Sudan University of Science and Technology

College of Graduate Studies

College of Computer Science and Information Technology

Multiclassification for Medical Images Using Voting Method

التصنيف المتعدد للصور الطبية باستخدام طريقة الانتخاب

A Thesis Submitted to the Collage of College of Graduate in Partial Fulfillment of the Requirement for the Degree of Master of Computer Science

By

Nosayba Mustafa Hummaida

Supervisor
Dr. Ali Ahmed AlfakiAbdalla

Apr 2017

الآية

(وَقُل رَّبِّ زِدْنِي عِلْمًا)

سورة طه (114)

Dedication

This thesis is dedicated to:

The sake of Allah, my Creator and my Master,

My great teacher and messenger, Mohammed (May Allah bless and grant him),

who taught us the purpose of life,

The Sudan University of Science and Technology; my second magnificent home;

My great parents, who never stop giving of themselves in countless

Ways,

My dearest husband, who leads me through the valley of darkness

With light of hope and support,

My beloved brothers and sisters,

My friends who encourage and support me,

All the people in my life who touch my heart,

I dedicate this research.

Acknowledgement

In the Name of Allah, the Most Merciful, the Most Compassionate all praise be to Allah, the Lord of the worlds; and prayers and peace be upon Mohamed His servant and messenger.

First and foremost, I must acknowledge my limitless thanks to Allah, the Ever-Magnificent; the Ever-Thankful, for His help and bless. I am totally sure that this work would have never become truth, without His guidance.

I owe a deep debt of gratitude to our university for giving us an opportunity to complete this work.

I am grateful to some people, who worked hard with me from the beginning till the completion of the present research particularly my supervisor Dr.

Ali Alfaki, who has been always generous during all phases of the research.

I would like to take this opportunity to say warm thanks to all my beloved friends, who have been so supportive along the way of doing my thesis.

I also would like to express my wholehearted thanks to my family for their generous support they provided me throughout my entire life and particularly through the process of pursuing the master degree. Because of their unconditional love and prayers, I have the chance to complete this thesis.

Abstract

Breast cancer is the disease most common malignancy affects female population and the number of affected people is the second most common leading cause of cancer deaths among all cancer types in the developing countries. Nowadays, there is no sure way to prevent breast cancer, because its cause is not yet fully known. But there are things you can do that might lower risk such as early detection of breast cancer can play an important role in reducing the associated morbidity and mortality rates. The basic idea of this study is to a proposed classification method based on multi classifier voting method that can aid the physician in a mammogram image classification. The study emphasis of five phases starting in collect images, preprocessing (image cropping of ROI), features extracting, classification and end with testing and evaluating. The experimental results using MIAS Dataset show that the voting method achieves accuracy of 76.47. We recommend to use more than three classifiers to achieve better performance in terms of accuracy.

المستخلص

سرطان الثدي هو أكثر أنواع السرطانات شيوعا عند النساء ويأتي بالمرتبة الثانية من حيث أعداد الوفيات عند النساء. في الوقت الحاضر، لا توجد طرق فعالة للوقاية من هذا السرطان، لأن سبب حدوثه لم يعرف بصورة كاملة حتى الآن. ولكن هنالك بعض الأشياء التي يمكن من خلالها تقليل نسبة الخطر وذلك من خلال، الكشف المبكر عن سرطان الثدي يلعب دورا مهما في تقليل نسبة انتشار المرض ومعدلات الوفيات.

الفكرة الأساسية لهذا البحث هو اقتراح مصنف يمكن أن يساعد الطبيب في عملية تصنيف الصور من خلال دراسة استخدام تنقيب البيانات على اساس العمليات المختلفة على صور الماموجرام وتصنيفهم كصور حميدة، أو خبيثة بالاعتماد على التصويت بين عده مصنفات. تحتوي هذه الدراسة على خمسة مراحل بدأت بمرحلة جمع الصور ومن ثم تهيئتها واستخلاص الخصائص ومن ثم مرحلة التصنيف انتهاء بمرحلة الاختبار والتقييم للنتائج. تظهر النتائج التجريبية التي إستخدمت مجموعة البيانات(MIAS) أن طريقة التصويت يحقق دقة 76.47%. يوصي بإستخدام المزيد من المصنفات لزيادة في الدقة والاداء.

الفهرس

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviations	Full Form
KDD	Knowledge Discovery in Databases
SVM	Support Vector Machine
BNC	Bayes Naïve Classifier
KNN	K-nearest Neighbors
ROI	Extracting the Region of Interest
MAIS	The Mammographic Image Analysis Society

Chapter One

Introduction

1.1 Introduction

Data mining refers to extracting or “mining” knowledge from large amounts of data. The term is actually a misnomer. Remember that the mining of gold from rocks or sand is referred to as gold mining rather than rock or sand mining. Thus, data mining should have been more appropriately named “knowledge mining from data,” which is un fortunately somewhat long. “Knowledge mining,” a shorter term may not reflect the emphasis on mining from large amounts of data.

Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge.

Data mining tasks can be classified into two categories: descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions. [1]

Knowledge Discovery in Databases, frequently abbreviated as KDD, typically encompasses more than data mining. The knowledge discovery process comprises six phases: data selection, data cleansing, enrichment, data transformation or encoding, data mining, and the reporting and display of the discovered information. Data mining is typically carried out with some end goals or applications. Broadly speaking, these goals fall into the following classes: prediction, identification, classification, and optimization. [2]

1.2 General Approach of Classification

Classification is the process of learning a model that describes different classes of data. The classes are predetermined. For example, in a banking application, customers who apply for a credit card may be classified as a poor risk, fair risk, or good risk. Hence this type of activity is also called supervised learning. Once the model is built, it can be used to classify new data. The first step—learning the model—is accomplished by using a training set of data that has already been classified. Each record in the training data contains an attribute, called the class label, which indicates which class the record belongs to. The model that is produced is usually in the form of a decision tree or a set of rules. Some of the important issues with regard to the model and the algorithm that produces the model include the model's ability to predict the correct class of new data, the computational cost associated with the algorithm, and the scalability of the algorithm. [2]

Classification is a fundamental issue in machine learning and data mining. In classification, the goal of a learning algorithm is to construct a classifier image of breast cancer. Classification has two-step Process Classifier building; Describing a set of predetermined classes and Classifier usage. [3]

