accuracy	**Recall**	**Specificity**	**Accuracy**	**Recall**	**Specificity**
60_40	0.71	**0.75**	0.70	0.72	0.67	0.75
70_30	0.77	0.47	**1**	0.80	**0.75**	0.81
85_15	**0.78**	**0.75**	0.8	**0.83**	0.63	**0.89**

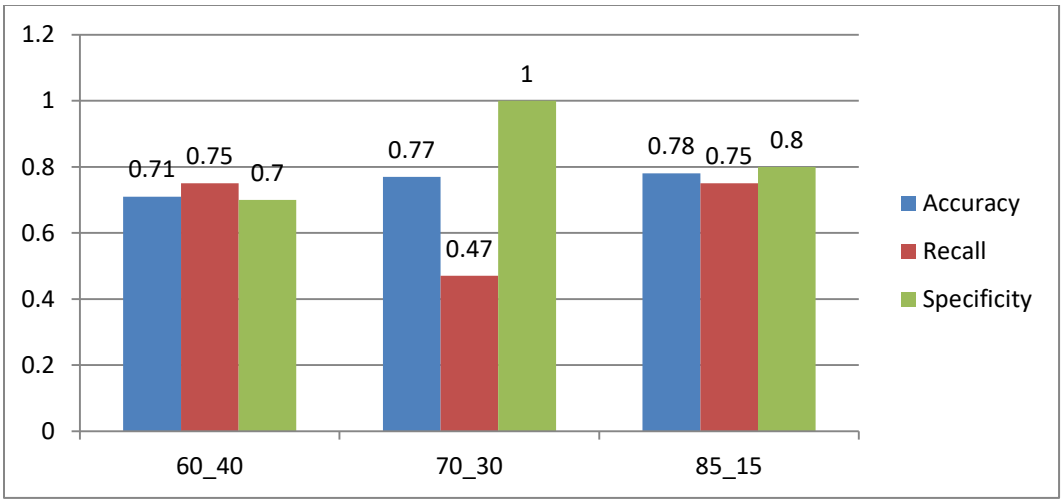


Figure (4-5): SVM Classifier result based on five Sub-Features

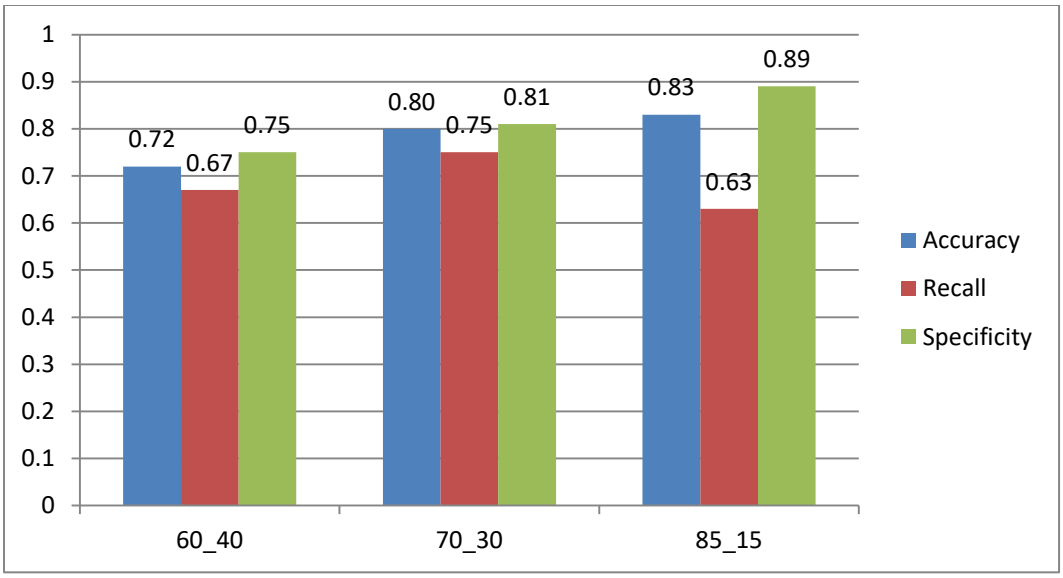


Figure (4-6): KNN Classifier result based on five Sub-Features

4.2.4 Result of apply KNN and SVM using top three Sub-Features

After applying the classification algorithm with the three sub-features, the result of classification accuracy is less than the result of classification accuracy for selecting five sub-feratures, and the reason for the low accuracy of classification is that there are important features that not selected in the process of selecting features selection , which led to a decrease in classification accuracy. The five-sub features are the best features that can be ranked because they provide us with the best classification accuracy in this study. Tables (4.4) show the classification results based on the SVM and KNN classifier after five exercises and with three specific sub-features, the graph (4.7_4.8) shows the result of accuracy, recall, and specificity of the SVM classifier and KNN based three Sub-Features.

Table 4-4): KNN, SVM classifications results after five training based on three Sub-Features

Split dataset	**SVM**			**KNN**		
	Accuracy	**Recall**	**Specificity**	**Accuracy**	**Recall**	**Specificity**
60_40	**0.55**	0.5	0.6	**0.51**	0.6	0.5
70_30	**0.57**	0.4	0.7	**0.66**	0.5	0.8

