10-28-2004

Application Of Support Vector Machines And Neural Networks In Digital Mammography: A Comparative Study

Nivedita V. Candade
University of South Florida

Application Of Support Vector Machines And Neural Networks In Digital Mammography: A Comparative Study

by

Nivedita V. Candade

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Biomedical Engineering
Department of Chemical Engineering
College of Engineering
University of South Florida

Major Professor: Wei Qian, Ph.D.
Barnali Dixon, Ph.D.
William E. Lee III, Ph.D

Date of Approval:
October 28, 2004

Keywords: breast cancer, microcalcifications, pattern recognition, feature selection, Free Receiver Operating Characteristics (FROC)

ACKNOWLEDGEMENTS

I take this opportunity to express my gratitude to all the people who have helped me accomplish this work, directly or indirectly. I would like to appreciate all the people who have contributed to my life, in the present and the past.

I thank my major professor, Dr. Wei Qian, Associate Professor, Imaging Science Research Division, Moffitt Cancer Research Center, for giving me an opportunity to work on this project. I am immensely grateful to his group, Dr. Xue Jun Sun and Dansheng Song for helping me a great deal with my work. Both of them have been very involved in this project and have been more than willing to help me out.

I would like to thank Dr. Barnali Dixon for being an exceptional guide and advisor. She has not only helped me accomplish this work in many ways, but given me fantastic opportunities in my academic career and opened doors to prospects that I never dreamt of. I am greatly indebted to her for what I am today, for she has truly made a difference in my life. I thank her for being a great friend, philosopher and guide. Without her support and constructive criticism, this work would never have reached completion. I also thank Dr. William Lee for being on my committee, for his time and comments.

I would like to thank some of my friends who have seen me through this work. Dr. Arun Karpur, a close friend and a statistics pro for his invaluable time and help with data analysis. He has never hesitated to help me out or teach me and I really appreciate his time and patience. I would like to thank Sruthi for helping me with any question on my thesis. She has been my 24 hour ‘help line’ and a great friend. I wish to thank my friends and roommates Jyothi, Rashmi and Preksha for being my ‘family’ and for the great times I had in Tampa. Bharath and Balaji for their help and concern and for cheering me up during all my adversities. All these guys and Nishant for always welcoming me with a smile during times I gate crashed at their place. This semester has not been easy with hurricanes and health problems and all these people have helped me pull through these obstacles. I thank my friends for being there for me and for moments that I will cherish throughout my life.

Finally, I would like to thank my Mom and Dad for being my pillars of strength, love and encouragement. They are the wind beneath my wings; and mere words cannot articulate my gratitude and respect for them. Good upbringing, education and unconditional love... these were the greatest gifts they could give me!

This space is probably not enough to remember and thank all the people who have been a part of my life. There is a part of me in two ends of the world and it goes unsaid that wherever I may go, whatever I may do, my family and friends back home would remain the most precious jewels in my treasure trove! I feel blessed and fortunate to have such wonderful people as part of my life’s journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACR	American College of Radiology
ACS	American Cancer Society
AFP	Average False Positive
ANN	Artificial Neural Network
BIRADS	Breast Imaging Reporting and Data System
CAD	Computer Aided Diagnosis
CC	Cranial Caudal
CV	Cross Validation
CWMF	Central Weighted Median Filters
DN	Digital Number
FCM	Fuzzy C-Means algorithm
FN	False Negative
FNR	False Negative Rate
FP	False Positive
FPR	False Positive Rate
FROC	Free Receiver Operating Characteristic
LL	Log Likelihood
LM	Latero Medial
LOO	Leave One Out

MC	Micro Calcification
ML	Medio Lateral
MLE	Maximum Likelihood Estimation
MLO	Medio Lateral Oblique
NN	Neural Network
OSH	Optimal Separating Hyperplane
PC	Principal Components
PCA	Principal Component Analysis
QP	Quadratic Programming
RBF	Radial Basis Function
ROC	Receiver Operating Characteristic
ROI	Region of Interest
SBE	Stepwise Backward Elimination
SBP	Standard Back Propagation
SFS	Stepwise Forward Selection
SMO	Sequential Minimal Optimization
SRM	Structural Risk Minimization
SSE	Sum of Square Error
SVM	Support Vector Machine
TN	True Negative
TNR	True Negative Rate
TP	True Positive
TPR	True Positive Rate

TSF	Tree Structured Filter
VC	Vapnik Chervonenkis dimension
WT	Wavelet Transformation

APPLICATION OF SUPPORT VECTOR MACHINES AND NEURAL NETWORKS IN DIGITAL MAMMOGRAPHY: A COMPARATIVE STUDY

Nivedita V. Candade

ABSTRACT

Microcalcification (MC) detection is an important component of breast cancer diagnosis. However, visual analysis of mammograms is a difficult task for radiologists. Computer Aided Diagnosis (CAD) technology helps in identifying lesions and assists the radiologist make his final decision.

This work is a part of a CAD project carried out at the Imaging Science Research Division (ISRD), Digital Medical Imaging Program, Moffitt Cancer Research Center, Tampa, FL. A CAD system had been previously developed to perform the following tasks: (a) pre-processing, (b) segmentation and (c) feature extraction of mammogram images. Ten features covering spatial, and morphological domains were extracted from the mammograms and the samples were classified as Microcalcification (MC) or False alarm (False Positive microcalcification/ FP) based on a binary truth file obtained from a radiologist's initial investigation.

The main focus of this work was two-fold: (a) to analyze these features, select the most significant features among them and study their impact on classification accuracy and (b) to implement and compare two machine-learning algorithms, Neural Networks (NNs) and Support Vector Machines (SVMs) and evaluate their performances with these features.

The NN was based on the Standard Back Propagation (SBP) algorithm. The SVM was implemented using polynomial, linear and Radial Basis Function (RBF) kernels. A detailed statistical analysis of the input features was performed. Feature selection was done using Stepwise Forward Selection (SFS) method. Training and testing of the classifiers was carried out using various training methods. Classifier evaluation was first performed with all the ten features in the model. Subsequently, only the features from SFS were used in the model to study their effect on classifier performance. Accuracy assessment was done to evaluate classifier performance.

Detailed statistical analysis showed that the given dataset showed poor discrimination between classes and proved a very difficult pattern recognition problem. The SVM performed better than the NN in most cases especially on unseen data. No significant improvement in classifier performance was noted with feature selection. However, with SFS, the NN showed improved performance on unseen data. The training time taken by the SVM was several magnitudes lesser than the NN. Classifiers were compared on the basis of their accuracy and parameters like sensitivity and specificity. Free Receiver Operating Curves (FROCs) were used for evaluation of classifier performance.

The highest accuracy observed was about 93% on training data and 76% for testing data with the SVM using Leave One Out (LOO) Cross Validation (CV) training. Sensitivity was 81% and 46% on training and testing data respectively for a threshold of 0.7. The NN trained using the 'single test' method showed the highest accuracy of 86% on training data and 70% on testing data with respective sensitivity of 84% and 50%. Threshold in this case was -0.2. However, FROC analyses showed overall superiority of SVM especially on unseen data.

Both spatial and morphological domain features were significant in our model. Features were selected based on their significance in the model. However, when tested with the NN and SVM, this feature selection procedure did not show significant improvement in classifier performance. It was interesting to note that the model with interactions between these selected variables showed excellent testing sensitivity with the NN classifier (about 81%).

Recent research has shown SVMs outperform NNs in classification tasks. SVMs show distinct advantages such as better generalization, increased speed of learning, ability to find a global optimum and ability to deal with linearly non-separable data. Thus, though NNs are more widely known and used, SVMs are expected to gain popularity in practical applications. Our findings show that the SVM outperforms the NN. However, its performance depends largely on the nature of data used.

CHAPTER 1

INTRODUCTION

Breast cancer is the second leading cause of cancer deaths in women today (after lung cancer). According to the World Health Organization, more than 1.2 million people will be diagnosed with breast cancer this year worldwide (Imaginis, 2004). Currently, approximately 3 million women in the US are living with the disease (Center, 2004). According to American Cancer Society (ACS) estimates, 215,990 cases of invasive breast cancer will be diagnosed in 2004. In the same year, it is also estimated that 1,450 men will be diagnosed with breast cancer. Year 2004 estimates include nearly 40,580 deaths occurring from breast cancer in US alone. According to the National Cancer Institute, one out of eight women will develop breast cancer during her lifetime.

Breast cancer stages range from Stage 0 (very early form of cancer) to Stage IV (advanced, metastatic breast cancer) (Imaginis, 2004). Early stage breast cancers are associated with high survival rates than late stage cancers.

The key to surviving breast cancer is early detection and treatment. According to the ACS, when breast cancer is confined to the breast, the five-year survival rate is almost 100%. Breast cancer screening has been shown to reduce breast cancer mortality (Society, 2004). Currently, 63% of breast cancers are diagnosed at a localized stage, for which the five-year survival rate is 97%. The high survival rates of early detection of breast cancer can be attributed to utilization of mammography screening as well as high levels of awareness of the disease symptoms in the population.

1.1 Motivation

Mammography is used for breast cancer screening and diagnosis for the detection and characterization of abnormalities that maybe malignant (Association, 2002). Approximately 85% sensitivity (proportion of positives detected correctly as a disease) is achieved with conventional film-screen mammography, though results are operator dependent and may vary with reader expertise. A lot of research has gone into finding techniques that can improve sensitivity and reduce variability among readers.

One method of reducing missed MCs or the false-negative (FN) rate in screening mammography is the double reading of mammograms (Anttinen I, 1993; Thurfjell E.L., 1994). Investigations of this method reported increase in cancer detection rates by as much as 15% (Hendee WR, 1999). However, this method is both time consuming and not cost-effective.

The incorporation of computer algorithms to increase sensitivity in screening mammography has gained popularity in recent years (Chan HC, 1990; Kregelmeyer WP, 1994; Nishikawa RM, 1995; te Brake GM, 1998; Vyborny, 1994; Warren Burhenne LJ, 2000). Findings indicated the potential of Computer Aided Diagnosis (CAD) to reduce the false negative (FN) rate by 50%-70%.

CAD systems use computerized algorithms for identifying suspicious regions of interest (ROIs). The motivation behind CAD systems is to reduce both the False Positive (FPR) and False Negative rates (FNR). When used as intended, CAD would be expected to increase the number of mammograms interpreted as positive to the extent that it points out abnormalities previously overlooked by the radiologist. On the other hand, the cost of missed or undetected abnormalities (FNs) is very high.

This work presents a part of a CAD scheme for the detection of microcalcifications in mammograms using NNs and SVMs. This would be an aid to a radiologist who would have already outlined suspected abnormalities. This system provides a classification scheme which would aid the radiologist make his final diagnosis.

Research (Edwards DC, 2000; Woods KS, 1993; Zhang W, 1996) has shown that the use of classifiers based on Artificial NNs (ANNs or simply NNs) can improve the performance of a detection scheme. NNs (Hagan MT, 1996) have been successful in many applications, especially for clustering (Park, 2000) and pattern recognition (Gader PD, 1997). In recent years, the SVM (Chapelle O, 1999; Pontil M., 1998; Vapnik, 1995, 1998) has become an effective tool for pattern recognition, machine learning and data mining, because of its high generalization performance.

Given a set of points that all belong to one of the two classes, an SVM can find the hyperplane that leaves the largest possible fraction of points of the same class on the same side, while maximizing the distance of either class from the hyperplane. This optimal separating hyperplane can minimize the risk of misclassifying examples of the test set. On the other hand, NNs are based on the minimization of empirical risk, which is the minimization of the number of misclassified vectors of the training set.

SVMs are attracting increasing attention because they rely on a solid statistical foundation and appear to perform quite effectively in many different applications (Lecun Y, 1995; M. Pontil, 1998; Osuna E, 1997). After training, the separating surface is expressed as a certain linear combination of a given kernel function centered at some of the data vectors (named support vectors). All the remaining vectors of the training set are effectively discarded and the classification of new vectors is obtained solely in terms of the support vectors. SVMs also offer other advantages over multivariate classifiers. They are free of optimization problems of NNs because they present a convex programming problem, and guarantee finding a global solution. They are much faster to evaluate than density estimators (like maximum likelihood classifiers), because they make use of only the relevant data points, rather than looping over each point regardless of its relevance to the decision boundary. Recent research has suggested that the SVM is superior to the NN (Burbidge R, 2001; Ding CH, 2001; Liang H, 2001). In this study, both the algorithms were used to classify microcalcifications from false positive signals (or false alarms) and evaluated.

1.2 Objectives and Approach

CAD systems consist primarily of the following processing stages: (a) Pre-processing, (b) Segmentation (c) Feature extraction and (d) classification. Stages (a)-(c) were a part of previous work conducted on this dataset. The mammograms were first studied for abnormalities before they were given to the CAD system. Pre-processing was performed to reduce noise and artifacts and to enhance the image (Qian W, 1994). Segmentation was used to identify suspicious areas from the whole image. Feature extraction and selection is a crucial part of the CAD classification process and has a significant impact on classification accuracy. Ten features were extracted and given as inputs to the classification stage (Qian W, 2001). This work focused on the classification (Stage (d)) and feature selection. The database consisted of 22 mammograms, which included Cranial Caudal (CC) and Medio Lateral Oblique (MLO) view images of the breast.

The NN and SVM algorithms were implemented and evaluated for their performance. The NN was constructed using the MATLAB NN toolbox. The network used the Standard Back Propagation (SBP) algorithm for training. The SVM classifier was obtained in C using the LIBSVM toolbox (Chih-Chung Chang, 2001). LIBSVM is an integrated software for support vector classification, regression and distribution estimation. It supports multi-class classification. The basic algorithm is a simplification of both Sequential Minimal Optimization (SMO) by Platt (Platt, 1999), and SVMLight by Joachims (Joachims, 1999). It is also a simplification of the modification of SMO by Keerthi et al (Keerthi, 1999). Several kernel options are supported by the classifier.