1.3 Research Background

Medical Images Classification is a form of data analysis that extracts models describing important data classes. Mammography is a good method used for early cancer detection. Breast Cancer is a disease that threat women in the whole world especially in the developed country.
The process of manually interpreting the mammography is a very difficult task for the specialists it take time, effort and it depend on the experience of the specialist. The general approach for classification stage considered as a two-steps process. In the first step, we build a classification model based on training data. In the second step, we determine if the model's accuracy is acceptable, and if so, we use the model to classify new data.
Classification is the process of taking decisions that best matches the membership of the object. The task is a complex process that is influenced by many factors. The goal is to associate the appropriate class labels with the test image. Classification is a data mining (machine learning) technique used to predict group membership for data instances. [4]
Mammogram is an X-ray image used to screen for breast cancer. Mammograms play a key role in early breast cancer detection and help decrease breast cancer deaths. Mammograms are medical images that are difficult to interpret, thus a pre-processing phase is needed in order to improve the image quality and improve the performance in breast cancer screening in computer aided diagnosis (CAD), the CAD process can mainly help in decrease the number of errors in diagnosis phase. In CAD process, the classification algorithms or software can help in searching for suspicious signs, or could help in classifying lesions in benign or malignant types. CAD system consists of several modules, such as preprocessing, segmentation and

classification of pathological cases. The medical image classification procedure usually consists of three steps: Texture Feature Extraction, noise removal and Classification. [5]

1.4 Problem Statement

The main reason of the low accuracy comes as a result of using either not accurate features or not a proper single classifier method. Most of the classifiers use different set of medical image features. The accuracy of classification result depends on two things the accurate features and the classifier method used. Accurate features play an important role in classification accuracy. This study use six statistical features function extracted from each mammogram images and propose multi-classifier approach based on SVM, Bayes Naïve and K-nearest Neighbors classifiers and voting method.

1.5 Objectives of the study

i. To get Region of Interest from each mammogram images.
ii. To extract Mean, Standard Deviation, Skewness, Kurtosis, Contrast and Smoothness features for each image.
iii. To apply SVM, Bayes Naïve and K-nearest Neighbors classifiers for mammogram image classification.
iv. To develop mammogram image multi-classifier using the results obtained by the three classifiers mentioned above and based on voting method.

1.6 Significant of the study

The high incidence of breast cancer in women, especially in developed countries, has increased significantly in the last years. Though much less common, the etiologies of this disease are not clear and neither are the reasons for the increased number of cases. Currently there are no methods to prevent breast cancer, which is why early detection represents as a very important factor in cancer treatment and allows reaching a high survival rate. Mammography is considered the most reliable method in early detection of breast cancer. Due to the high volume of mammograms to be read by physicians, the accuracy rate tends to decrease, and automatic reading of digital mammograms becomes highly desirable. It has been proven that double reading of mammograms (consecutive reading by two physicians or radiologists) increased the accuracy, but at high costs. That is why the computer aided diagnosis systems are necessary to assist the medical staff to achieve high efficiency and effectiveness.

In Sudan, cancer was the third leading cause of death after malaria and viral pneumonia , according for 5% of all deaths. There is an urgent need for better early detection of cancer in Sudan to make treatment more effective.[20]

1.7 Expected Results

This study aims to increase the accuracy result of the mammogram images by apply the multi classifier voting, by the end of this study the expected result is automate the classification of mammography images with acceptable accuracy to detect either the image is in beginning or is malignant.

1.8 Scope of Study

This study covers the Offline not online 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.

1.9 Thesis Organization

Chapter one 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 Literature reviews of 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 explains the implementation of the Voting based classification method, build the individual classifiers and run it in training data, test it after that in test data then applying voting method to determine the accuracy of combine the classifiers and compare the result with other studies. In the last chapter the conclusion and recommendation of the study are mentioned.

Chapter Two

Literature Review

2.1 Introduction

Classification is the process of learning a model that describes different classes of data. The classes are predetermined. For example, in a banking application, customers who apply for a credit card may be classified as a poor risk, fair risk, or good risk. Hence this type of activity is also called supervised learning. Once the model is built, it can be used to classify new data. The first step—learning the model—is accomplished by using a training set of data that has already been classified. Each record in the training data contains an attribute, called the class label, which indicates which class the record belongs to. The model that is produced is usually in the form of a decision tree or a set of rules. Some of the important issues with regard to the model and the algorithm that produces the model include the model's ability to predict the correct class of new data, the computational cost associated with the algorithm, and the scalability of the algorithm. [2]

2.2 Classification Methods or Approaches

The methods or approaches of medical image classification has span into different techniques of data mining. This section provides a brief overview of different classification methods to classify or to detect abnormalities in medical images.

2.2.1 Neural Network

Artificial neural networks are relatively crude electronic networks of neurons based on the neural structure of the brain. They process records one at a time, and learn by comparing their classification of the record (i.e., largely arbitrary) with the known actual classification of the record. The errors from the initial classification of the first record is fed back into the network, and used to modify the networks algorithm for further iterations.

2.2.2 Decision Tree

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy).

Leaf node (e.g., Play) represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

2.2.3 Bayes Classification Methods

The Naive Bayesian (NB) is based on the Bayesian theorem .The Naïve Bayesian Classifier assumes that features are independent .This method is important for several reasons. It is very easy to construct, does not need any complicated iterative parameter estimation schemes. This means it may be readily applied to huge data sets. This classification technique analyses the relationship between each attribute and the class for each instance to derive a conditional probability for the relationships between the attribute values and the class [6].

2.2.4 K-Nearest Neighbor Classification

Neighbors-based classification is a type of instance-based learning or non-generalizing learning: it does not attempt to construct a general internal model, but simply stores instances of the training data. Classification is computed from a simple majority vote of the nearest neighbors of each point: a query point is assigned the data class which has the most representatives within the nearest neighbors of the point.

2.2.5 Support Vector Machine Classification

It is a machine learning method which uses a hyperplane that maximizes the margin in the training data to classify binary classes. Support vectors are the training data along the hyperplane. The distance between the support vectors and the class boundary is the margin. The decision planes that define decision boundaries are the basic idea of SVM. Therefore, the optimal separating hyperplane maximizes the margin of the training data.

2.3 Classification of Medical Image

Medical image Classification can play an important role in diagnostic and teaching purposes in medicine. For these purposes different imaging modalities are used. There are any classifications created for medical images using both grey-scale and color medical images. One way is to find the texture of the images and have the analysis.

Texture classification is an image processing technique by which different regions of an image are identified based on texture properties [8].

2.3.1 Texture Classification

In texture classification the goal is to assign an unknown sample image to one of a set of known texture classes. Texture classification is one of the main domains in the field of texture analysis. Texture analysis is important in many applications of computer image analysis for classification or segmentation of images based on local spatial variations of intensity or color. A successful classification or segmentation requires an efficient description of the image texture. An important application area in texture classifications are industrial and biomedical surface inspection, for example finding the defects and disease, ground classification and segmentation of satellite or aerial imagery, segmentation of textured regions in document analysis, and content-based access to image databases etc. However, despite of many potential areas of application for texture analysis in industry, there are only a limited number of successful examples available. A wide variety of techniques for describing image texture have been proposed. Texture analysis methods can be divided into four categories called statistical, geometrical, model-based and signal processing [9].