85_15	**0.61**	0.5	0.7	**0.72**	0.6	0.8

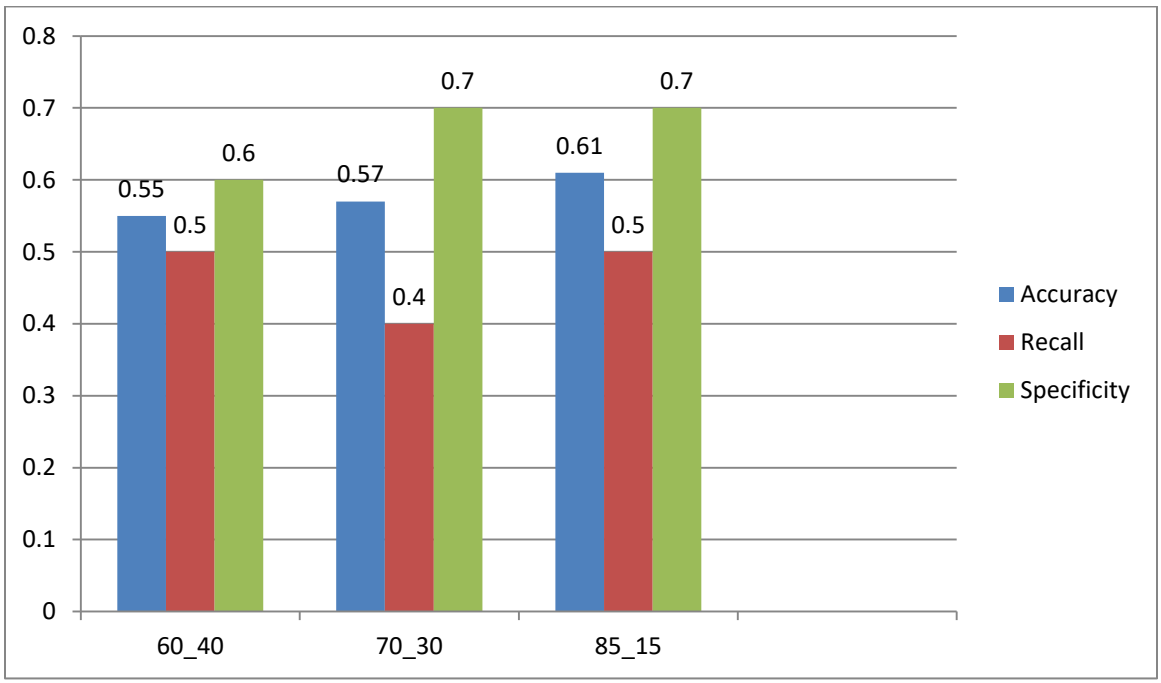


Figure (4-6): SVM Classifier result based on three Sub-Features

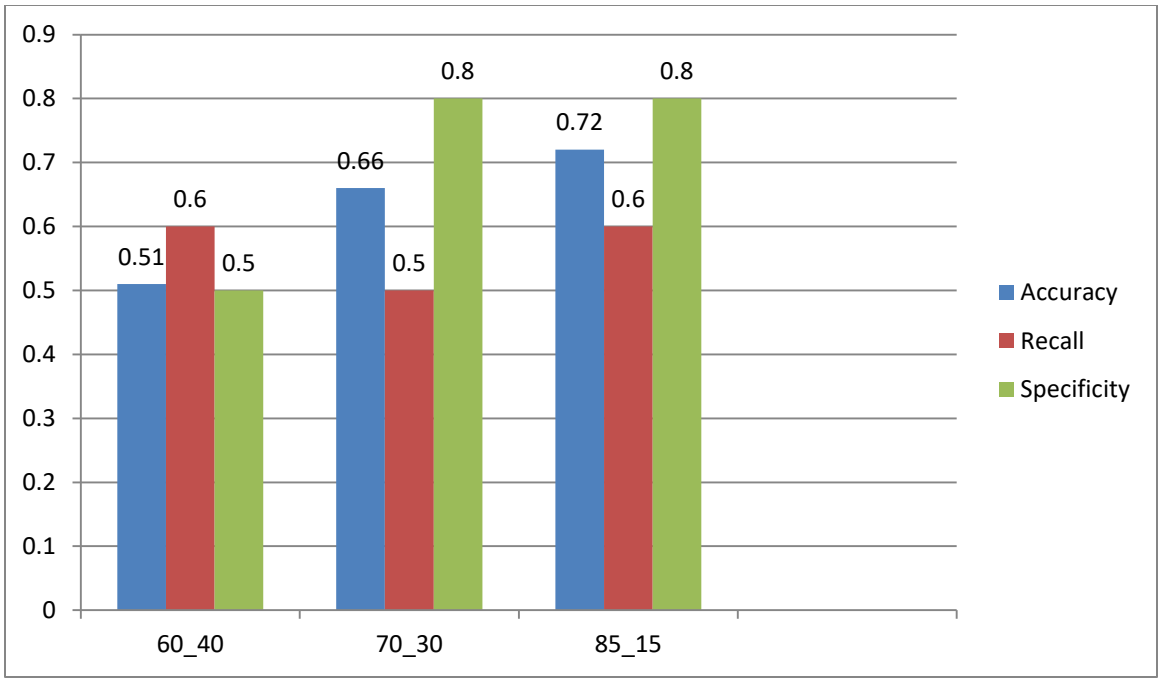


Figure (4-7): KNN Classifier result based on three Sub-Features

4.3 Results discussions

When applying the SVM and KNN classification algorithm with all the features; obtained The best classification SVM algorithm results accuracy (0.61)) and the best classification KNN algorithm results accuracy (0.67).

After selecting the important features by Mutual information(MI) algorithm then applying the KNN and SVM classification algorithm (seven Sub-Features) increased the classification results than more results of all features ; The best classification SVM algorithm results accuracy (0.74) when split dataset (85_15).The best classification KNN algorithm results (accuracy (0.80) when split dataset (85_15).

When applying the KNN and SVM classification algorithm with (five Sub-Features); The best classification SVM algorithm results accuracy (0.78) when split dataset (85_15).The best classification KNN algorithm results accuracy (0.83) .

The reason for the high results of classification accuracy after features selection by Mutual information (MI) algorithm is the selection of the important features that increase the accuracy of

classification and the exclusion of the unimportant features that reduce the accuracy of classification.

After applying the SVM and KNN classification algorithm with three sub-features, the classification accuracy result is lower for SVM and KNN algorithm respectively (0.61_0.72)when split dataset (85_15) . Selecting fewer than five features leads to lower classification accuracy, The five features are the most appropriate features to can be apply classification algorithms on her and give good classification accuracy.