A detailed feature analysis was performed to evaluate the relationships between the input features and the outcome. From this feature analysis, the most significant features were selected and tested with the classifiers. Free Receiver Operating Characteristic

(FROC) curves were plotted for each experiment and compared. The classification algorithms were compared and the feature selection process was assessed.

1.3 Study Outline

This document has been organized as follows. Chapter 2 discusses the medical background of breast cancer, literature review of detection methods and CAD systems in mammography. Chapter 3 gives a description of the classification algorithms studied. Chapter 4 gives a description of the developed CAD module and the Materials and Methods used in this study. Results and discussion are presented in Chapter 5. Chapter 6 presents the conclusions and future work.

CHAPTER 2

LITERATURE REVIEW

As the second leading cause of cancer-related mortality in women, it is crucial that breast cancer be detected in its early stages of development. Mammography has been used as a screening and diagnostic tool for the early detection of breast cancer. Screening mammography has proven to be effective for women 50-75 years of age (Kerlikowske K, 1995). A recent study showed that in women aged 40-49 years; screening mammography reduces breast cancer mortality by 16-18% (Rajkumar S, 1999). 80-85% of breast cancers are visible on a mammogram as a mass, calcification or combination of both (Mckenna RJ, 1994). CAD methods play an important role in improving diagnostic accuracy in mammogram interpretation. This chapter provides a background on mammography, types of mammograms, types of abnormalities and an introduction to automated methods in breast cancer detection.

2.1 Background: Mammography

A mammogram is a test that is done to look for any abnormalities in a woman's breasts. The test uses an X-ray machine to take pictures of both the breasts. With digital mammography, once the images are taken, they can be electronically manipulated. Digital mammography offers certain advantages over film mammography. Results can be obtained much faster; the doctor can electronically manipulate the images (zoom in, magnify etc.) and transmit the images to another site for viewing and printing (Systems, 2003).

Mammograms look for breast lumps and changes in breast tissue that may develop into problems over time. They can find abnormalities that a woman or a health care provider cannot feel during a physical examination. Breast lumps can be *benign* (non-cancerous) or *malignant* (cancerous). A *biopsy* is done if a lump is found, where a small amount of tissue is taken from the lump and the area around the lump. This tissue is then tested for cancer. Early detection of breast cancer increases the chances of a woman surviving the disease.

Figure 1 shows a sample mammogram.

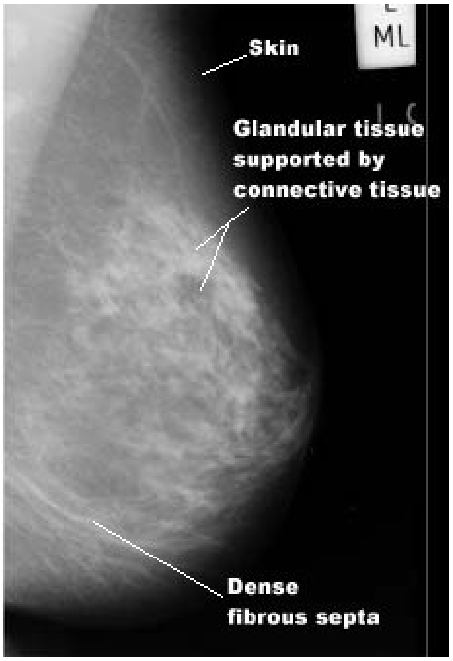


Figure 1 ***Mammographic Anatomy Of The Breast ("Interactive Mammography Analysis Web Tutorial", 1999)***

2.2 Types of mammography

Two types of mammography exams are in practice today: *Screening* and *Diagnostic*.

2.2.1 Screening mammography

This is performed to detect breast cancer when it is too small to be felt by a physician or a patient. It is performed on women with no complaints or symptoms of breast cancer (Imaginis, 2004). The procedure involves taking x-ray images of two views for each breast. These views are typically from above (Cranial-Caudal view, CC) and from an angled view (Medio Lateral Oblique, MLO). The MLO is probably the most important and most common view taken followed by the CC. These views are represented in Figure 2.

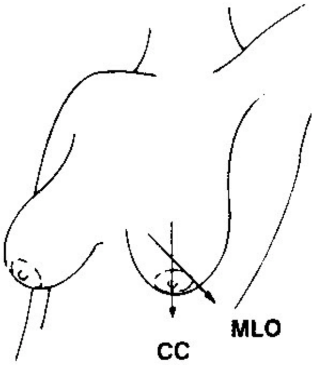


***Figure 2** Views In Screening Mammography- Cranio- Caudal (CC) And Mediolateral Oblique (MLO) Views (Imaginis, 2004)*

2.2.2 Diagnostic mammography

This is performed on a patient who has been evaluated as symptomatic by a physical exam or screening mammography. Additional views of the breast are usually taken as against two in screening mammography, hence making it a more time-consuming and costly procedure. The objective here is to determine the exact size and location of abnormality and to image the surrounding tissue and lymph nodes. Diagnostic mammography helps determine malignancy, following which a biopsy maybe ordered. Biopsy is the only definitive way to ascertain breast cancer (Imaginis, 2004).

Diagnostic mammography typically involves two additional views, the Latero Medial (LM) and the Medio Lateral view (ML) apart from the CC and MLO views discussed earlier. Additional views maybe taken depending on the nature of the problem.

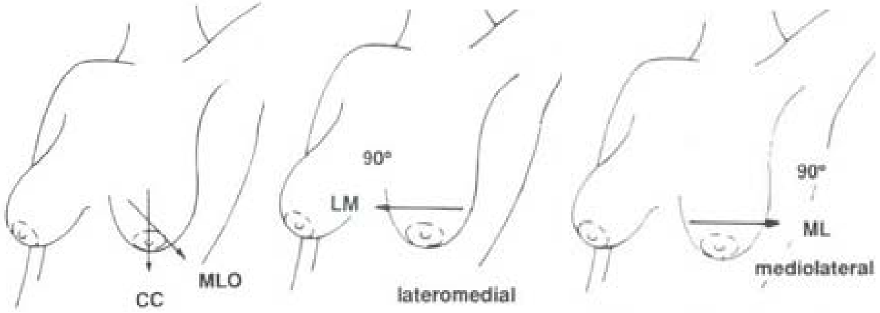


***Figure 3** Views In Diagnostic Mammography. (Left) Cranio-Caudal (CC) And Mediolateral Oblique (MLO) Views, (Center) Latero Medial (LM) View, (Right) Medio Lateral (ML) View (Imaginis, 2004)*

2.3 Mammographic Abnormalities

A suspicious abnormality normally falls into three broad categories: (1) Asymmetric density, (2) Masses (including architectural distortion) and (3) Calcifications (Imaginis, 2004). Masses often have distinguishing shape, size and margin characteristics. Likewise, calcifications can be characterized by their size, number, morphology, distribution and heterogeneity. These are the distinguishing characteristics based on which a mammogram maybe classified as benign or possibly malignant. Masses and calcifications are the most common features associated with cancer. They are discussed below.

2.3.1 Mass

Masses are three-dimensional lesions which may represent a localizing sign of breast cancer. A mass is a group of cells clustered together more densely than the surrounding tissue. A (non-cancerous) cyst may appear as a mass in a mammographic film. Masses can be caused by benign breast conditions or by breast cancer (Imaginis, 2004). They are characterized by their location, size, shape, margin characteristics, x-ray attenuation, effect on surrounding tissue, and other associated findings like architectural distortion, associated calcifications and skin changes. A mass could be round, oval, lobular, irregular or have architectural distortion. Mass margins as defined by Breast Imaging Reporting and Data System (BI-RADS) include: circumscribed, obscured, micro-lobulated, ill-defined and speculated (Figure 4).

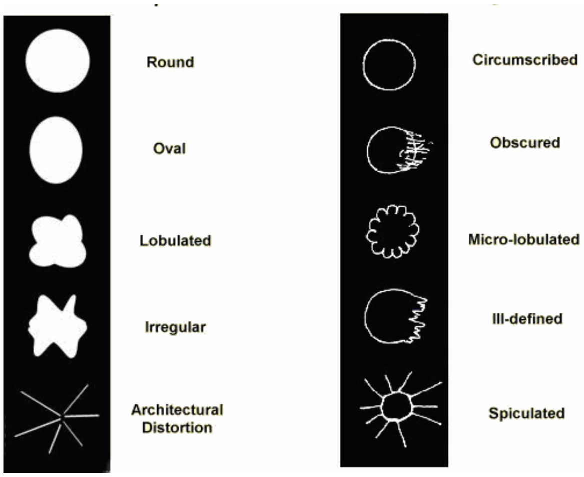


Figure 4 Descriptors For (Left) Shape, (Right) Margins (Imaginis, 2004)

2.3.2 Calcification

Microcalcifications are tiny (less than 1/50 of an inch or ½ of a millimeter) specks of Calcium that maybe found in an area of rapidly dividing cells (Nagel Rufus H, 1998). Calcifications are often important and common findings in mammograms. They may be intramammary, within and around the ducts, within the lobules, in vascular structures, in interlobular connective tissue or fat. When many are seen in a cluster, they may indicate a small cancer. About half the cancers detected appear as these clusters. Microcalcifications are the most common mammographic sign of ductal carcinoma in situ (an early cancer confined to the breast ducts).

Most breast calcifications are benign. The term microcalcification is often used for calcifications found with malignancy, which are usually smaller, more numerous, clustered, and variously shaped (rods, branches, teardrops). Calcifications associated with benign conditions are usually larger, fewer in number, widely dispersed and round. These are termed macro-calcifications. In the middle are hard-to-tell calcifications that are often labeled indeterminate. The number of calcifications that make up a cluster can be used as an indicator of benign and malignancy. While the actual number itself is arbitrary, a minimum number of either four, five or six calcifications per cluster is considered to be of significance. The morphology of calcifications is considered to be the most important indicator in differentiating benign from malignant. As discussed earlier, round and oval shaped calcifications are more likely to be benign. Those associated with malignant processes resemble small fragments of broken glass and are rarely rounded or smooth (Imaginis, 2004).

The American College of Radiology (ACR) BIRADS has classified findings of calcifications into three categories (Table 1):

(a) Typically benign
(b) Intermediate concern
(c) High probability of malignancy

Table 1 Summary Of BIRADS Classification Of Calcifications

	Type of calcification	Characteristics
Typically benign	Skin	typical lucent center and polygonal shape
	Vascular	parallel tracks or linear tubular calcifications that run along a blood vessel
	Coarse or pop-corn like	Involuting fibroadenomas
	Rod-shaped	Large rod-like structures usually > 1mm
	Round	Smooth, round clusters
	Punctuate	Round or oval calcifications
	Spherical or lucent centered	Found in debris collected in ducts, in areas of fat necrosis
	Rim or egg-shell	Found in wall of cysts.
	Milk or calcium	Calcium precipitates
	Dystrophic	Irregular in shape but usually large > 0.5mm in size
Intermediate concern	Indistinct or amorphous	Appear round or flake shaped, small and hazy uncertain morphology
High risk	Pleomorphic or heterogenous	Cluster of these calcifications irregular in shape, size and < 0.5mm raises suspicion
	Fine, linear or branching	Thin, irregular that appear linear from a distance

2.4 Limitations of Mammograms

Mammography can help detect breast cancer at an early stage, when the chances for successful treatment and survival are the greatest. Mammography can detect about 85% to 90% of breast cancers. However, mammographic films maybe difficult for the radiologist to read and in some cases, abnormalities maybe overlooked. Also, False Negatives (FN) and False Positives (FP) are possible. FN means even though the mammogram may look normal, cancer is actually present. An FP occurs when the results shows the presence of cancer, even though this is not the case (4woman.gov, 2002). Younger women are more likely to have an FN mammogram because the breast tissue is denser, making cancer harder to spot. In such cases where there is ambiguity in results, a second interpretation would help the radiologist make his final decision.

The CAD technology works as a “second reading” for radiologists, alerting them to areas on the image that require his attention. The following section describes the CAD system, its benefits and limitations and its components in detail.

2.5 Computer Aided Diagnosis (CAD) for Mammography

CAD is a recent advance in the field of breast imaging. Studies on CAD technology estimate that for every 100,000 breast cancers currently detected with screening mammograms, the CAD technology could result in the detection of an additional 20,500 breast cancers.

In CAD, the computer marks abnormalities on the digitized films. After reviewing the results from CAD, the radiologist decides whether the marked area is indeed an abnormality that is of concern.

Mammograms are first loaded into a special processing unit that digitizes the mammogram images. The CAD unit incorporates special pattern recognition algorithms to highlight any detected breast abnormalities. In the meantime, the radiologist reviews the patient’s mammogram and makes his interpretation. He then views the mammogram from the CAD system and modifies his/ her interpretation if appropriate. CAD technology is designed to detect masses and calcifications in digital mammograms.

2.5.1 Components of CAD

The goal of a CAD system in this work is the detection of MCs and the reduction of false positive MCs on mammograms. The goal is also to achieve high sensitivity in order to detect MCs that a radiologist might miss. Clinical utility would depend strongly on the number of FPs per image, since radiologists must take extra time and care to read areas of the mammograms with FPs (Rufus H. Nagel, 1995). FPs can also reduce the confidence a radiologist has in using a CAD system. Therefore, it is important to reduce the number of computer FPs, while maintaining high sensitivity.

There are many methods that can be used to classify MCs. Rule based methods (Chan HP, 1987; Davies DH, 1992) and NNs (Yoshida H, 1994; Zhang W, 1996) are two examples of these methods. The overall process involves several steps that include pre-processing, segmentation, feature extraction and classification (Figure 5). Each module of the CAD process is discussed in sections below with emphasis on classification and evaluation modules.