2.3.2 Important of Medical Image Features

Medical images in their raw form are represented by arrays of numbers in the computer, with the numbers indicating the values of relevant physical quantities that show contrast between different types of body tissue. Processing and analysis of medical images are useful in transforming raw images into a quantifiable symbolic form for ease of searching and mining, in extracting meaningful quantitative information to aid diagnosis, and in integrating complementary data from multiple imaging modalities. [9]

2.3.3 Medical Image Feature Extraction

In various computer visions, feature extraction applications widely used is the process of retrieving desired images from a large collection on the basis of features that can be automatically extracted from the images themselves. The feature extraction process is started by edge and shape information extraction from original medical x- ray images.

These systems called CBIR (Content- Based Image Retrieval) have received intensive attention in the literature of image information retrieval. The algorithms used in these systems are commonly divided into three; Extraction, Selection and Classification. The extraction task transforms rich content of images into various content features. Feature extraction is the process of generating features to be used in the selection and classification tasks. Feature selection reduces the number of features provided to the classification task. There are adaptive normalization and Feature compensation method. Adaptive normalization is applied after the classification stage, and yields a normalization size appropriate for the input pattern with fluctuation in aspect ratio by repeating the processes of normalization and classification while feature compensation is applied to the candidates output at the classification stage, and offsets the feature values corrupted by image degradation to obtain higher recognition accuracy in the final recognition stage. [9]

Extracting the Region of Interest (ROI) using the function from [x] position to [y] position and [radius] depend of the MIAS dataset. This stage means to apply the six functions to extract the features from each image. The statistical image features such mean, standard deviation, smoothness, contract, kurtosis and skewnes are extracted from the ROI of the mammogram images.

2.4 Multi Classifier

The concept of Multi Classifier System (MCS) based on set of individual classifiers combining their decisions to classify new patterns, using MCS might not be better than the single classifier but can reduce the risk of picking an inadequate single classifier.

In the fact any classifier has its own strengths and weakness, so a group of classifiers could make strong decision and the final output would become more reliable. The most reason for combining Multi Classifier to improve the accuracy, there are two main strategies for combine classifiers Selection and fusion. A very simple and widely-used method of the second paradigm is voting. According to this procedure, each model outputs a class value and the class with the most votes is the one proposed by the ensemble.

There are two main paradigms in combining different classification algorithms: classifier selection and classifier fusion. The first one selects a single algorithm for classifying a new instance, while the latter combines the decisions of all algorithms. A very simple but effective method of the first paradigm is Evaluation and Selection. This method evaluates each of the models (typically using 10-fold cross-validation) on the training set and selects the best one for application to the test set. A very simple and widely-used method of the second paradigm is voting. According to this procedure, each model outputs a class value (or ranking, or probability distribution) and the class with the most votes (or the highest average ranking, or average probability) is the one proposed by the ensemble. This method is expected to have better accuracy than that of the best individual model, due to the correction of uncorrelated errors through voting. However, the fact that all models (even the less accurate ones) participate with equal vote, has eventually an overall negative effect on the performance of the method.

In Weighted Voting, the classification models are not treated equally. Each model is associated with a coefficient (weight), usually proportional to its classification accuracy. This partly amends the problem of inferior models, but it does not eliminate it completely as it still allows them to affect the final decision. [10]

Voting algorithm can divided into two types: those that adaptively change the distribution of the training set based on the performance of previous classifiers (as in Boosting method) and those that do not(as in Bagging) .[21]

2.5 Related Work

Several works have been done in the area of mammography image classification, among them; we describe the most recent researches.

In [6] they propose automatic process of mammography classification. They use several machine learning algorithms such as Random forest (RF), The Naive Bayes (NB), C4.5, The multi-layer perceptron (MLP) and Decision Table (DT).The best result were obtained when using Second-Order Statistics together with Random forest as a classification technique.

Also, in [11] attempts to study about pre-processing is the most important step in the mammogram analysis due to poor captured mammogram image quality. Filters are used to improve image quality, remove the noise, preserves the edges within an image, enhance and smoothen the image. The experimental results conclude that the adaptive median filter is best for mammogram image noise removal and gives better performance by estimating the PSNR values.

Another study [12] leads to analysis an efficient method by diagnosing the mammogram using Naïve Bayes classifier. The proposed method has a) ROI extraction (Chain code) b) Pre-processing (Enhancement), c) Feature extraction (HOG) and d) Classification using Naive Bayes classifier. The test of the proposed system yield 96.5% micro-calcification detection in mammograms. Experimental results show that the proposed method using Mammogram Image Analysis Society (MIAS) Database clinical mammogram.

Also, [8] survey focus on highlighting different techniques on enhancement, detection and classification of breast cancer along with its accuracy. Concludes that there are several techniques that deal with pre-processing, enhancement, segmentation, feature extraction and classification of diagnosing images that gave different accuracies. Also found that the mammographic images are giving better accuracy than ultrasound images, MRI images etc. and most of the works used MIAS database which contain 322 mammographic images.

Another work in [7] they propose a Multiple Classifier System (MCS) for classifying breast lesions in Dynamic Contrast Enhanced-Magnetic Resonance Imaging (DCE-MRI). The proposed MCS combines the results of two classifiers trained with dynamic and morphological features respectively. Twenty-one malignant and seventeen benign breast lesions, histologically proven, were analyzed. Volumes of Interest (VOIs) have been automatically extracted via a segmentation procedure assessed in a previous study. The performance of the MCS have been compared with histological classification. Results indicated that with automatic segmented VOIs 90% of test-set lesions were correctly classified. Also in [13] they proposed a multi classifier system composed of three classifiers.

That used dynamic features to classify breast lesion in DCE-MRI, Several neural networks classifiers like MLP, PNN, GRNN, and RBF has been presented on a total of 112 histopathologically verified breast lesions to classify into benign and malignant groups. Also, support vector machine has been considered as classifier. Before applying classification methods, feature selection has been utilized to choose the significant features for classification. Finally, to improve the performance of classification, three classifiers that have the best results among all applied methods have been combined together that they been named as multi-classifier system. For each lesion, final detection as malignant or benign has been evaluated, when the same results have been achieved from two classifiers of multi-classifier system. The results show that the proposed methods are correctly capable to feature selection and improve classification of breast cancer.

And in [14] a theoretical model is derived for estimating the misclassification error probability of MCS based on majority vote combiner. In the derivation, they assumed that each classifier produces at its output an estimation of the posterior class probability that has a Gaussian distribution. In addition, they assumed that each classifier has two classes, and the outputs of classifiers are dependent and identically distributed. They validated our model using computer simulations. Results show that the ensemble performance is highly sensitive to class variance while exhibits a smoother behavior against class mean. Also, results show that as the correlation among classifiers' outputs increases, the probability of classification error degrades exponentially. The trend continues until the performance reaches the behavior of a single classifier regardless of the number of base classifiers used in the ensemble. The proposed model provides a better understanding of the behavior of majority vote combiner in MCS.