The experiments approve that the classification accuracy results is improved when selected the most important features using Mutual information(MI) algorithm(five Sub-Features) compared with all features. Table (4.4) shows the result for each stage, Figure (4.7_ 4.8): diagram show SVM, KNN experiments result.

By comparing the results of this study with previous studies (Study No. 27), the accuracy of this study is higher than the accuracy of previous studies using the same data set, and the reason for the high accuracy of the classification of this study is the use of the feature selection algorithm mutual information (MI).

Table (4-5): Accuracy results according to the number of features

Split dataset	**SVM**				**KNN**			
	all features	**seven Sub-Features**	**five sub features**	**three sub features**	**all features**	**seven Sub-Features**	**five sub features**	**three sub features**
60_40	0.57	0.70	0.71	055	0.59	0.68	0.72	0.51
70_30	0.57	0.71	0.77	057	0.58	0.74	0.80	0.66
85_15	**0.61**	**0.74**	**0.78**	**061**	**0.67**	**0.80**	**0.83**	**0.72**

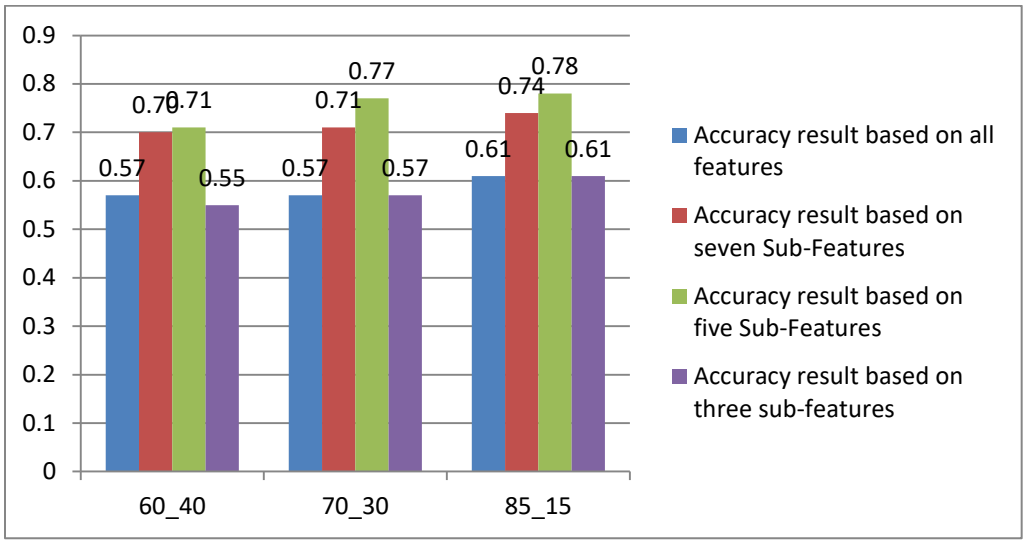


Figure (4-8): SVM experiments result

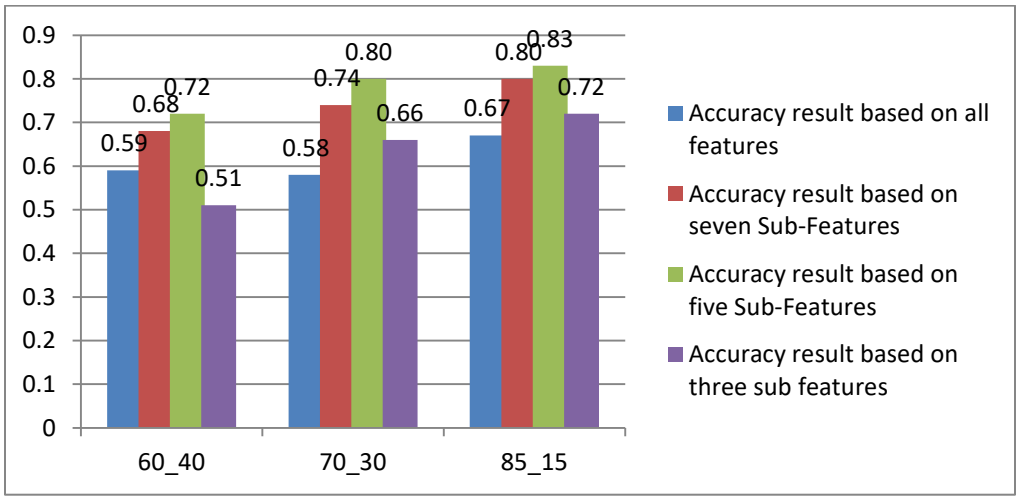


Figure (4-9): diagram show KNN experiments result

4.4 Summary

In this chapter the proposed methodology for the classification of the mammogram images used K-Nearest Neighbor and Support Vector Machine classification method and used the feature selection method (MI) to select and rank the most important features (seven Sub-Features and

five Sub-Features).The best accuracy obtained when select the most important features (five Sub- Features) the accuracy is (83%) and (78%) by K-NN and SVM respectively .

Chapter Five
Conclusion and Recommendation

5.1 Conclusion

Breast cancer threatens the lives of many women, so it has to be paid attention and research on how to predict it, but over there is a problem which is how to obtain a high rating accuracy. This study is an attempt at how to increase the accuracy of classification for mammogram image (breast cancer) using the feature selection method (MI) of selecting to important feature that contributes to increasing accuracy.

This study aimed to increase the accuracy of classifying the mammogram images is five phases starting in downloaded from the MIAS Mini data set (Mammographic Image Analysis Society) The MIAS ,second phases is preprocessing by extracting the Region of Interest (ROI) using the function from [x] position to [y] position and [radius] depend of the MIAS dataset ,third phase is features extraction , fourth phase is features selection by mutual information (IM) to important features selection ,and fifths phase is evolution measurement by confusion matrix.