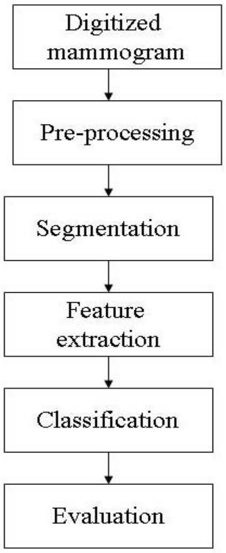


Figure 5 Stages In A CAD Process

2.5.1.1 Pre-processing

This module involves noise and artifact reduction, and intensity adjustment. Image enhancement is usually performed by noise reduction or contrast enhancement. Increase in contrast is very essential in mammograms, especially for dense breasts (Ted C. Wang, 1998). Contrast between the malignant tissue and the normal dense tissue maybe present in the mammogram but may not be discernable to the human eye. As a result, defining the characteristics of MCs is difficult (Ted C. Wang,1998).

Conventional image processing techniques may not work well on mammographic images because of the large variation in feature size and shape (W. Morrow,1992). There are two possible approaches to enhancing mammographic features. One is to *suppress background noise* and the other is to *increase the contrast of suspicious areas*.

Noises due to intrinsic characteristics of imaging device and from imaging process will impact detection sensitivity of CAD. Several types of filters have been reported (Qian W, 1994).

Non- linear filtering has proven more robust than linear filtering in preserving details of the image during noise reduction. Median filtering and selective median filtering locally adapt to the image gray scale using empirically derived threshold criteria (Lai SM, 1989). Selective median filtering is generally based on restricting the set of pixels within the selected window to those pixels with a difference in gray level not greater than an empirically derived threshold. However, detail preservation maybe lost since some pixels might be ignored within the filter window (Lai SM, 1989). Other methods like straight line windowing (Chan HP, 1987) and hexagonal windows (Glatt A, 1992) have been introduced to non-linear filtering. Though these methods were more successful for noise suppression than linear approaches, they did not necessarily show significant improvements in image detail preservation.

Multi-stage filtering is introduced in order to combine the properties of single filters. The tree-structured nonlinear filter, a symmetric multistage filter combining the advantages of *Central Weighted Median Filters* (CWMF), linear and curved windows, shows more robust characteristics for noise suppression and detail preservation. This filter is a three-stage filter designed with CWMFs as subfiltering blocks (Qian W, 1995; Qian W, 1999) applied to each pixel within the filter window (Bamberger RH, 1992; Qian W, 1994). CWMFs are a class of *median filters* where the basic principle involves replacing a pixel value with the median of the neighboring pixel values (Ko SJ, 1991).

The *weighted median filter* is an extension of the median filter, which gives more weight to some values within the window (Ko SJ, 1991), i.e. a weight coefficient is assigned to each position in a window. The filter output is the median of the sequence of pixel values; additionally, if weight coefficient is *n* at a position, the value at this position appears *n* times. As more emphasis is placed on the central weights, the filter's ability to suppress noise and preserve image details increases (Qian W, 1994).

The *Tree Structured Filter (TSF)* is a symmetric multistage filter that sequentially compares filtered and raw image data with the objective of obtaining more robust characteristics for noise suppression and detail preservation (Arce GR, 1989; Bauer PH, 1991). The TSF architecture consists of cascaded CWMFs (Qian W, 1994). Since noise is suppressed at each stage, the overall performance of the TSF is considered to be superior (Arce GR, 1989; Bauer PH, 1991).

2.5.1.2 Enhancement and Segmentation

Following noise suppression and artifact removal, image enhancement is performed to improve digital image quality. Enhancement algorithms using the wavelet transformation (WT) are used where the data is cut up into different frequency components using mathematical functions called '*wavelets'*. Each component is then studied with a resolution matched to its scale (Graps, 2004). This method has advantages over other enhancement techniques like the Fourier transform in analyzing physical situations where the signal contains discontinuities and sharp spikes.

Segmentation is used to identify suspicious areas from the whole image. Mammographic lesions are extremely difficult to identify because their radiographic and morphological characteristics resemble those of normal breast tissue. As a mammogram is a projection image, lesions do not appear as isolated densities but are overlaid over parenchymal tissue patterns.

The fuzzy C-means (FCM) algorithm was used for soft segmentation based on fuzzy set theory. It allows for fuzzy pixel classification based on iterative approximation of local minima to global objective functions. This has two advantages over other segmentation approaches, namely it is unsupervised and is robust to missing and noisy data. This algorithm helps differentiate small size suspicious regions.

2.5.1.3 Feature extraction and Classification

Feature extraction and selection is an important part of supervised classification. The number of features selected for breast cancer detection reported in literature varies with the CAD approach employed. It is desirable to use an optimum number of features since a large number of features would increase computational needs, making it difficult to define accurate decision boundaries in a large dimensional space. Features in different domains (morphological, spatial, texture etc.) are extracted. In this process, the most important characteristics of the ROI are studied. Among the most important characteristics reported by radiologists are given below (Wouter J, 2000).

(a) *Polymorphism vs. monomorphism*: MCs that are malignant tend to polymorph while benign clusters are mostly characterized by monomorphous calcifications of uniform size (Lanyi, 1988).

(b) *Size and contrast*: some benign calcifications have larger size and contrast compared to malignant calcifications.

(c) *Branching vs. round and oval type*: linear calcifications maybe an indication of Ductal Carcinoma in situ, since such calcifications are located in the glandular ducts. Benign calcifications are mostly round or oval in shape and are often located in the lobules.

(d) *Orientation*: malignant calcifications often have shapes that are oriented to the nipple (Lanyi, 1988)

(e) *Number*: A cluster with very few MCs is regarded as less suspicious. Five or more calcifications, measuring less than 1 mm, in a volume of one cubic centimeter, are considered to form a cluster (Popli, 2001).

(f) *Location*: About 48% of the cancerous processes are located in the outer upper quadrant of the breast. Lesions located in this quadrant are more suspicious (Harris JR, 1991).

Several methods for feature extraction have been proposed in literature. The use of wavelet features and gray level statistical features was proposed by Songyang Yu et al (Songyang Yu, 2000). MCs are considered to be relatively high-frequency components buried in the background of low-frequency components and very high-frequency noise in the mammograms. Wavelets have a multiresolution property since they are localized in both space and frequency domains. This property makes it suitable for extracting MCs from low-frequency backgrounds and high-frequency noise. Spatial features which describe gray level statistics like median contrast (Kong, 1998) and normalized gray level value (Stetson PF, 1997) are used in combination with wavelet features to describe MCs.

Huai Li et al (Huai Li, 1997) suggested a deterministic fractal approach to the enhancement of MCs. Since MCs can be characterized by different shapes, and possess structures with high local self-similarity, these tissue patterns can be constructed by fractal models (Huai Li, 1997).

Features in morphological and spatial domain are most commonly used for MC detection. Once the feature extraction is complete, these features are used for classification.

Several automated classification techniques have been investigated for the detection of MCs in mammograms. The k-Nearest Neighbor approach is a relatively simple and fast classification method (Wouter J, 2000). A statistical method based on the use of statistical models and the general framework of Bayesian image analysis was developed by Karssemeijer et al (Karssemeijer, 1993; N.Karssemeijer, 1991). Another method is based on a difference image technique in which a signal suppressed image is subtracted from a signal enhanced image to remove structured background noise in the mammogram (Chan HP, 1987). Global and local thresholding were then used to extract potential MC signals. Yoshida et al (Yoshida H, 1994) used decimated wavelet transform and supervised learning for the detection of MCs. Zheng et al (Zheng B, 1994) proposed a method for the detection of MCs using mixed feature-based NNs. A fuzzy logic based approach was proposed by Cheng et al (Cheng HD, 1998). Issam El-Naqa et al (Issam El-Naqa, 2002) used the SVM to detect MCs based on finite image windows. Their approach relies on the capability of the SVM to automatically learn the relevant features for optimal detection. In their work, a sensitivity of as high as 98% was achieved.

Recent studies have shown the superiority of SVM over other techniques, suggesting that SVM is a promising technique for MC classification. A detailed description of the NN and SVM approaches to MC/ FP classification used is given in Chapter 3.

2.5.2 Feature Selection

Feature selection is an important part of any classification scheme. The success of a classification scheme largely depends on the features selected and the extent of their role in the model. The objective of performing feature selection is three fold: (a) improving the prediction performance of the predictors, (b) providing faster and more cost effective predictors and (c) providing a better understanding of the processes that generated the data (Isabelle Guyon, 2003).

There are many benefits of variable and feature selection: it facilitates data visualization and understanding, reduces the storage requirements, reduces training times and improves prediction performance. The discrimination power of the features used can be analyzed through this process. The goal is to eliminate a feature if it gives us little or no additional information beyond that subsumed by the remaining features (Daphne Koller, 1996). Only a few features may be useful or 'optimal' while most may contain irrelevant or redundant information that may result in the degradation of the classifier's performance. Irrelevant and correlated attributes are detrimental because they contribute noise and can interact counter- productively to a classifier induction algorithm (Chun-Nan Hsu, 2002).

The information about the class that is inherent in the features determines the accuracy of the model (Daphne Koller, 1996). Theoretically, having more features should give us more discriminating power. However, the real world provides us with many reasons why this is generally not the case. Irrelevant and redundant features cause problems in this context as they may confuse the learning algorithm by obscuring the distributions of the small set of truly relevant features for the task at hand. In light of this, a number of researchers have recently addressed the issue of feature subset selection in machine learning. As defined by (John G, 1994) this work is often divided along two lines: *filter* and *wrapper* models.

In the *filter* model, feature selection is performed as a pre-processing step to induction (Figure 6). Induction refers to the classification algorithm.

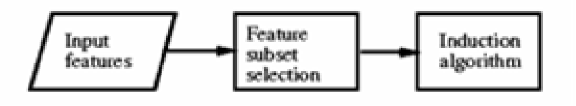


Figure 1 The Filter Model

Methods using criteria such as correlation coefficients and entropy measures that do not involve the inducer come under the category of filter models.

Many researchers in machine learning found difficulties in this classical definition of the "optimal" feature subset and the filter model. John et al (John G, 1994) point out that to measure the relevance of a given feature, one must take the existence and relevance of other features into account. In follow up work, Kohavi et al (Kohavi, 1995) consider that the optimality of a feature subset depends on both the specific induction algorithm and the training data at hand. This implies that an "optimal" feature subset for a given induction algorithm should be defined as a subset such that the induction algorithm can generate a hypothesis with the highest predictive accuracy. Feature selection should focus on finding features that are "useful" for improving the predictive accuracy rather than necessarily finding the "theoretically optimal" ones. Since the filter model ignores the effect of the feature subset on the performance of the classifier induction algorithm, an alternative method of feature selection called the *wrapper model* is proposed. The Wrapper model "wraps" around the induction algorithm (Figure 7). The idea is to generate a set of candidate feature subsets, use the induction algorithm to generate a hypothesis for each candidate feature subset, and evaluate candidate feature subsets by the classification performance of the resulting hypotheses. Methods like Forward Selection and Backward Elimination come under this category.

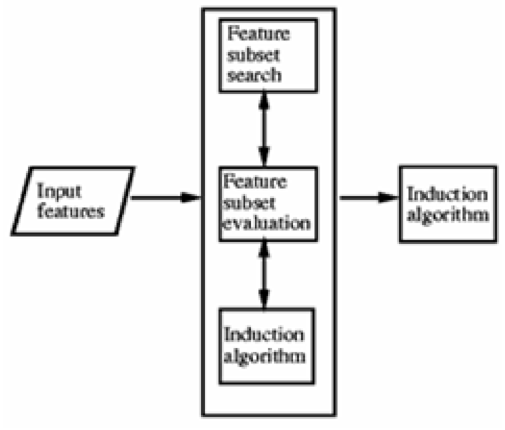


Figure 2   The Wrapper Model

The disadvantage of the wrapper model is that since a large number of training cycles is required to search for the best performing feature subset, it can be prohibitively expensive.

Wrappers try to solve the real world problem, hence optimizing the desired criterion. They are very time consuming. Filters on the other hand are much faster. Also, filters provide a generic selection of variables, not tuned for/ by a given learning machine (Isabelle Guyon, 2003). Another justification is that filtering can be used as a preprocessing step to reduce space dimensionality and overcome over fitting.

Several feature selection techniques have been discussed in literature. All these methods determine the relevancy of the generated feature subset candidate towards the classification task. There are five main types of evaluation functions (Dash M, 1997):

(a) distance (Euclidean distance measure)
(b) information (entropy, information gain, etc.,)
(c) dependency (correlation coefficient)
(d) consistency (minimum features bias)
(e) classifier error rate (based on a classification algorithm)

The first four are filter models while the last one comes under the wrapper model. Within the filter model, different feature selection algorithms can be further categorized into two groups, namely *feature weighting algorithms* and *subset search algorithms* depending on whether they evaluate the goodness of features individually or through feature subsets.

The *distance measure* calculates the physical distance (Dash M, 1997), where the main assumption is that instances of the same class must be closer than those in different class.

Entropy is a measure of the uncertainty of a feature (Lei Yu, 2003). The entropy of a variable (or feature) X is defined in Equation 1.

$$H(X) = -\sum_i P(x_i)\log_2(P(x_i)) \qquad (1)$$

And the entropy of a variable X after observing the value of another variable Y is defined by Equation 2.

$$H(X \mid Y) = -\sum_j P(y_i)\sum_i P(x_i \mid y_j)\log_2(P(x_i \mid y_j)) \qquad (2)$$

Where *P(x$_i$)* is the prior probabilities of all values of *X*, and *P(x$_i$|y$_i$)* is the posterior probability of *X* after observing the values of *Y*. *Information gain* (Quinlan, 1993) gives the amount by which the entropy of X decreases and reflects the additional information about X provided by Y (Equation 3).

$$IG(X \mid Y) = H(X) - H(X \mid Y) \qquad (3)$$

In Equation 3, a feature Y is regarded more correlated to feature X than to feature Z, if

$$IG(X \mid Y) > IG(Z \mid Y)$$

Another feature weighting criteria is the correlation measure which measures the correlation between a feature and a class label. The Pearson's correlation coefficient is given by Equation 4.

$$r_{X,Y} = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{(n-1)\sigma_X \sigma_Y} \qquad (4)$$

A positive correlation implies an simultaneous increase in X and Y (Struble). A negative correlation indicates increase in one variable as other decreases. If the $r_{X,Y}$ has a large magnitude, X and Y are strongly correlated and one of the attributes can be removed (Struble). On the other hand, variables that have a strong correlation with the outcome are retained in the model.