2.6 General Discussion

The vast majority of the literature evaluates the performance of classification models using only the criterion of predictive accuracy. For instance, contrasting the interpretability of decision trees and classification rules, if the dataset contains many attributes having a single value that is relevant for class prediction, decision trees would have the drawback of including irrelevant values.

Bayesian network classifiers is a graphical models, the correct interpretation of the former seems more difficult from users. Nearest neighbor algorithms usually do not produce an explicit, turning to evaluation issues, there have been few experiments where users subjectively evaluated the comprehensibility of classification models expressed in different types of knowledge representations.

Ensemble methodology, which builds a classification model by integrating multiple classifiers, can be used for improving prediction performance. Ensemble neural networks to generally improve the accuracy and robustness of sample classification. Ensemble classifier shows predominate performance. Furthermore, the characteristics and different effect issues to classification performance. If a single SVM can obtain satisfactory classification performance, an ensemble SVM is hardly capable to improve it. Otherwise, an ensemble of SVM is superior to the best single SVM. We also investigated the effect of kernel functions, feature selections and type of classifiers on the classification.

Chapter Three
Research Methodology

3.1 Introduction

The classification process involves five major steps namely image collection, Image Cropping, feature extraction, Individual Classification and Multi-classifier based on voting method. Image collection (all mammogram images collected from MIAS), Image Cropping applied crop technique to the images; cut the interest parts of the image and remove the unwanted parts of the image, 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 Individual Classification we apply three classifiers (BNC, KNN and SVM), Multi -classifier based on the individual results obtained by each single classifier.

3.2 Research Phases

This study emphasis of five phases starting with collect images, preprocessing, features extracting, individual classification and end with testing and evaluating. The research phases is described in Figure (3.1) followed by detailed about each phase

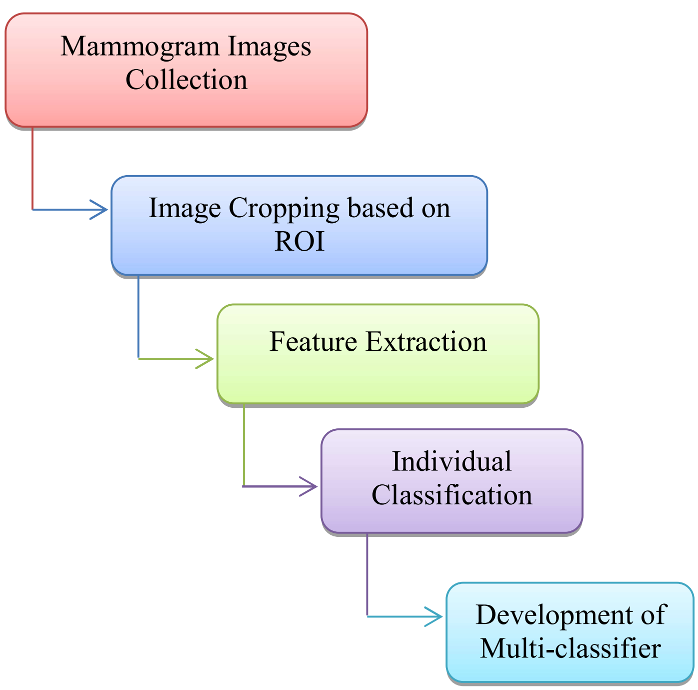


Figure (3.1): Research Phases

3.2.1 Phase (1) Mammogram Images Collection

Dataset used in this study downloaded from the MIAS (Mammographic Image Analysis) database website [15], this dataset was recently used by many researcher [16, 17].

3.2.2 Phase (2) Image Cropping based on ROI

Images processing techniques applied to the images before the feature extraction phase. Regions of interest ROI's are defined as regions containing user defined objects of interest. Here we applied crop technique to the images; a cropping operation was employed in order to cut the interest parts of the image. Cropping removed the unwanted parts of the image usually peripheral to the regions of interest.

3.2.3 Phase (3) Features Extraction

This stage means to apply the five moment function to extract the six features from each image, these moments are: mean, standard deviation, skewness, kurtosis, contrast and smoothness. Feature extraction is a process to analyze objects and images to extract the most prominent features that are correspondence of various classes of objects.

3.2.3.1 Mean

The Mean is a measure of the average intensity of the neighboring pixels of an image. (Equation 3.1)

$$Mean = \sum_{i=0}^{L-1} Z_i * P(Z_i)$$

Equation (3.1): Mean [22]

3.2.3.2 Standard Deviation

The Standard Deviation is a measure of how spread out numbers are. (Equation 3.2)

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

Equation (3.2): Standard Deviation [22]

3.2.3.3 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 a Skewness near zero. Negative values for the Skewness indicate data that are skewed left and positive values for the Skewness indicate data that are skewed right (Equation 3.3) .

$$skewness = \sum_{i=0}^{l-1} (z_i - m)^3 * p(z_i)$$

Equation (3.3): Skewness [22]

3.2.3.4 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. (Equation 3.4)

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

Equation (3.4): Kurtosis [22]

3.2.3.5 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.(Equation 3.5)

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

Equation (3.5): Contrast [22]

3.2.3.6 Smoothness

Measures the relative intensity variations in a region

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

Equation (3.6): Smoothness [22]

3.2.4 Phase (4) Individual Classification

The result of the previous three phases converts the data to numeric values. In the stage we apply three individual classifiers, namely Support Vector Machine, Bayes Naïve and K-nearest Neighbors. Each of them classify Data in 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. A short detail about each classifier is discussed below.

3.2.4.1 SVM classifier

Support Vector Machine performs classification by finding the optimal hyperplane that maximizes the margin between the two classes. The vectors (cases) that define the hyperplane are the support vectors. Is critical point close to decision boundary?

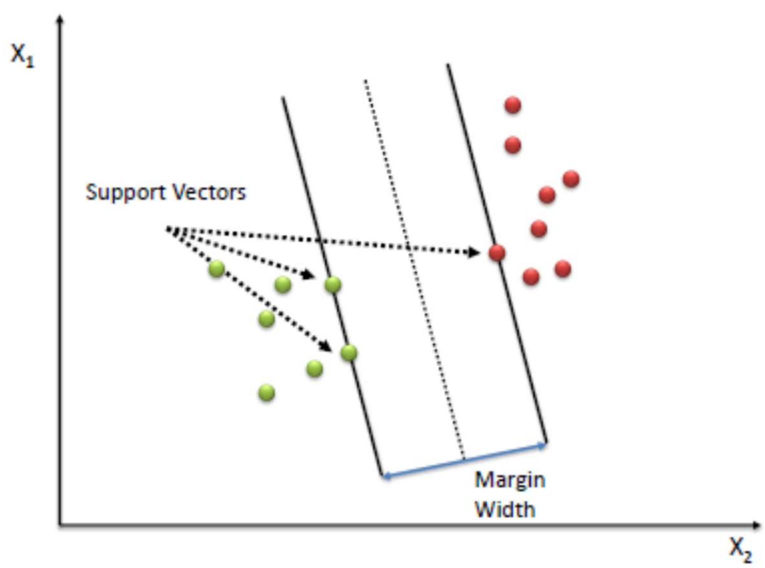


Figure (3.2) select the separating hyperplane that maximizes the margin

The three main ideas of the Algorithm:

1. Define what an optimal hyperplane is (in way that can be identified in a computationally efficient way): maximize margin.
2. Extend the above definition for non-linearly separable problems: have a penalty term for misclassifications.
3. Map data to high dimensional space where it is easier to classify with linear decision surfaces: reformulate problem so that data is mapped implicitly to this space.