Used the study the K-Nearest Neighbor and Support Vector Machine classification algorithm, the KNN and SVM classification accuracy enhanced after selecting the most important features and the best accuracy obtained when select the most important features (five Sub-Features) the accuracy is (83%) and (78%) by K-NN and SVM respectively. Classification accuracy decreased when selecting the three most important features being (72%) and (61%) by K-NN and SVM respectively. This indicates that there are important features that have been dispensed with, which increase the classification accuracy and accordingly the best classification accuracy is obtained (83% for KNN algorithm and 78% for SVM algorithm) when classifying the five most important features. The classification accuracy results proved in this study, the accuracy of the KNN algorithm is higher than that of the SVM.

5.2 Recommendation

This study contributed to improving the results of the accuracy of classification of mammograms using feature selection (Mutual Information Algorithm (MI)). I recommend :

- increasing and improving accuracy by comparing more than one classifier with other features selection method instead of MI,
- Compare the classification results with this study results.

References

Siddhartha Sankar Nath et al. (2014). A survey of image classification methods and techniques. International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 554-557.

Abeer Alzubaidi ,Georgina Cosma,David Brown,A. Graham Pockley. (2016). Breast cancer diagnosis using a hybrid genetic algorithm for feature selection based on mutual information. International Conference on Interactive Technologies and Games, (pp. 70-76.).

Animesh Hazra, Subrata Kumar Mandal, Amit Gupta. (2016). Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms. International Journal of Computer Applications, 39-45.

Animesh Hazra,Surbrata Kumar Mandal,Gubta. (2016). Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms . International Journal of Computer Applications.

B.Padmapriya , T.Velmurugan. (2016). Classification Algorithm Based Analysis of Breast Cancer Data. International Journal of Data Mining Techniques and Applications, 43-49.

Bazila Banu , and Ponniah Thirumalaikolundusubramanian. (2018). Comparison of Bayes Classifiers for Breast Cancer Classification. Asian Pacific journal of cancer prevention: APJCP, 2917-2920.

Chhabra, G. Kaur and A. (2014). "Improved J48 classification algorithm for the prediction of diabetes". International Journal of Computer Applications , 13-17.