A limitation of all the methods listed above is that they may lead to the selection of a redundant subset of variables. Hence subset search methods are preferred over feature weighting methods. Isabelle et al (Isabelle Guyon, 2003) have shown that variables that are independently and identically distributed are not truly redundant. Noise reduction and better class separation can be obtained by adding variables that are presumably redundant. They have also shown that a variable that is completely useless by itself can provide a significant improvement in performance when taken with others. In other words, two variables that are useless by themselves can be useful together. Thus selecting subsets of variables could together have good predictive power, as opposed to ranking the variables according to their individual predictive power.

The wrapper methodology is based on using the prediction performance of a learning machine to assess the relative usefulness of subsets of variables. However, in practice it is necessary to decide on a search strategy that is computationally advantageous and robust against overfitting. Greedy search strategies like forward selection and backward elimination are the most popular search strategies while genetic algorithms, best-first and simulated annealing are among the others (Kohavi R, 1997).

In this work, the wrapper approach with logistic regression as an induction algorithm was used to find the best subset of features. The two most commonly used variable selection strategies are *Stepwise Forward Selection (SFS)* and *Stepwise Backward Elimination (SBE).*

The SFS begins with no features in the model. At each step, it enters the feature that contributes most to the discriminatory power of the model as measured by the likelihood ratio criterion. When none of the unselected features meets the entry criterion, the SFS process stops. The SBE on the other hand begins with all the features in the model and at each step eliminates the feature that contributes least to the discriminatory power of the model. The process stops when all the remaining features meet the criterion to stay in the model. The SFS was used in this work, details of which are given in Chapter 4.

2.5.3 Limitations of CAD

Though the use of CAD is becoming widespread, a great deal of time and effort is required to digitize the films (Imaginis, 2004). Some radiologists also believe that the CAD technology marks a fairly high number of “normal” areas as abnormalities leading to additional unnecessary and costly breast imaging and/ or biopsies.

In addition, the high cost of CAD technology may hinder its widespread use. A CAD system costs approximately $200,000, in addition to the cost of a mammography system. The price of mammograms may also rise from $10 to $15 per exam with the usage of CAD technology.

In spite of these limitations, studies continue to evaluate the advantages and disadvantages of CAD technology. The disadvantages stated above are weighed against the CAD system’s ability to diagnose cancers early, which dramatically reduces long-term treatment costs.

CHAPTER 3

CLASSIFICATION ALGORITHMS

The focus of this work was to examine the suitability of using the NN and SVM algorithms in the detection of MCs in mammograms and study their impact on classification accuracy. Support Vector Machines (SVMs) and Neural Networks (NNs) are the mathematical structures, or models, that underlie learning. They are both machine learning techniques that learn patterns based on training data, fit the models to this training data and predict or classify unseen (or future) data. The active development of NNs research started in 1970s and that of SVMs started in 1980s. Currently, both techniques are used widely even though SVMs demonstrate superior performance in various problems compared to NNs. The applications of SVMs are expected to expand even though NNs are more widely known. The following sections describe these algorithms in detail.

3.1 Machine Learning Principles

Learning tasks are usually divided into *supervised*, *unsupervised* and *reinforcement* learning (Hiep Van Khuu, 2003). We discuss the supervised learning procedure which is an approach that uses examples to model input output relationships. The input/ output pairings typically reflect a functional relationship mapping of inputs to outputs (Cristianini N, 2000). When there exists an underlying function between the inputs and outputs, it is referred to as the *target function*. The estimate of this target function is known as the *solution* of the learning problem. This is also called the *decision function* in case of a classification problem (Cristianini N, 2000). The solution is chosen from a set of candidate functions that map the input to the output domain. These set of candidate functions are termed the *hypotheses*.

The quality of learning algorithms is assessed in terms of the number of mistakes it makes during the training phase. However, it is not always possible to verify the validity of the training process especially if the function we are trying to learn does not have a simple representation. Also, frequently the training data are noisy and the input-output mapping does not guarantee the existence of an underlying function. The fundamental problem of machine learning is not just to find a hypothesis that is consistent with the training data but also works well on unseen data. This is known as the *generalization* capability which these algorithms try to optimize. It is possible that with a difficult training dataset, the hypothesis behaves like a rote learner i.e. the data in the training dataset are correctly classified, but predictions on unseen data are uncorrelated. Hypotheses that become too complex in order to become consistent are said to *overfit* (Cristianini N, 2000). The VC theory due to Vapnik and Chervonenkis gives a better insight of choosing a hypothesis space and hypothesis (Hiep Van Khuu, 2003). Assuming that the data are drawn from an unknown probability distribution *P(x,y)* and *l(.)* is some loss function signifying the error of a hypothesis, the risk of the hypothesis is given by Equation 5.

$$R[h] = \int l(h(x), y)dP(x, y) \tag{5}$$

Where *h(x)* is the hypothesis function. The risk of hypothesis over the training set is termed the *empirical risk* given in Equation 6.

$$R_{emp}[h] = \frac{1}{n}\sum_{i=1}^{n} l(h(x_i), y_i) \tag{6}$$

The primary goal is to minimize the empirical risk (error on training data). Unfortunately, this is not possible since the probability distribution is unknown. However, the risk is bounded by the inequality given in Equation 7.

$$R[h] \le R_{emp}[h] + \sqrt{\frac{d(l_n \frac{2n}{d} + 1) - l_n(\frac{\delta}{4})}{n}} \tag{7}$$

Where *d* is the VC dimension of the function class of h, and is a measure of the classifier's 'power'. This power does not depend on the choice of the training dataset and hence is a true representation of the classifier’s generalization performance. The VC dimension is the maximum number of data points a function can shatter given all possible labels. A complex function will have a higher VC dimension. This gives us a way to estimate the error on the future data based only on the training error and the VC-dimension of h. The goal is to choose a hypothesis that minimizes the empirical risk.

3.2 Neural Networks

The ANN is an information processing system inspired by the biological nervous system. It is composed of a large number of highly interconnected processing elements called *neurons*. The principle of ANN learning systems is much the same as the biological neuron; it involves adjustments to the synaptic connections that exist between the neurons.

An artificial neuron is a device with many inputs and one output. The neuron has two modes of operation, the training mode and the testing mode. In the training mode, the neuron can be trained to fire (or not) for a particular set of input patterns. In the testing mode, when a pattern is presented at the input the firing rule decides whether to fire the neuron or not. These neurons form the nodes of the NN. Each node is assigned a threshold and each interconnection between the nodes is assigned a weight that represents the strength between the neurons.

The simplest NN has a set of inputs and one output. Figure 8 shows a 1-level NN also called a *perceptron*.

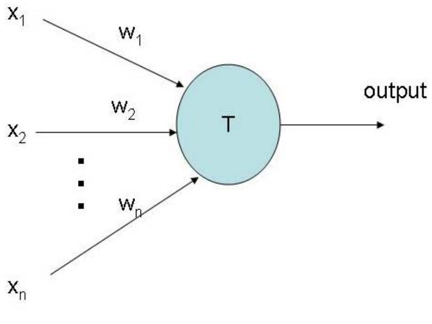


Figure 8   A Perceptron

In the above figure, *x* refers to the inputs, *w* the weights, *y* the output and *T* the threshold of the node. The strength of signals a node receives is calculated as the weighted sum of inputs

$$w_1 x_1 + w_2 x_2 + \ldots + w_n x_n \qquad (8)$$

If this value overcomes the threshold T of the node, then the signal is transmitted to other connected nodes. The value of the output of the node is decided by the activation function *f*, which decides whether the perceptron should fire or not. Thus, the output y is given as

$$y = f(w_1 x_1 + w_2 x_2 + \ldots + w_n x_n - T) \qquad (9)$$

Since Equation 9 can be interpreted as an equation of a linear line or a hyperplane, it classifies data ($x_1, x_2, \ldots x_n$) into two classes, one above the plane and one below the plane (Hiep Van Khuu, 2003).

A dataset is considered linearly separable if it requires only a single hyperplane to classify two classes. If the dataset is not linearly separable, we need more than one hyperplane. Multiple hyperplanes are represented by introducing more nodes in another layer to a perceptron. This is known as a multi-layer perceptron network as shown in Figure 9.

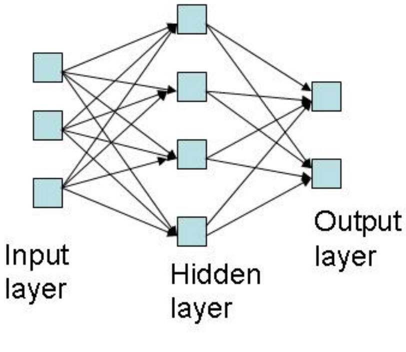


Figure 9 Multilayer Perceptron

The nodes between the input and the output layer are called hidden nodes and the layers are called hidden layers. Once the number of layers, the number of units in each layer has been established, the network's weights and thresholds must be set so as to minimize the prediction error made by the network. This is the role of the *training algorithms*. The training cases are run through the network and the output generated is compared to the desired outputs or the targets. The differences are combined together by an error function to give the network error. The most common error function is the Sum of Square Error (SSE) where the individual errors of output units on each case are squared and summed together.

3.2.1 The Standard Back Propagation Algorithm (SBP)

The SBP is the most popular NN training algorithm. Other examples of training algorithms are the conjugate gradient descent, Quasi-Newton, quick propagation etc. In BP, the gradient vector of the error surface is calculated. The vector points along the line of the steepest descent from the current point so any move in the shortest distance decreases the error. A sequence of such moves, will eventually find a minimum of some sort (Statsoft Inc., 1984-2003). Large steps converge more quickly but might overstep the solution. Small steps would require a large number of iterations. The step size is defined by the *learning rate* of the algorithm. The algorithm progresses through a number of *epochs* iteratively, the error between the target and actual outputs calculated for each epoch. This error is used to adjust the weights, and the process repeats. The initial weights are random and training stops at a set convergence criterion like a predefined number of epochs, or an acceptable level of SSE.

The BP algorithm consists of two phases: The *Forward* phase and the *Backward* phase. The feed-forward phase is where the inputs x are fed into the network. All outputs are computed using sigmoid (activation function) thresholding of the inner product of the corresponding weights and the input vectors. All the outputs at stage *n* are connected to all the inputs at stage *n+1.* Errors are then propagated backwards by apportioning them to each unit according to the amount of error the unit is responsible for (Anand, 1999).

Let *(x,t)* denote a training example where *x* and *t* are vectors representing the inputs and targets respectively. η is the learning rate. n_i, n_o and n_h are the input, output and hidden nodes respectively. Input from unit *i* to unit *j* is denoted as x_{ji} and weight is denoted by w_{ji}. The SBP algorithm is stated as follows (Anand, 1999):

(a) Create a feed-forward network with n_i inputs, n_o outputs and n_h hidden units.

```
(b) Initialize all the weights to random values (say between -0.05 and 0.05)
(c) Until convergence do
        For each training sample $(x,t)$, do
            (a) Compute the output $o_u$ of every unit for instance $x$
            (b) For each output unit $k$ calculate
                $\delta_k = o_k(1-o_k)(t_k - o_k)$
            (c) For each hidden unit $h$ calculate
                $\delta_h = o_h(1-o_h) \sum_{k \in downstream(h)} w_{kh}\delta_k$
            (d) Update each network weight $w_{ji}$ as
                $w_{ji} \leftarrow w_{ji} + \Delta w_{ji}$
                Where $\Delta w_{ji} = \eta \delta_j x_{ji}$
```

Thus the weights of the network are updated until the convergence criterion is met.

3.2.2 Over-learning and Generalization

One major problem of the above learning approach is that it doesn't actually minimize the error that we are actually interested in, the generalization error. In reality the network is trained to minimize the error on the training set. The most important manifestation of this problem is that of over fitting. A network with more weights models a more complex function, and is therefore prone to this problem. On the other hand, a network with fewer weights may not be sufficiently powerful to model the underlying function. For example, a network with no hidden layers actually models a simple linear function. Thus, it is important to select the optimum number of hidden units. In view of this, a simple model is preferred to a highly complex network.

The performance of the NN depends on other factors such as nature of the datasets. It is of relevance here to mention the problem of having unbalanced datasets. Since a network minimizes the overall error, the proportion of the classes of data in the set is critical. A network trained with 1000 positive cases and 100 negative cases will bias its decision towards the positive case, as it allows the algorithm to lower the overall error. It is also important that the training and testing data are representative of the underlying model.

3.3 Support Vector Machines

The foundations of SVM have been developed by Vapnik, and are gaining popularity due to many attractive features, and promising empirical performance. The formulation of SVM embodies the Structural Risk Minimization (SRM) principle, as opposed to Empirical Risk Minimization (ERM) commonly employed with other statistical methods. SRM minimizes the upper bound on the generalization error, as against ERM which minimizes the error on the training data. Thus, SVMs are known to generalize better.

The SRM technique consists of finding the optimal separation surface between classes due to the identification of the most representative training samples called the *support vectors*. If the training dataset is not linearly separable, a kernel method is used to simulate a non-linear projection of the data in a higher dimensional space, where the classes are linearly separable. Here, we first introduce the foundation of SVMs- the linear learning machine. SVM kernels and other components are then explained.

3.3.1 Structural Risk Minimization (SRM)

A linear learning machine learns a linear classifier (Hiep Van Khuu, 2003) or *hyperplane* from the training data (Equation 10).

$$h(x) = w.x + b,\ \ w \in R^N,\ \ b \in R \qquad (10)$$

Thus the hyperplane divides the data so that that all the points with the same label lie on the same side of the hyper plane. This amounts to finding *w* and *b* so that

$$y_i(w.x_i + b) > 0 \qquad (11)$$

It is possible to rescale w and b so that

$$y_i(w.x_i + b) \geq 1 \qquad (12)$$

This system of equations can have several solutions as shown in Figure 10.