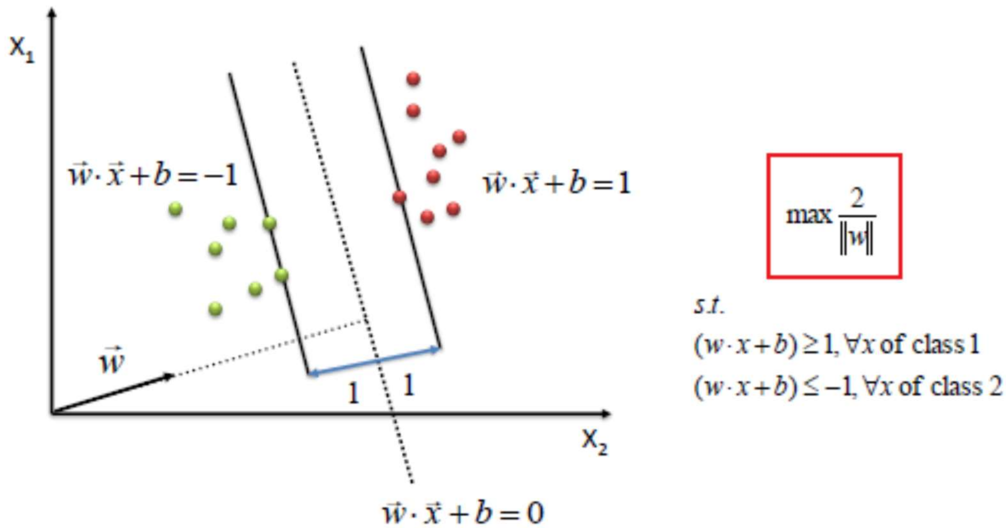


Figure (3.3) the width of the margin is max 2/w

3.2.4.2 Bayes Naïve classifier

In this study Naïve Bayes rule will be applied to convert them into posterior probabilities used to the algorithm discrete valued features and Continuous-valued Features; Numberless values for a feature and Conditional probability often modeled with the normal distribution.

The discrete model uses the following equation:

$$p(c_j|d) = \frac{p(d|c_j) * p(c_j)}{p(d)}$$

Equation (3.8): Discrete model

Where:

$p(c_j|d)$ = Probability of instance d being in class c_j

$p(d|c_j)$ = Probability of generating instance d given class cj,

$p(c_j)$ = Probability of occurrence of class cj,

$p(d)$ = Probability of instance d occurring

In the testing phase we used the previous probability to predict each new instance based on the following equation:

$$p(x/c_i) = p(x_1/c_i) * p(x_2/c_i) * \dots .* p(x_n/c_i)$$

Equation (3.9): Predict model

While the continuous model based on the following equation:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\sigma = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2\right]^{0.5}$$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Equation (3.9): Continues model

Where:

μ Is mean

σ Is standard deviation

σ^2 Is variance

3.2.4.3 K-nearest Neighbors classifier

,

The K-nearest-neighbor (KNN) algorithm measures the distance between a query scenario and a set of scenarios in the data set. We can compute the distance between two scenarios using some distance function $d(\ x\ y)$, where $x\ y$ are scenarios composed of N features, such that $x = \{x_1 \ldots x_N\}\ y = \{y_1 \ldots y_N\}$...

Two distance functions are discussed in this summary:

- Absolute distance measuring:

$$d_A(x, y) = \sum_{i=1}^{N} |x_i - y_i|$$

Equation (3.10)

- Euclidean distance measuring:

$$d_E(x, y) = \sum_{i=1}^{N} \sqrt{x_i^2 - y_i^2}$$

Equation (3.11)

Because the distance between two scenarios is dependent of the intervals, it is recommended that resulting distances be scaled such that the arithmetic mean across the dataset is 0 and the standard deviation 1. This can be accomplished by replacing the scalars $x\ y$ with $x'\ y'$ according to the following function:

$$x' = \frac{x - \bar{x}}{\sigma(x)}$$

Equation (3.12)

Where x is the unscaled value, $\bar{x}$ is the arithmetic mean of feature Equation 3.13), $\sigma(x)$ is its standard deviation (see Equation 3.14), and x' the arithmetic mean is defined as:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Equation (3.13)

x Across the data set (see is the resulting scaled value).

We can then compute the standard deviation as follows:

$$\sigma(x) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Equation (3.14)

KNN can be run in these steps:

1. Store the output values of the M nearest neighbors to query scenario q in vector $r = \{r_1 \dots r^M\}$ by repeating the following loop M times:
   a. Go to the next scenario s^i in the data set, where i is the current iteration within the domain $\{1 \dots P\}$
   b. If q is not set or $q < d(q\ s^i)$: $q \leftarrow d(q\ s^i)\ t \leftarrow o^i$
   c. Loop until we reach the end of the data set (i.e. $i = P$)
   d. Store q into vector c and t into vector r
2. Calculate the arithmetic mean output across r as follows:

$$\bar{r} = \frac{1}{M}\sum_{i=1}^{M} r_i$$

3. Return $\bar{r}$ as the output value for the query scenario q

3.2.5 Phase (5) Development of Multi-classifier based on voting method

In this phase we propose multi-classifier based on the individual results obtained by each single classifier discussed above. The concept of our proposed approach depends on the voting method as we will described in following algorithm:

```
Begin
        Set bnc_output to BNC output xsl file
        Set svm_output to SVM output xsl file
        Set knn_output to KNN output xsl file
        Set s1 to the length of output
        For counter=1 to s1
                Compare the sum of the 3 values
                If sum >=5 Then
                        Set voting_output =2
                Else
                        Set voting_output =1
                End if
        End for
End algorithm
```

3.3 Methods and Tools

3.3.1 Dataset

The dataset of digital mammogram images are collected from the Mammographic image Analysis Society (MIAS) database along with the clinical report. It consists of totally 68benign images and 51 malignant images it is two class dataset. Table (3.2) below gives some description of this dataset

Table (3.2): 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.	

3.3.2 MATLAB 2013b

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. [18]

3.4 Evaluation Measures

The performance of the classifier is estimated by using confusion matrix, running time and classification accuracy. A confusion matrix is a table that is often used to describe the performance of a classification model or "classifier" on a set of test data for which the true values are known.