Dr. R. J. Ramteke, Khachane Monali. (2012). Automatic Medical Image Classification and Abnormality Detection Using KNearest Neighbour. International Journal of Advanced Computer Research, 190-196.

Elkan, C. (2012). Evaluating classifiers. 250. California, retrieved.

El-Sayed Ahmed El-Dahshana, Tamer Hosnyb, Abdel-Badeeh M. Salem. (2010).

Grant, M., S. Boyd, and Y. Ye, CVX. (2008). Matlab software for disciplined convex programming.

Jiawe Hana, Jian Pei , Micheline KAMBER. (2011). Data mining: concepts and techniques. morgan kaufmann publishers is an imprint of elsevier.

Jorge R. VergaraPablo A. Estévez. (2013). "A review of feature selection methods based on mutual information".

KUMARI, Praveen, et. (2016). Web Mining-Concept, Classification and Major Research Issues: A Review. Asian J. Adv. Basic Sci. ISSN JORNAL, 41-44.

M Manoj krishna, M Neelima, M Harshali and M Venu Gopala Rao . (2018). Image classification using Deep learning. International Journal of Engineering & Technology.

M Manoj krishna, M Neelima, M Harshali and M Venu Gopala Rao. (2018). Image classification using Deep learning. International Journal of Engineering & Technology, 614-617.

Mohsen, H., El-Dahshan, E. S. A., El-Horbaty, E. S. M., & Salem. (2018). Classification using deep learning neural networks for brain tumors. Future Computing and Informatics Journa, 68-71.

Obuandike Georgina N, Audu Isah, John Alhasan. (2015). Analytical Study of Some Selected Classification Algorithms in WEKA Using Real Crime Data", IJARAI. International Journal of Advanced Research in Artificial Intelligence, Vol. 4, No.12,.

R. Kumar, B. Singh, D. Shahani, A. Chandra, and K. Al-Haddad,. (2015). "Recognition of power-quality disturbances using S-transform-based ANN classifier and rule-based decision tree,. " IEEE Transactions on Industry Applications,, vol. 51, pp. 1249-1258,, pp. 1249 - 1258.

S. Sharma, J. Agrawal, and S. Sharma. (2013.). "Classification through machine learning technique: C4. 5 algorithm based on various entropies,". vol. 82,.

S. Sharma, J. Agrawal, and S. Sharma,. (2013). "Classification through machine Learning technique: C4. 5 algorithm based on various entropies,". International Journal of Computer Applications (0975 – 8887) , vol. 82,, 20-26.

S. Sharma, J. Agrawal, S. Agarwal, and S. Sharma,. (2013). "Machine learning techniques for data mining: A survey,". Computational Intelligence and Computing Research (ICCIC),.

Saira Charan, Muhammad Jaleed Khan, Khurram Khurshid. (2018). Breast Cancer Detection in Mammograms using Convolutional Neural Network. IEEE International Conference on Mathematics and Engineering Technologies (iCoMET), (pp. 1-5).

Sandhya G 1, D Vasumathi2, G T Raju3. (2015). Classification of Mammogram Images for Detection of Breast Cancer. IOSR Journal of Computer Engineering (IOSR-JCE), 11-17.

Sertan Kaymaka , Abdulkader Helwana, Dilber Uzun. (2017). Breast cancer image classification using artificial neural networks. Procedia Computer Science , 126–131.

Shofwatul ‘Uyun, Lina Choridah. (2018). Feature Selection Mammogram based on Breast Cancer Mining. International Journal of Electrical and Computer Engineering, 60-69.

SUMBALY, Ronak. (2014). Diagnosis of breast cancer using decision tree data mining technique. International Journal of Computer Applications, 16-24.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, . (2006). Classification: basic concepts, decision trees, and model evaluation. Introduction to data mining. 1: 145-205.

Xiaoming Liu,et al. (2012). Mass Diagnosis in Mammography. International Conference on Intelligent Computing (pp. 1-8). Nanchang, China: Springer, Berlin, Heidelberg.