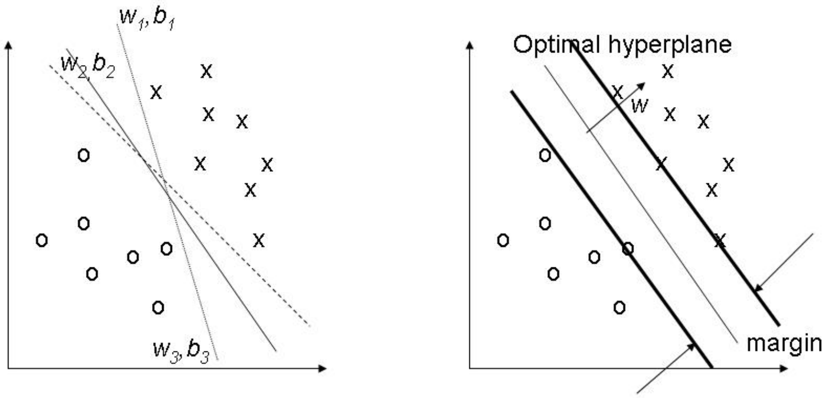


Figure 10 (Left) Several Feasible Hyperplanes, (Right) Optimal Separating Hyperplane

The SRM approach is based on minimizing both the terms in the RHS of Equation 7. The classifier that has the maximal margin to the training set is the preferred solution among all other feasible hyperplanes shown in Figure 10 (Left). This choice of hyperplane gives a tighter bound on the VC dimension and reduces the risk. Thus determining the classifier involves the Quadratic Optimization Problem (QP) of minimizing $\frac{1}{2}\|w\|^2$ under constraints (12). Thus, the N dimensional vector *w* and the real vector *b* define the OSH.

This concept can be extended to the case when the classes are not linearly separable, i.e. when Equation 12 has no solution. A non-linear mapping $\Phi : R^N \rightarrow R^L$ which maps the input data to a high dimensional space (also called the *feature space*) is introduced. Here, L is usually much larger than N. We can then try to find a linear classifier in feature space.

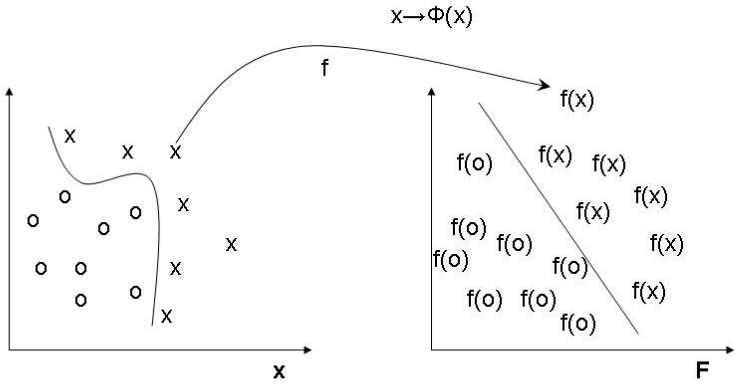


Figure 11 Kernel Mapping From Input Space To Feature Space

The problem of finding a hyperplane in feature space is one of reformulating the linear case. Thus the problem is one of minimizing $\frac{1}{2}\|w\|^2$

Subject to $y_i.((w.\Phi(x_i))+b)\geq 1,\ i=1,.....,n.$ (13)

We introduce Lagrange multipliers $\alpha_i \geq 0$, *i=1,...,n,* for each constraint in Equation (13) and find the saddle point (or minimum) of the Lagrangian

$$L(w,b,\alpha)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{n}\alpha_i(y_i((w.\Phi(x_i))+b)-1) \quad (14)$$

At the saddle point we have

$$\frac{\partial L}{\partial b}=0\ \ and\ \ \frac{\partial L}{\partial w}=0$$

Which translate into

$$\sum_{i=1}^{n}\alpha_i y_i=0\ \ and\ \ w=\sum_{i=1}^{n}\alpha_i y_i\Phi(x_i) \quad (15)$$

Substituting (15) in (14), we have the dual quadratic optimization problem

Maximize $\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j\Phi(x_i)\Phi(x_j)$ (16)

Subject to $\alpha_i\geq 0,\ i=1,....,n$ (17)

$$\sum_{i=1}^{n}\alpha_i y_i=0$$

In Equation (16), the inner product $\Phi(x_i)\Phi(x_j)$ can be replaced with a kernel function $K(x_i,x_j)$ that obeys *Mercer's condition*. Mercer's condition states that any positive semi-definite kernel $K(x_i,x_j)$ can be expressed as a dot product in high-dimensional space. Thus we avoid translating the input data to feature space first and then finding their inner products. This is equivalent to mapping the feature vectors into a high-dimensional feature space before using a hyper plane classifier there (Figure 11). The use of kernels makes it possible to map the data implicitly into a feature space and to train a linear machine in such a space, potentially side-stepping the computational problems inherent in evaluating the feature map (Cristianini N, 2000)

In a high dimensional feature space R^L the hyperplane is defined by the L dimensional vector w and real number b (Hiep Van Khuu, 2003). L can however be very large, hence storing w and b explicitly is expensive and sometimes impossible. In equations (15) and (17) the vector w is defined by the input vectors that have the non-zero Lagrange multipliers associated with them. These non-zero coefficients are called the *support vectors*, which together implicitly define the hyperplane. New data $\bar{x}$ is classified with

$$\bar{y} = \mathrm{sgn}(\sum_{i=1}^{l} \alpha_i y_i k(x_i, \bar{x}) + b) \qquad (18)$$

where l is the number of support vectors.

In this research, three kinds of kernels are studied. These kernels are mathematically defined in Equations 19-21 (Chang and Lin, 2003):

1. Polynomial kernel. $K(x, y) = ((\gamma * x' * y) + 1)^d$ (19)

2. RBF kernel. $k(x, y) = \exp\left(-\gamma * |u - v|^2\right)$ (20)

3. Linear kernel. $k(x, y) = x' * y$ (21)

There is no theory regarding which kernel is the best, given a problem domain. It is important to select the appropriate kernel based on the specific application.

Training of SVMs requires the solution of a very large Quadratic Programming (QP) optimization problem which is very time-consuming (Platt, 1999). Sequential Mimimal Optimization (SMO) is an algorithm for training the SVM where this large QP problem is broken down into a series of smallest possible QP problems which are solved analytically. SMO can handle very large training datasets and considerably speeds up training times.

SMO solves the smallest possible optimization problem at every step. The smallest possible optimization problem involves two Lagrange multipliers. At every step, the SMO chooses two Lagrange multipliers to jointly optimize, finds the optimal values for these multipliers, and updates the SVM to reflect the new optimal values.

3.4 Comparison of SVMs and NNs

Both NNs and SVMs are based on the concept of linear learning models, using linear hyperplanes to classify data. For non-linear models, the approach of these two algorithms is different. SVMs use non-linear mappings to find a decision hyperplane in feature space. On the other hand, NNs use activation functions such as sigmoid, radial or Gaussian to handle non-linear data, so that the BP algorithm can compute the weight change depending on the error on the output. These activation functions in effect, create some non-linear decision boundary classifying the input data into different classes.

SVMs minimize the structural risk (error on unseen data) while NNs minimize only the empirical risk (training data). Hence, SVMs are known to generalize better with a better learned hypothesis function that approximates more closely to the true classification function.

NNs are known to have longer training times since the learning process involves training with the dataset repeatedly to better learn the hypothesis function that will perform the classification task. NNs learn better the more times they get trained. SVMs on the other hand, handle data simultaneously, without losing the degree of accuracy.

NNs converge to local minima while the SVMs find a global solution. The problem of overfitting in NNs might get them stuck at local optima while with SVMs, the bound on the true risk and the QP solution always ensures a global solution.

With a good understanding of the mathematical foundations of these algorithms, we explain our methodologies and results in Chapter 4 and 5.

CHAPTER 4

MATERIALS AND METHODS

This chapter presents the overall approach used in the CAD system. A description of the database and the features used is included. Detailed feature analysis as well as feature selection is performed prior to classification. Classification using NN and SVM algorithms with and without feature selection are studied and evaluated.

4.1 Schematic of Proposed CAD System

The objective of this thesis project was three-fold:

(a) Analyze the input features and their importance in predicting the outcome. Perform feature selection by selecting the most significant features.

(b) Use all the ten features with the NN and SVM to classify MC and FP and compare their performances.

(c) Use the most significant features and their interactions (from Step 1) with the NN and SVM to classify MC and FP and compare their performances.

The overall procedure however consists of multiple steps like pre-processing, segmentation, feature extraction, classification and evaluation. The schematic of the entire procedure is shown in Figure 12.

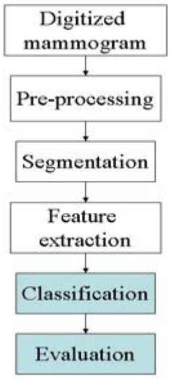


Figure 12 Schematic Of CAD System

The ‘Classification’ and ‘Evaluation’ stages were the main focus of this work. Classification was performed using the NN and SVM algorithms, the schematic for which is given in Figure 13.

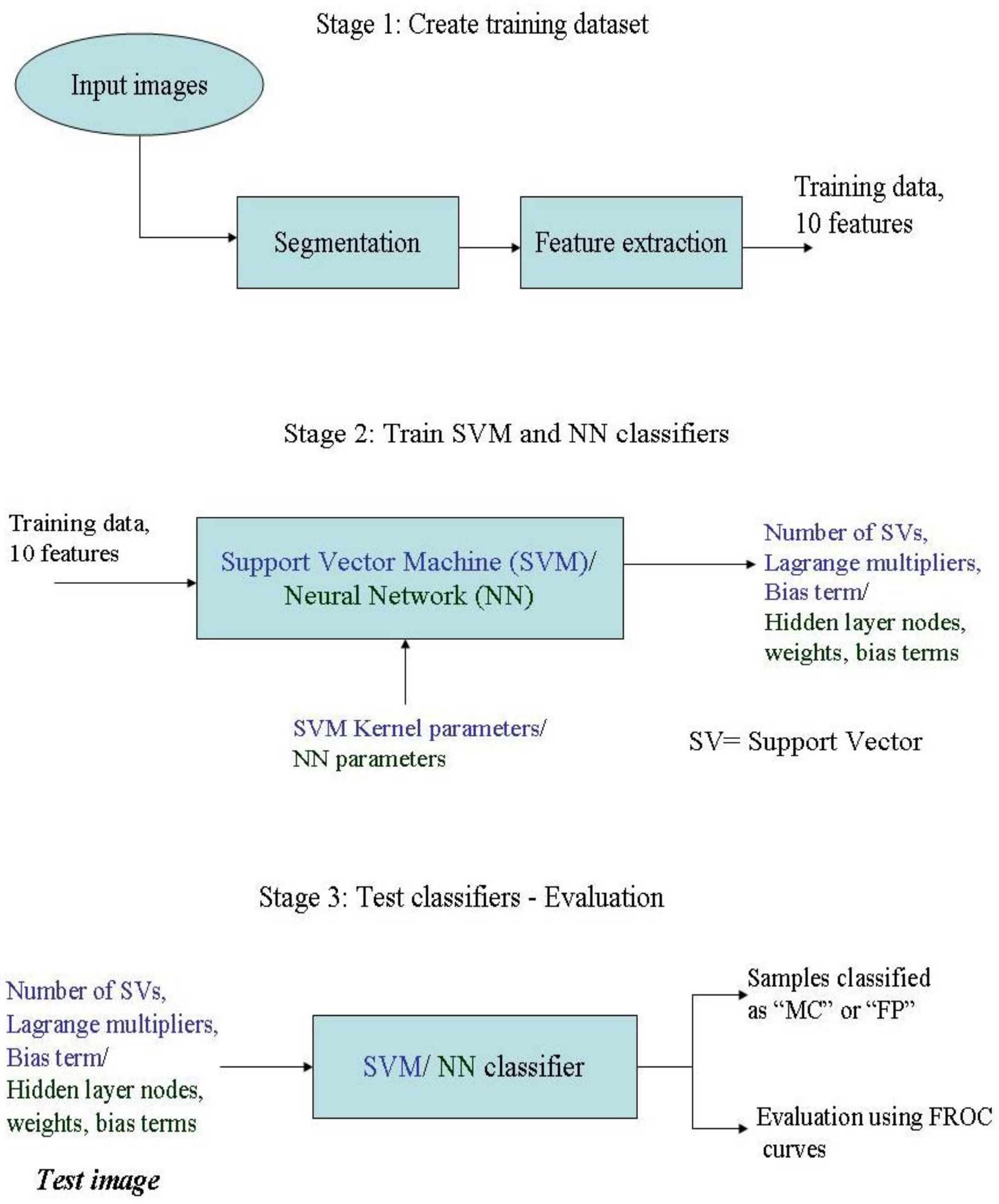


Figure 13 Detailed Schematic Of Training And Testing Of SVM And NN Algorithms

4.2 Database Description

The database consisted of 22 images of 60 micron resolution of which all the images had a case of abnormality, marked out by a radiologist. These images included the CC and MLO views of each breast. Figures 14 and 15 show two examples of abnormal cases. The images shown are the raw mammogram, the Region of Interest (ROI) marked out by a radiologist and a portion of the image after segmentation. Figure 16 shows other examples of MCs that were identified by the radiologist in this database.