Table (3.3): General procedure to construct confusion matrix for a classifier.

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

Where terms to the confusion matrix:

i. True negatives (TN): We predicted B, and they have the disease in the Beginning.
ii. True positives (TP): These are cases in which we predicted M (they have the disease in Malignant), and they do have the disease really in Malignant state.
iii. False positives (FP): We predicted M, but they actually have the disease in Beginning.
iv. False negatives (FN): We predicted B, but they actually do have the disease in Malignant.

A number of different measures are commonly used to evaluate the performance of the proposed method. These measures including Accuracy, sensitivity and specificity calculated from confusion matrix using the following equations:

Accuracy = (TP+TN)/ (TP+TN+FP+FN), (Number of correct assessments)/Number of all assessments).

$$CR = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

Equation (3.15): The Accuracy [1]

Sensitivity: is the percentage of positive records classified correctly out of all positive records.

Recall= TP/ (FP+TN), (Number of true positive assessment)/(Number of all positive assessment)

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

Equation (3.16): the Sensitivity (Recall) [1]

Specificity: is the percentage of positive records classified correctly out of all positive records.

TN/(TN + FP) = (Number of true negative assessment)/(Number of all negative assessment)

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

Equation (3.17): The Specificity [1]

Chapter Four

Voting based classification method

4.1 Introduction

This chapter includes experiments that led to the results of applying simple voting method on the three classifiers shown in the table 4.1, also it includes the experiments results and discussion of the result.

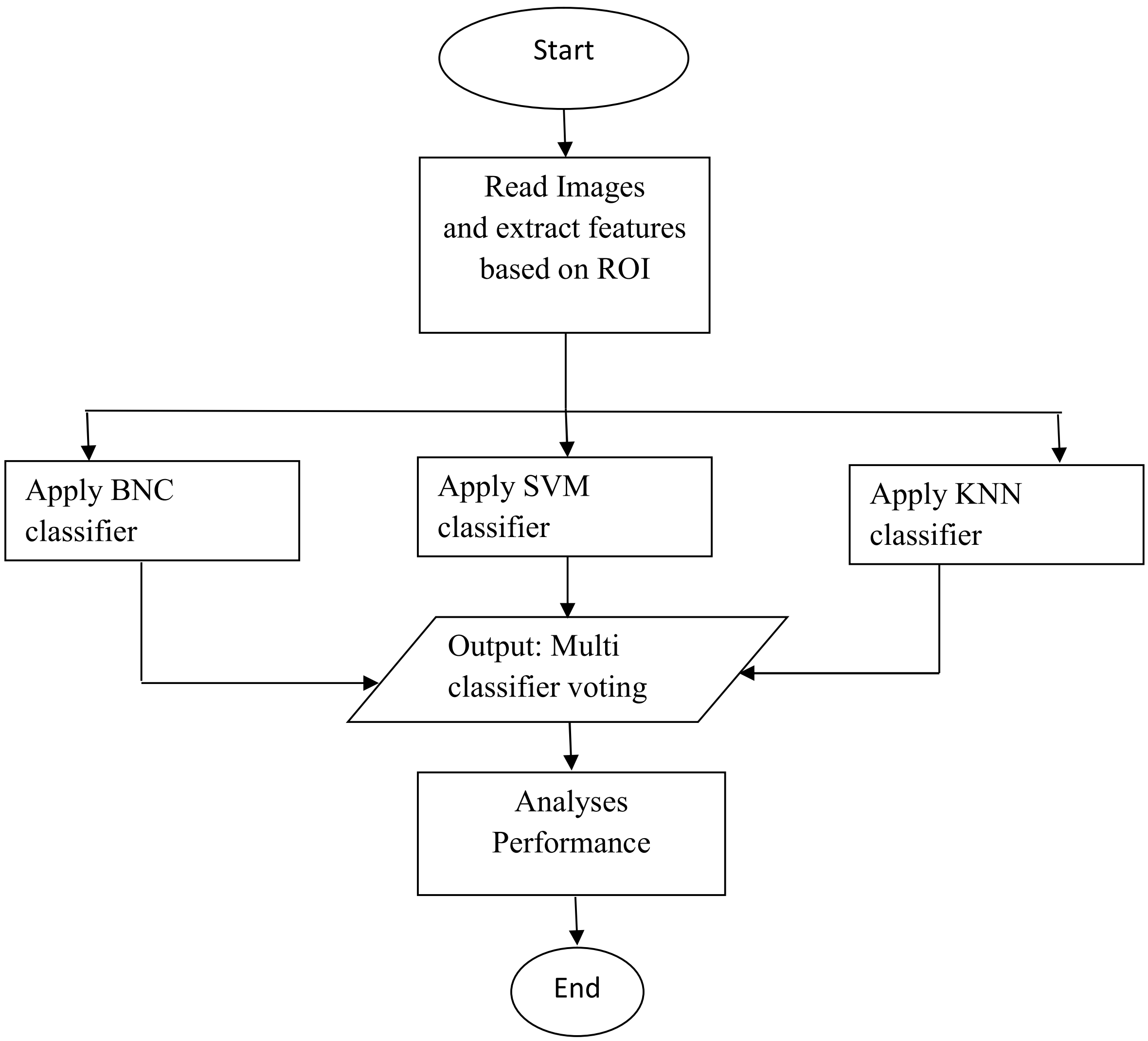


Figure (4.1): Flow Diagram for Medical Images Classification Process using multi voting classifiers

4.2 Data Description

The Mammographic Image Analysis Society MIAS Table (4.1) shows a brief description of the dataset that is being considered.

Table (4.1): Description of Breast Cancer Dataset

Dataset	**No.Of Attributes**	**No. Of Instances**	**No. Of Classes**
MIAS Breast Cancer Dataset	7	119	2

Details of the attributes present in the dataset are shown in Table (4.2)

Table (4.2): MIAS Breast Cancer Dataset Attribute

S No	**Attribute**	**Domain**
1	Mean	0.03 – 196.66
2	Stda	0.20 – 13.25
3	Skew	0.34 – 5.60
4	Smoothness	0.16 – 0.92
5	Kurtosis	0.50 – 3.77
6	Contrast	0.44 – 3.64
7	Class	0 (Benign) or 1(Malignant)

The dataset comprises of 119 instances of breast cancer patients with each, either having malignant or benign type of cancer. Figure (4.2) shows the distribution of the patient based on the class label (malignant or benign).