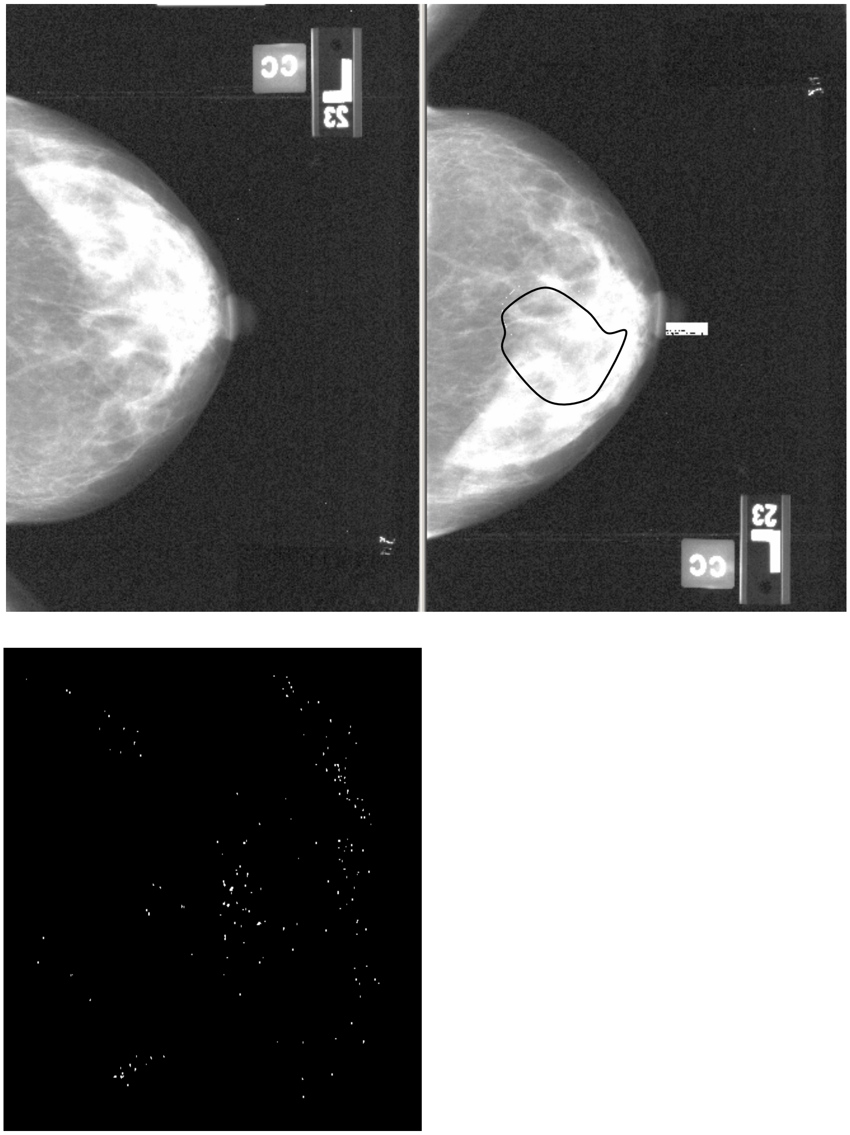


***Figure 14** Arch Distortion With Suspected Microcalcification, (Top Left) Raw Image, (Top Right) ROI, (Bottom Left) Section Of Segmented Image Including MCs And FPs*

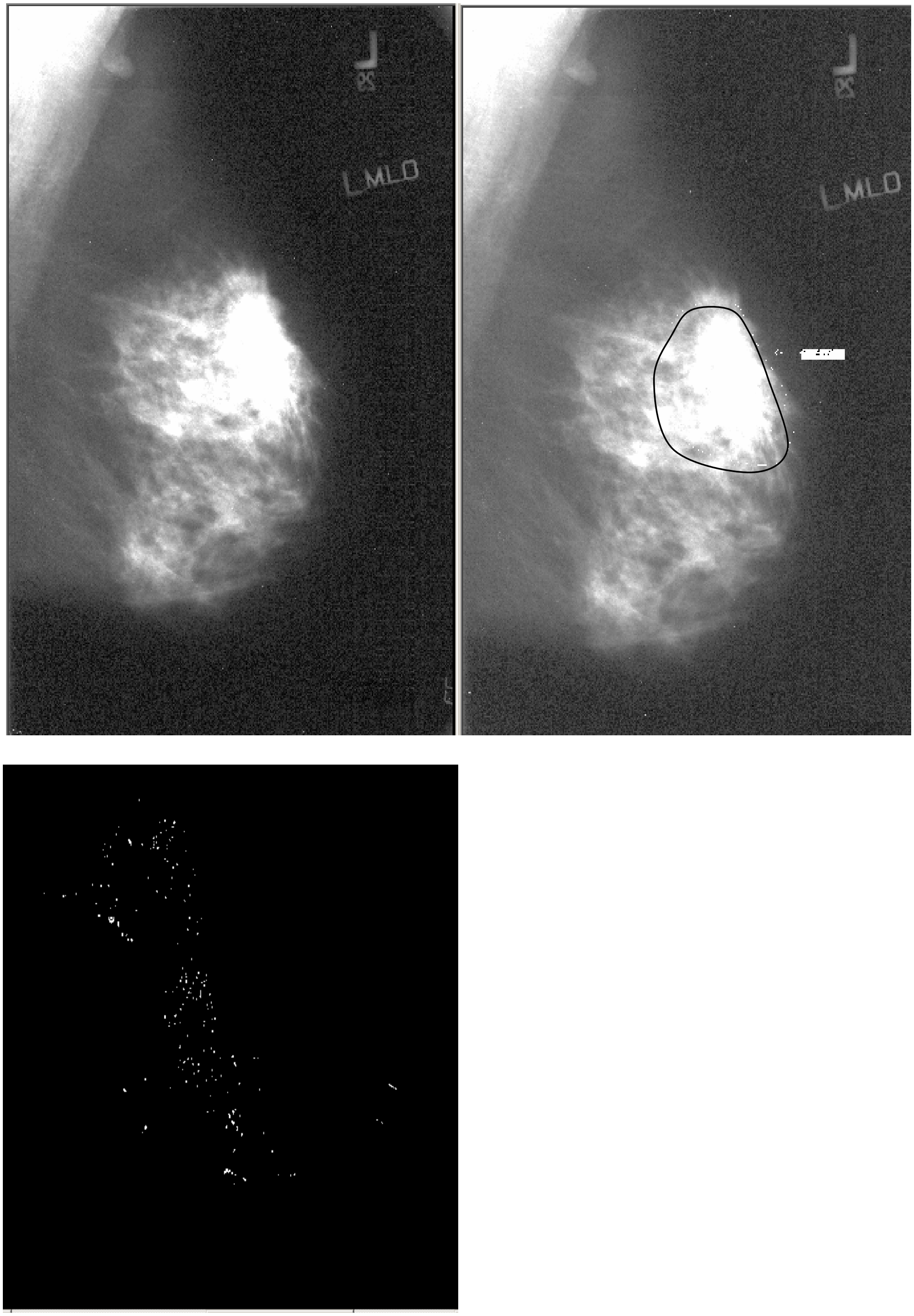
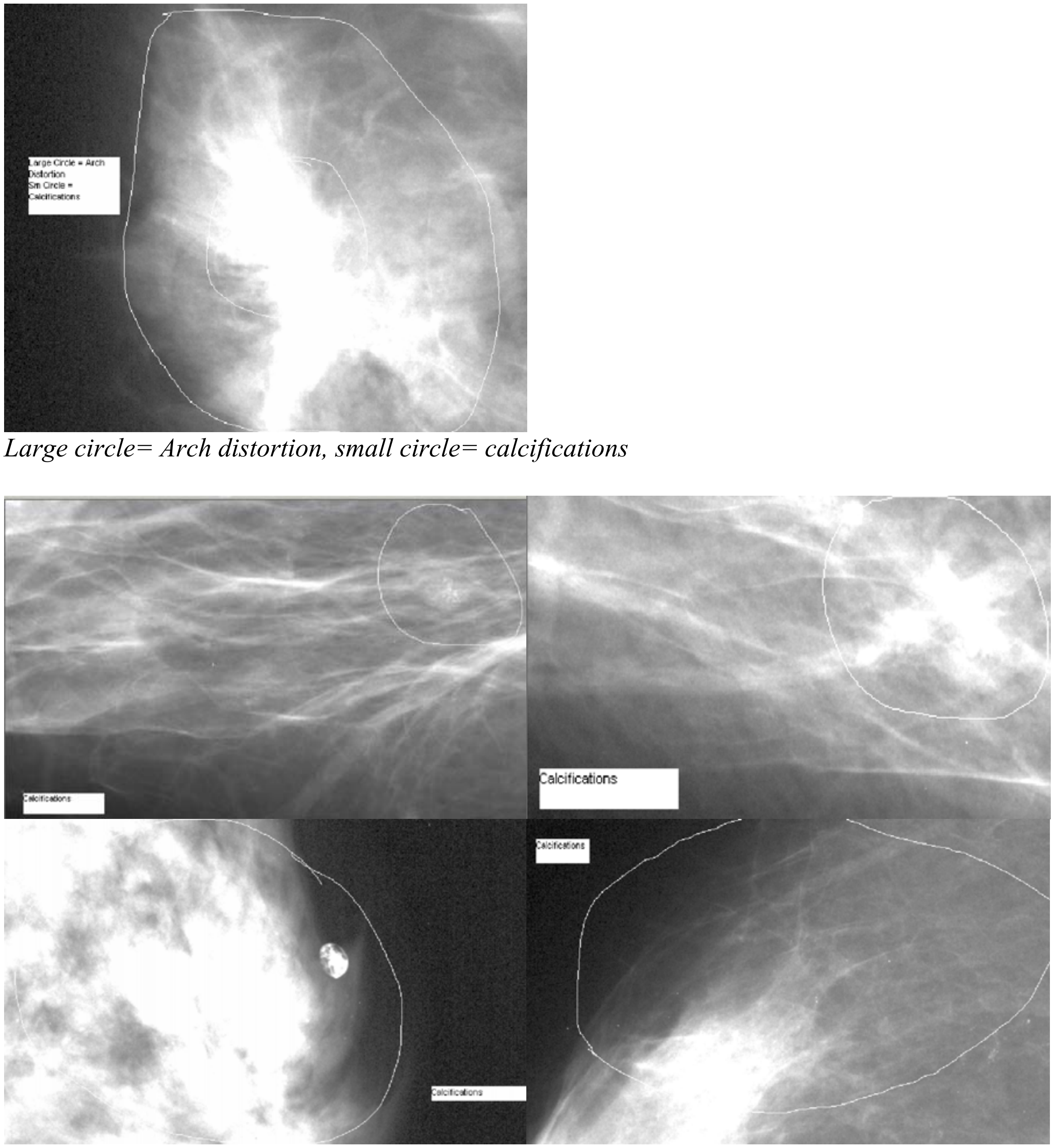


Figure 15 Image Containing Both The Arch Distortions And Faint Microcalcifications, (Top Left) Raw Image, (Top Right) ROI, (Bottom Left) Section Of Segmented Image, Includes MCs And FPs



Large circle= Arch distortion, small circle= calcifications

Calcifications

Figure 16 Other Examples Of Microcalcifications Outlined By The Radiologist

The above images consist of both the MC and FP signals. According to Takehiro et al (Takehiro Ema, 1995), false-positive MC signals in mammograms can be classified into four major categories: (a) MC-like noise pattern, (b) artifacts, (c) linear pattern and (d) FP signals appearing on ducts, step like edges or ring patterns. These False Positive MCs vary with database, but overall look like subtle MCs. However, careful observation would reveal their differences. Artifacts are caused by dusts or scratches in films or noise in the digitization process (Takehiro Ema, 1995). False positive MC signals have higher contrast than true MCs. MC-like noise pattern is most commonly seen, while factors (b)-(d) mentioned above also contribute to false positive MCs.

4.3 Image Pre-processing and Segmentation

Though not a direct part of this thesis, image pre-processing and segmentation were performed prior to classification and is mentioned here for completeness. Image preprocessing was performed using TSFs that used cascaded CWMFs. Adaptive WT-based enhancement algorithms were developed for digitized CAD methods. Segmentation was performed using the fuzzy C-means algorithm.

4.4 Feature Description

Subsequent to image segmentation, feature extraction was performed. Ten features that cover spatial and morphological domain and that are believed to be representative of the two classes were extracted from the segmented image. These features are listed in Table 2.

Table 2 Input Features

Feature No.	Feature	Type of feature
3	Mean entropy	Spatial
4	Deviation of entropy	
7	Average foreground	
8	Deviation foreground	
9	Mean contrast	
5	Moment	Morphological
6	Compactness	
1	Eccentricity	Describes the margins
2	Spread	
10	Boundary gradient	

4.4.1 Spatial Domain Features

These features are extracted from the enhanced output image. They describe the *entropy* and *gray-levels* of the image. Entropy refers to the disorder of a system. The entropy of a system is related to the amount of information it contains. Low entropy images, such as those containing a lot of black sky, have very little contrast and large runs of pixels with the same or similar Digital Number values (Brien). An image that is perfectly flat will have entropy of zero. On the other hand, high entropy images such as an image of heavily cratered areas on the moon have a great deal of contrast from one pixel to the next. In short, the entropy refers to the Information content of the gray values. The entropy for each ROI can be calculated using Equation 22.

$$Entropy = -\sum_{0}^{255} rel[i] * l_n\left(rel[i]\right) \tag{22}$$

Where *rel[i]* = histogram of the relative gray value frequencies

i = gray value of input image (0...255)

(a) Mean entropy: This is the average entropy value given by Equation 23.

$$\overline{Entropy} = \frac{1}{n}\sum_{i=1}^{n} Entropy_i \tag{23}$$

(b) Deviation of entropy: standard deviation of entropy values from the mean entropy, given by Equation 24.

$$SD_{entropy} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Entropy_i - \overline{Entropy}\right)^2} \tag{24}$$

(c) Average foreground: This is the average gray-level of foreground in enhanced image (Qian W, 2001) given by Equation 25.

$$Avg_{foreground} = \frac{1}{sum(pixel_{foreground})}\sum_{(m,n)\in foreground} x(m,n) \tag{25}$$

(d) Deviation foreground: Standard deviation of gray-levels of the foreground in enhanced image given by Equation 26.

$$Stdev_{foreground} = \left(\sum_{(m,n)\in foreground}\left[x(m,n) - Avg_{foreground}\right]^2\right)^{\frac{1}{2}} \tag{26}$$

(e) *Mean contrast*: Difference in gray level values of foreground and background given by Equation 27.

$$Contrast = \frac{Avg_{foreground} - Avg_{background}}{Avg_{background}} \qquad (27)$$

The above features are based on the fact that MC spots have different gray levels compared to the background tissues.

4.4.2 Morphology Domain Features

These features focus on the shape description. They are extracted from the segmented image.

(a) Compactness

Compactness is a dimensionless quantity that provides a measure of contour complexity versus the area enclosed (Gavrielides, 1996; Shen L, 1994). It is one of the most commonly used feature in pattern recognition and classification techniques (Tembey, 2003). Compactness can be defined by Equation 28.

$$\gamma = \frac{(perimeter)^2}{4\pi(area)} \qquad (28)$$

For a disc, γ would be a minimum and equals to 1.

A larger value of compactness describes an irregular and elongated object while a smaller value is representative of a more symmetric object (Tembey, 2003)

(b) Moment

The moment refers to the roughness of a contour and increases as the irregularity of the shape increases.(Castleman, 1979; Tembey, 2003). It gives information regarding the shape roughness and is used to distinguish between the different shape categories of calcifications.

For a two-dimensional image f(x,y), the moments m_{pq} of order (p+q) are defined in Equation 29 (Qian W, 2001)

$$m_{pq} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x^p y^q f(x,y)\,dxdy \qquad \text{for } p,q\text{=0,1,2....} \qquad (29)$$

While the central moments are defined as

$$\mu_{pq} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (x-\bar{x})^p (y-\bar{y}) f(x,y)\,dxdy \qquad (30)$$

Where $\bar{x} = m_{10} / m_{00}$ *and* $\bar{y} = m_{01} / m_{00}$

For a binary image, the above formula can be rewritten as

$$\bar{m} = \frac{1}{N} \sum_{(m,n)\in\Re} \sum m . \bar{n} = \frac{1}{N} \sum_{(m,n)\in\Re}\sum n$$

$$\mu_{pq} = \sum_{(m,n)\in\Re}\sum (m-\bar{m})^p (n-\bar{n})^q \qquad (31)$$

4.4.3 Boundary Definitions

(a) Boundary gradient

This feature is obtained by calculating the gradient of each boundary pixels 8 connected neighbors and taking the average of its neighbor's gradient value as its gradient.

The gradient operators are represented by a pair of masks H1 and H2, which measure the gradient of the image u(m,n) in two orthogonal directions. By defining the bi-directional gradients as g1(m,n)=<U,H1>$_{m,n}$ and g2(m,n)=<U,H2>$_{m,n}$ the gradient vector magnitude and direction are given by Equation 32.

$$g(m,n) = \sqrt{g_1^2(m,n) + g_2^2(m,n)} \quad \theta_g(m,n) = \tan^{-1}\frac{g_2(m,n)}{g_1(m,n)} \qquad (32)$$

Using the above formulae, the segmented image is first screened, labeled all the boundary pixels of each calcification, and then mapped back to the enhanced image to get their boundary pixel gradient. The gradient feature is based on the optimized algorithm, which use an initially given value and initially defined searching direction to find the optimized convergence solution for the problem. The Sobel gradient operator was used for calculating the gradient description feature used with the masks defined as:

$$H_1 = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \qquad H_2 = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \tag{33}$$

(b) Eccentricity

Eccentricity (ε) measures the degree to which an object's mass is concentrated along a particular axis. The range of values for ε is [0-1] where 0 defines a circular object and 1 a linear object. It is defined in Equation 34.

$$\varepsilon = \frac{(m_{2,0} - m_{0,2})^2 + 4m_{1,1}^2}{(m_{2,0} + m_{0,2})^2} \tag{34}$$

Where m_{pq} is the moment of order *(p+q)*

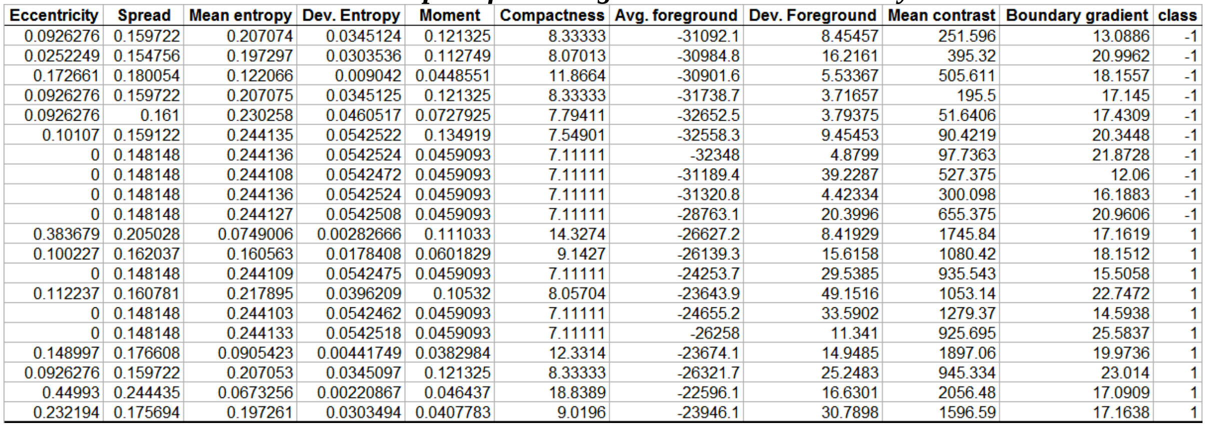
(c) Spread (S) is based on the central moments of the boundary pixels. It measures how unevenly an object's mass is distributed along its centroid and takes values in the range of [0-1]. Again, a lower value represents a circular object while a large value defines a linear and non-uniform object. Spread is defined in Equation 35 (Tembey, 2003).

$$S = \mu_{0,2} + \mu_{2,0} \tag{35}$$

where μ_{pq} is the central moment

These 10 features extracted are classified into MC and FP categories based on the truth file (marked by radiologist). A sample of the training dataset used is shown in Table 3 below. Here '-1' stands for class FP and '1' for class MC.

Table 3   Sample Of Training Data Used In The Study



Eccentricity	Spread	Mean entropy	Dev. Entropy	Moment	Compactness	Avg. foreground	Dev. Foreground	Mean contrast	Boundary gradient	class
0.0926276	0.159722	0.207074	0.0345124	0.121325	8.33333	-31092.1	8.45457	251.596	13.0886	-1
0.0252249	0.154756	0.197297	0.0303536	0.112749	8.07013	-30984.8	16.2161	395.32	20.9962	-1
0.172661	0.180054	0.122066	0.009042	0.0448551	11.8664	-30901.6	5.53367	505.611	18.1557	-1
0.0926276	0.159722	0.207075	0.0345125	0.121325	8.33333	-31738.7	3.71657	195.5	17.145	-1
0.0926276	0.161	0.230258	0.0460517	0.0727925	7.79411	-32652.5	3.79375	51.6406	17.4309	-1
0.10107	0.159122	0.244135	0.0542522	0.134919	7.54901	-32558.3	9.45453	90.4219	20.3448	-1
0	0.148148	0.244136	0.0542524	0.0459093	7.11111	-32348	4.8799	97.7363	21.8728	-1
0	0.148148	0.244108	0.0542472	0.0459093	7.11111	-31189.4	39.2287	527.375	12.06	-1
0	0.148148	0.244136	0.0542524	0.0459093	7.11111	-31320.8	4.42334	300.098	16.1883	-1
0	0.148148	0.244127	0.0542508	0.0459093	7.11111	-28763.1	20.3996	655.375	20.9606	-1
0.383679	0.205028	0.0749006	0.00282666	0.111033	14.3274	-26627.2	8.41929	1745.84	17.1619	1
0.100227	0.162037	0.160563	0.0178408	0.0601829	9.1427	-26139.3	15.6158	1080.42	18.1512	1
0	0.148148	0.244109	0.0542475	0.0459093	7.11111	-24253.7	29.5385	935.543	15.5058	1
0.112237	0.160781	0.217895	0.0396209	0.10532	8.05704	-23643.9	49.1516	1053.14	22.7472	1
0	0.148148	0.244103	0.0542462	0.0459093	7.11111	-24655.2	33.5902	1279.37	14.5938	1
0	0.148148	0.244133	0.0542518	0.0459093	7.11111	-26258	11.341	925.695	25.5837	1
0.148997	0.176608	0.0905423	0.00441749	0.0382984	12.3314	-23674.1	14.9485	1897.06	19.9736	1
0.0926276	0.159722	0.207053	0.0345097	0.121325	8.33333	-26321.7	25.2483	945.334	23.014	1
0.44993	0.244435	0.0673256	0.00220867	0.046437	18.8389	-22596.1	16.6301	2056.48	17.0909	1
0.232194	0.175694	0.197261	0.0303494	0.0407783	9.0196	-23946.1	30.7898	1596.59	17.1638	1

4.5 Data Analysis and Classification

This was performed in three different stages:

1. Input feature analysis and feature selection using Forward Selection method.
2. Use all the ten features with the NN and SVM and compare their performances.
3. Include the most significant features and their interactions (from Step 1) with the NN and SVM and compare their performances.

4.5.1 Data Analysis

The first step was a detailed analysis of input data. Since the medical implications of these features (or domain knowledge) were not known precisely, we seek a statistical explanation for the effects of the predictors. Data analysis includes the univariate statistics as well as multiple regression analyses. *Logistic regression* is a form of regression that gives us an insight into the independent variable effects, their significance and the extent of their role in the model, and their relationship with the outcome. With this understanding we continue with training and classification using NN and SVM.

Logistic regression is used to predict a dependent variable on the basis of independents (Hosmer, 1989). In logistic regression, the dependent variable is binary or dichotomous. The goal is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (outcome) and a set of independent (or predictor) variables (Cox, 1989). Logistic regression generates the coefficients of a formula to predict a logit transformation of the probability of presence of the characteristic of interest (MedCalc, 2004).

$$\log it(p) = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + ... + b_k x_k \tag{36}$$

where p is the probability of the presence of the characteristic of interest. The logit transformation is defined as the logged odds (Equation 37) where odds are given as:

$$odds = \frac{p}{1-p} = \frac{probability\, of\ presence\, of\ characteristic}{probability\, of\ absence\, of\ characteristic}$$

and

$$\log it(p) = l_n \left[\frac{p}{1-p} \right] \tag{37}$$

Rather than choosing parameters that minimize the sum of squares errors (like in ordinary regression), estimation in logistic regression chooses parameters that maximizes the likelihood of observing the sample values. This is called the Maximum Likelihood Estimation (MLE) which is a method used to calculate the logit coefficients. MLE seeks to maximize the log likelihood (LL), which reflects the likelihood of predicting the odds of the observed values of the dependent from the observed values of independents.

Logistic regression gives us the univariate effects of the variables on the outcome i.e. it gives us an idea as to how each input feature affects the classification as MC/ FP, as well as the strength of association between each input and the outcome.

4.5.2 Feature Selection

Feature selection was performed using the wrapper method explained in Section 2.5.2. The induction algorithm used in this case was logistic regression with Stepwise Forward Selection (SFS) as the search strategy. Logistic regression was used previously for data analysis to study the significance of each variable in our model. The same concept is extended to a procedure for selecting the best subset of features based on the likelihood ratio criterion. Variables are tested for individual significances (*main effect model*) and in combination (*interaction effect model*) by adding each variable stepwise into the model. It is to be noted here that the term ‘variables’ and ‘features’ are used interchangeably.

The SFS was implemented in SAS. The algorithm starts out with no predictors (features) in the model. The test is based on the “chi-square” test which is a non-parametric test of statistical significance. The initial chi-square reflects the error associated with the model when only the intercept is included in the model i.e. the initial chi-square is *-2LL* for the model which accepts the null hypothesis that all the predictors’ coefficients are zero. This statistic is then compared with the corresponding *-2LL* for the model with the predictors included. The chi-square value represents twice the difference in log likelihoods between fitting a model with only an intercept term and a model with an intercept and a predictor (independent variable). This value (difference) is compared with a chi-square distribution with degrees of freedom equal to the difference in the number of terms between the two models. If the difference is significant (p-value > chi-square is lesser than 0.05), the null hypothesis that knowing the independents makes no difference in predicting the dependent, is rejected. Thus the new variable is added into the model.

As stated earlier, it is important to study both the effects of individual independents as well as their interactions. An interaction effect is a change in the simple main effect of one variable over levels of the second. All possible two way interactions are included to test for their significance. Only two-way interactions are used since anything more than two way would not be significant due to issues of power and sample size.

The main effect and the interaction effect models give feature subsets that are optimal. These feature subsets are tested with the NN and SVM algorithms. ROC curves for both these models are plotted and evaluated based on the c-statistic. The c-statistic indicates the area under the ROC curve.

4.5.3 Classification

Classification was performed with all the (a) all ten features and (b) features selected from SFS procedure.

Ideally, the extracted features are representative of the classes that they represent. Supervised classification involves two stages: *Training* and *Testing*. Three different training techniques were used, the single test method, Leave One Out (LOO) Cross Validation (CV) and the alternate class training method.

In the single test method, 12 images out of the 22 were used for training and 10 images for testing. The training images were selected after careful optimization of the training dataset. They were selected based on the NN's performance on the selected training set and the remaining images which formed the testing set. The images that gave the best performance on the testing set were used as the training images.