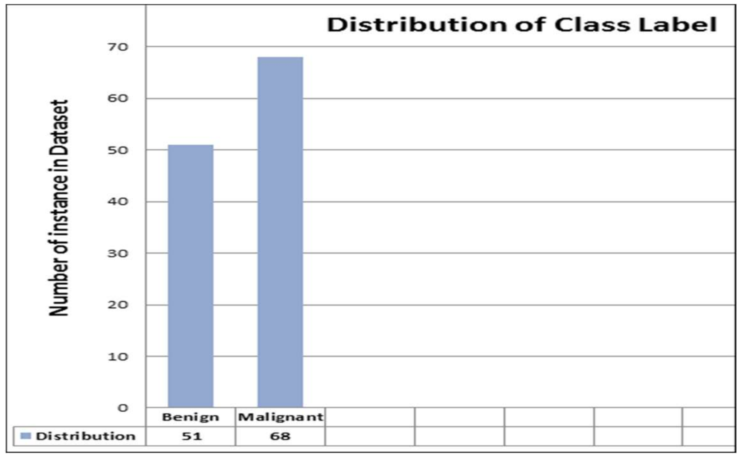


Figure (4.2): Distribution of Class Label in Dataset

4.3 Feature extraction functions

4.3.1 Region of Interest (ROI)

To create (ROI) using the function from [x, y] and [radius] depend of the MIAS dataset using this formula:

```
function [ roi_img ] = my_roi( img,x,y,r )
%this function create the ROI of an image from (x,y) and radius
roi_img=img((x-r):(x+r),(y-r):(y+r));

end
```

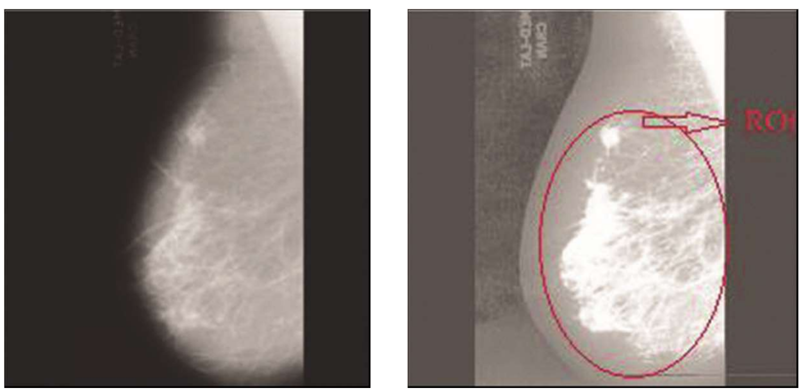
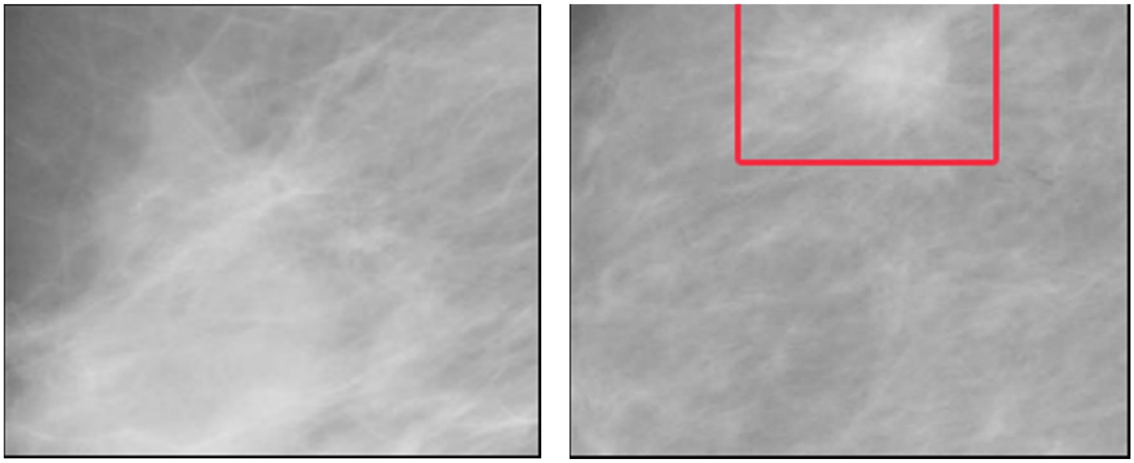


Image Cropping ROI detected

Figure (4.3): Image cropping and detecting Region of Interest (MIAS mdb023) (Red circle indicating ROI)

The size of ROI is 256×256 pixels is extracted by entering the coordinates of X, Y and radius in pixel. Then the images are classified respectively. The extracted portion of ROI of mammography with benign and malignant is shown in Figure(4.4)



(A) (B)

Figure (4.4) : ROI extracted images (a) Benign (b) Malignant

4.3.2 Mean

```
function [ m1 ] = moment1( A )
%this function compute the first moment for 2-D array
% m=size(A,1);
% n=size(A,2);
%m1=(sum(sum(A))/(m*n));
m1=mean(mean(A));
end
```

4.3.3 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
```

4.3.4 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
```

4.3.5 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
```

4.3.6 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
```

4.3.7 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
```

ROI apply the calls all functions

```
for i=1:s1(1)
    ii=num2str(active_img_index(i),'%03d');
    file1=strcat('D:\Mam\dataset\rawimages\mdb',ii,'.pgm');
    raw_img=imread(file1);
    %---------------------------------------------
    roi_of_img=my_roi(raw_img,x,y,r);
    %-------- noise removal code here ------------------------------
%     roi_of_img = medfilt2(roi_of_img1);
    %----------------------------------
    % ----- feature extraction function here --------
    f1=moment1(roi_of_img);
    f2=moment2(roi_of_img);
    f3=moment3(roi_of_img);
    f4=smoothness(roi_of_img);
    f5=kurtosis(roi_of_img);
    f6=contrast(roi_of_img);
    %-----------------------------------------------
    MyMatrix1(j,:)=[f1 f2 f3 f4 f5 f6];
    j=j+1;
end;
```

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

Mean	Stda	Skew	Smoothness	Kurtosis	Contrast
94.4637	11.17506	4.998425	0.917864868	3.351998	3.342912
140.489	11.4605	5.083183	0.919746413	3.428423	3.385336
101.8908	12.81015	5.474816	0.927589493	3.61695	3.579127
145.3989	13.16078	5.574266	0.929382406	3.763542	3.627779
132.2807	11.96595	5.231562	0.922874897	3.629847	3.459183
95.84858	11.10555	4.977677	0.91739327	3.348704	3.332499
95.43837	11.89393	5.210549	0.922444108	3.463183	3.448757
73.54247	10.18048	4.697271	0.910558437	3.1963	3.190687
132.7733	12.52617	5.393601	0.926069222	3.586039	3.539232
84.40241	10.7394	4.867657	0.914816752	3.300165	3.277102
102.3366	11.45809	5.08247	0.919730879	3.405534	3.38498
102.0926	11.52577	5.102464	0.920164577	3.55564	3.394962
114.5456	7.782105	3.927035	0.886132085	3.078205	2.789642
181.8119	12.77561	5.464971	0.927407957	3.778676	3.574299
178.6544	9.105499	4.360495	0.901043979	3.298117	3.017532
154.1629	8.820602	4.269059	0.898173244	3.251205	2.96995
93.156	11.01972	4.951998	0.916803405	3.347264	3.319597
97.57797	11.26653	5.025664	0.918477344	3.464465	3.356565
125.2002	12.82609	5.479356	0.927672977	3.599013	3.581353

4.4 Experiments

In this study we apply a simple voting method with three classifiers (BNC, KNN,SVM) , use continues form of data set , it was divided into 60% ,70% ,85% training and 40%,30%,15% testing datasets.

The dataset used for this experiment is composed of 119 mammograms from the MIAS dataset 71 images used for training and 48 images for testing in (60-40) case.

4.5 Result and Discussion

After apply three different sizes of training and testing and calculate the overall accuracy, the final results are shown in Table below

B	BNC	KNN	SVM	Voting
60-40	0.5532	0.5319	0.5106	0.5957
70-30	0.6000	0.5429	0.5714	0.6571
85-15	0.5882	0.6471	0.6471	0.7647

Table (4.3) Experiments result

Discussion

In this study MIAS data set was used three individual classifiers and applied multi classifier voting based on continues data set. The highest precision was given good accuracy in 85% the accuracy was 0.7647, while in 70% the accuracy was 0.6571 and in 60% the accuracy was 0.5957. Generally the accuracy was increased after applying voting in the three precisions.

4.6 Comparison with the Previous Studies

The main measurements of comparison is Accuracy ,previous study, They proposed a method to classify movie document into positive or negative opinions , consisted of three classifiers based on SVMs, ME and Score calculation. Using two voting method (naïve and weighted and integration with SVMs.

The dataset consists of 700 positive reviews and 700 negative reviews , they divided the data into three equal-size folds and tested their method with three folds cross validation. They applied the boosting algorithm to the integration process , in the method SVMs and the weighted voting method.

	[19] Kimitaka Tsustumi, Kazutaka Shimada and Tsutomu Endo		Our voting Method
Accuracy	Naïve voting	85.8%	76.47%
	Weighted voting	86.4%	
	SVM	87.1%	

Table (4.4) Comparison with the Previous Studies

Chapter Five

Conclusion and Recommendation

5.1 Conclusion

This study emphasis of five phases starting with collect images, preprocessing, features extracting, individual classification and end with testing and evaluating . six features extracted from each image, these moments are: mean, standard deviation, skewness, kurtosis, contrast and smoothness. we applied three individual classifiers, Support Vector Machine, Bayes Naïve and K-nearest Neighbors. Each of them classify Data in 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).

Our study aimed to build and implement the voting method on three classifiers (BNC, KNN, SVM) apply on medical image that is extracted from MIAS data set. The study contains two main processes the first one is build for each classifier using the 60,70,85 percentage to training set from the data set and after building the classifier, the 40,30,15 percentage of data is used in test stage The accuracy of the voting is 0.7647.

5.2 Recommendation

The current research resided mainly on classification accuracy as the main criteria for measuring the performance of proposed approaches. To achieve better performance in terms of accuracy we recommend to use more than six features and apply it into the decision tree model. However, future work will focus in other criteria such as classification speed and computational cost. In addition, breast cancer dataset used in this study has taken from MIAS Data set. We also hope get help from ministry of higher education and scientific research to obtain real dataset from Federal Ministry of Health .we also recommend using another features and using noise removal. To increase accuracy of the classifier.

References

[1] data mining concepts and techniques ,Jiawei Han and Micheline Kamber, 2nd edition .

[2] Fundamentals of Database Systems , Ramez Elmasri ,6th Edition .

[3] Freitas, A.A., Data mining and knowledge discovery with evolutionary algorithms. 2013: Springer Science & Business Media.

[4] Aggarwal, C.C., Data classification: algorithms and applications. 2014: CRC Press.

[5] Beura, S., B. Majhi, and R. Dash, Mammogram classification using two dimensional discrete wavelet transform and gray-level co-occurrence matrix for detection of breast cancer. Neurocomputing, 2015. 154: p. 1-14.

[6] "Classification of Mammography Image Using Machine Learning Classifiers and Texture Features",International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163 Issue 9, Volume 2 (September 2015) .

[7] Roberta Fusco , Mario Sansone , Antonella Petrillo, Carlo Sansone ,"A Multiple Classifier System for Classification of Breast Lesions Using Dynamic and Morphological Features in DCE-MRI" -2012.

[8] " Breast Cancer-Early Detection and Classification Techniques: A Survey", International Journal of Computer Applications (0975 – 8887) Volume 132 – No.11, December2015 9.

[9] P. Smitha, L. Shaji, and M. Mini, "A review of medical image classification technique," in International conference on VLSI, Communication &Intrumrnataiom, 2011, pp. 34-38.

[10] Grigorios Tsoumakas, Lefteris Angelis and Ioannis Vlahavas ,"Selective Fusion of Heterogeneous Classifiers ",Dept. of Informatics, Aristotle University of Thessaloniki.

[11] "A Study On Preprocessing A Mammogram Image Using Adaptive Median Filter"International Journal of Innovative Research in Science, Engineering and Technology (An ISO 3297: 2007 Certified Organization) Vol. 3, Issue 3, March 2014.

[12] "Study of Mammogram Microcalcificationtoaidtumour detection using Naive BayesClassifier", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering(An ISO 3297: 2007 Certified Organization)Vol. 3, Issue 3, March 2014 .

[13] Keyvanfard, F., et al.: Specificity enhancement in classification of breast MRI lesion based on multi-classifier. Neural Computing and Applications (2012), doi:10.1007/s00521-012-0937-y.

[14] Performance Analysis of Majority Vote Combiner for Multiple Classifier Systems - 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA) ,-11 Dec. 2015.

[15] http://peipa.essex.ac.uk/info/mias.html

[16] S. J. S. Antony and S. Ravi, "A New Approach to Determine the Classification of Mammographic Image Using K-Means Clustering Algorithm," 2015.

[17] R. Aarthi, K. Divya, N. Komala, and S. Kavitha, "Application of Feature Extraction and clustering in mammogram classification using Support Vector Machine," in Advanced Computing (ICoAC), 2011 Third International Conference on, 2011, pp. 62-67.

[18] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 1.21 (2011)," Available: cvxr. com/cvx, 2010.

[19] Kimitaka Tsustumi, Kazutaka Shimada and Tsutomu Endo ,"Movie Review Classification Based on a Multiple Classifier" ,Kyushu Institute of Technology , Lizuka , Fukoka 820-8502 Japan,2007.

[20] Overall survival of females with breast cancer in the national cancer institute University of Gezira, sudan , 2015.

[21] ERIC BAUER,RON KOHAVI, "An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting and Variants",1998.

[22] R.NTTHYA,B.SANTHI, "Comparative study on features extraction method for breast cancer classification" , Journal of Theretical and Applied Information Technology,2011.