The size of the dataset however was small and training with 12 images may not have produced the desired high accuracy. Also the classifier may perform well on the training dataset, but may not be able to generalize well i.e. may not produce good test results. *Cross Validation* is an alternate evaluation method to estimate how well the trained model is going to 'generalize' or perform on unseen data. This is done in order to avoid the possible bias introduced by relying on any one particular division into test and train components. The original set is partitioned in several different ways and an average score is computed over the different partitions. The extreme variant of this is to split *p* patterns into a training set of size *p-1* and a test set of size 1. This is performed *p* times and the squared error on the left out pattern is averaged over the iterations. This is called the LOO CV. In this work, LOO has been performed using the 12 images. In the first step of the procedure, the first 11 images were used for training and the last image for testing. In the next step, the next 11 images were used for training and the remaining one for testing. This procedure was carried out 12 times, so each image was used at least once for testing.

Training using alternate classes was done to account for the imbalance in the training data set since the number of FPs exceeded the MCs by almost five times. Training was performed using equal number of FPs and MCs and these classes were presented alternatively to the classifiers.

The NN was implemented in MATLAB using the NN toolbox. A feed-forward back propagation network was used which consists of the *forward* and the *backward* phases. The NN architecture consisted of 2 hidden layers with 13 units each, and an output layer with 1 unit. The transfer function for the hidden layers was 'tan-sigmoid' and for the output layer was 'linear'. The network was trained for 1000 epochs and the Sum of Squares Error (SSE) goal was set to 15. The architecture of the given NN is as shown in figure 17. Weights are initialized with random values. In the forward phase, the training inputs are given to the network. As the NN is learning, the value of error decreases. The error is propagated back to the hidden layer in the backward phase, thus modifying the weights of the network.

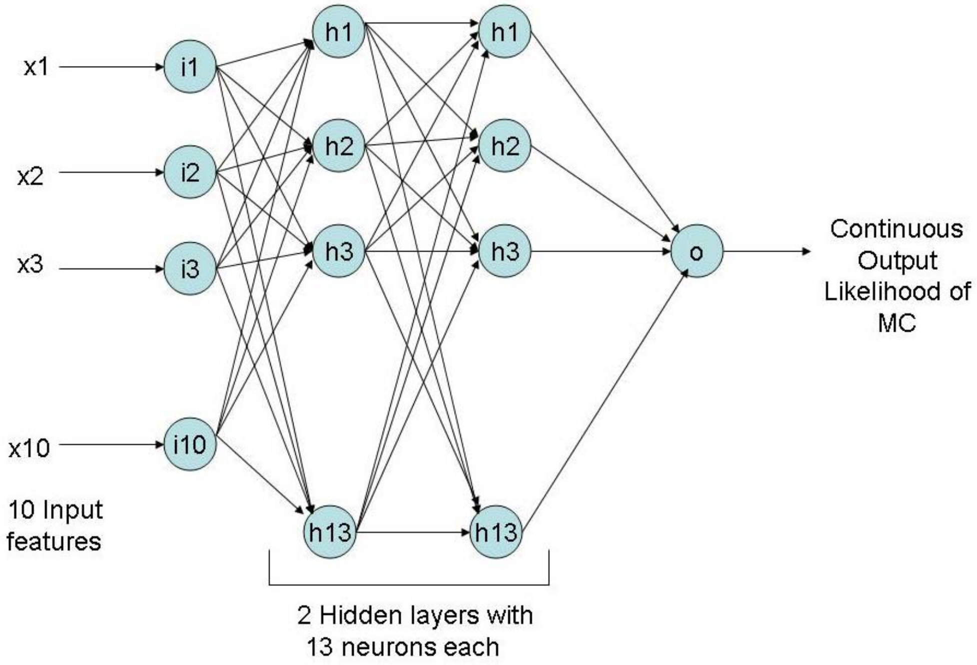

x1
x2
x3
x10
i1
i2
i3
i10
10 Input features
h1
h2
h3
h13
h1
h2
h3
h13
2 Hidden layers with 13 neurons each
o
Continuous Output Likelihood of MC

Figure 17 Architecture Of NN

The SVM was implemented using LIBSVM Version 2.6. This SVM classifier uses the Sequential Minimal Optimization (SMO) algorithm. The goal is to construct a binary classifier to derive a decision function from the available samples with the least probability of misclassifying a future sample. Different kernel functions and parameters were experimented with. The kernels included the polynomial, RBF and linear kernels with their different parameters. Initially these kernels and their parameters were compared. However, during the CV process, the best parameters are chosen by nested cross-validation procedures. The data was highly unbalanced i.e. the number of FPs outnumbered the number of MCs by 5 times. Thus they were weighted unequally to set the penalty for an MC higher than that for an FP. Also, the data was normalized and scaled before presenting it to the SVM to ease mathematical calculations as well as reduce the effect of larger attributes.

The final output of the SVM was a continuous vector ranging between 0 and 1, a value closer to 0 indicating a FP and a value closer to 1 indicating an MC. The output of the NN varied between -1 and +1. A threshold was specified on the output. If the likelihood value was greater than the threshold, then the predicted class would be ‘1’ or MC and if lesser than the threshold, the predicted class would be ‘-1’ or FP.

4.5.4 Evaluation

Evaluation of the classification algorithms was performed using two measures: *Accuracy* and *Confusion Matrix*. FROC curves were plotted by varying the threshold on the predicted output.

The confusion matrix (Kohavi, 1988) contains information about actual and predicted classifications done by a classification system. The following table shows the confusion matrix for a binary classifier:

Table 4 Confusion Matrix

	+1	-1
+1	***TP***	***FN***
-1	***FP***	***TN***

Where TP = number of correct predictions that an instance is positive

FP = number of incorrect predictions that an instance is positive

TN = number of correct predictions that an instance is negative

FN = number of incorrect predictions that an instance is negative

Based on the above values, the following evaluation criteria are defined:

(a) *Accuracy*: proportion of total number of predictions that were correct (Equation 38).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{38}$$

(b) *True Positive Rate (TPR)*: proportion of positive cases that were correctly identified (Equation 39).

$$TPR = \frac{TP}{TP + FN} \tag{39}$$

(c) *False Positive Rate (FPR)*: proportion of negatives that were incorrectly classified as positives (Equation 40).

$$FPR = \frac{FP}{FP + TN} \tag{40}$$

(d) *True Negative Rate (TNR)*: proportion of negatives that were correctly identified (Equation 41).

$$TNR = \frac{TN}{TN + FP} \tag{41}$$

(e) *False Negative Rate (FNR)*: proportion of positive cases that were incorrectly classified as negative (Equation 42).

$$FNR = \frac{FN}{FN + TP} \tag{42}$$

Accuracy alone is not an adequate measure of performance especially in our case where the number of negative cases is much greater than the number of positive cases (Kubat M, 1998). Suppose there are 100 cases, 95 of which are negative and 5 positive. If the system classified all the cases as negative, the accuracy would be 95%, even though the classifier missed all the positive cases. Thus it is important to study the other parameters described above. The FROC curve gives a graphical representation of these parameters for various thresholds on the output and encapsulates all the information contained in the confusion matrix. Here, the number of FPs/ image is plotted on the x-axis and the TPR on the y-axis. Each threshold results in an (FP, TP) pair and a series of such pairs are used to plot the FROC curve. In our case, the TPF would be the probability of correctly classifying a true MC as an MC. The FPF is the probability of incorrectly classifying a false positive (or 'false alarm' to avoid term confusion) as an MC. In medical diagnosis, these values are translated to produce two important indices of assessment: *Sensitivity* and *Specificity*. Sensitivity refers to the TPR or the proportion of patients with cancer who test positive. Specificity refers to TNR (or 1-FPR) or the proportion of patients without cancer who test negative. The position of the *cutoff* determines the number of TP, FP, TN and FN. As the sensitivity is increased, the specificity is also sacrificed. Thus, an optimum cut-off needs to be chosen, for which the sensitivity and specificity values are acceptable. Here, the TNR and TPR refer to the specificity and sensitivity of the classification stage. The *overall* specificity and sensitivity is affected by their respective values in the segmentation stage.

Classification and evaluation based on SVM and NN algorithms was carried out and their performances were compared. Also, the performances of these algorithms using all features, the most significant ones and their interactions were compared.

CHAPTER 5

RESULTS AND DISCUSSION

This chapter presents the results of various experiments conducted on the training and testing images. The results are presented as follows:

1. Detailed statistical analysis of input features
2. Logistic regression and Forward Selection
3. Classification results for different types of training methods with and without feature selection

5.1 Statistical Analysis of Features

The output of the segmentation and feature extraction process was a text file consisting of MC and FP cases. It was necessary to perform a detailed analysis of input data due to the lack of complete domain knowledge. The answers we seek are: Are the input features related to the outcome? Is there a pattern? Three types of statistical analyses were performed: *Univariate*, *Multivariate and Logistic Regression.*

Univariate analysis refers to the analysis of a single variable. This helps us get a ‘feel’ for the data by giving us an overall description of what we are working with. Univariate analysis included histogram plots and feature statistics. The simplest way to visualize the input variables in each class is to create a frequency distribution of the data on each input variable (independent) or feature. Histograms give us an idea about the distribution of data in a dataset. The vertical axis of the histogram gives the number of counts of the data in each data range or bin, the bins plotted on the horizontal axis. In this study, the histograms were plotted to give us an idea about the distribution of data values in each class.

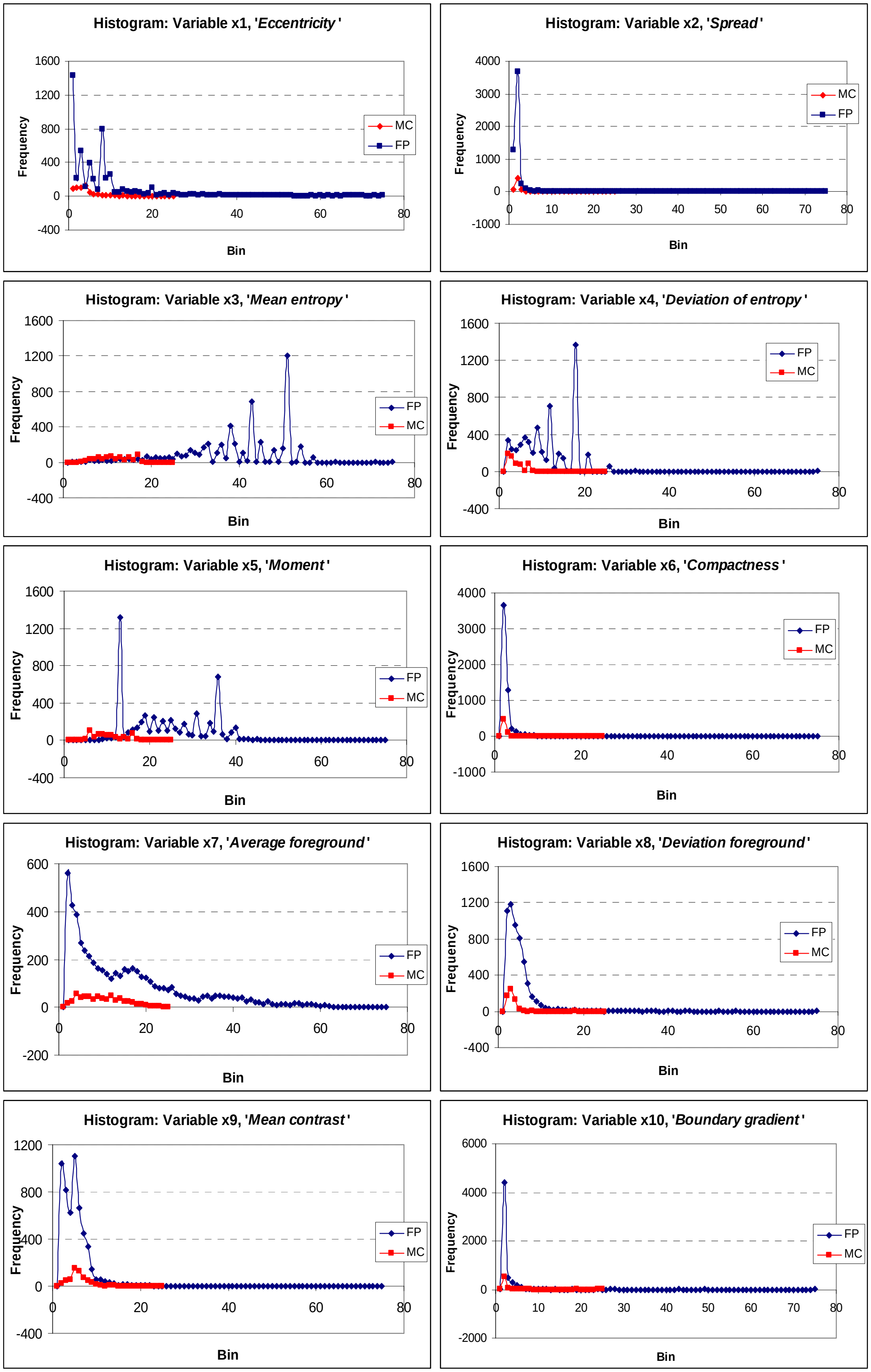


Figure 18 Histograms Of Individual Input Features

The histograms in Figure 18 show that most features are distributed in the same range for both the classes. This makes it impossible to use any one feature to distinguish between the two classes. Also, the distributions are heavily skewed (mostly to the right, in all cases except x3 and x5) i.e. the distribution of values is not symmetrical about the mean. Thus it is very difficult to estimate a "typical value" for the distribution. For instance, in a symmetric distribution, the typical value would be the center of the distribution. Data that is seriously skewed maybe an indication that there are inconsistencies in the process or procedures etc. Further decisions need to be made to determine if the skew is actually appropriate ("Histogram"). Among all the features, x3 (mean entropy) and x5 (moment) have reasonably (though not significant) different range of values.

Table 5 Univariate Statistics Of Input Features For Both Classes

	Feature	**Class FP, n=5500**				
		Mean	SD	Std. error mean	upper 95% mean	lower 95% mean
x1	Eccentricity	0.1213	0.1895	0.0025	0.1263	0.1162
x2	Spread	0.1789	0.1069	0.0014	0.1818	0.1761
x3	Mean entropy	0.1873	0.0576	0.0007	0.1888	0.1858
x4	Dev. Entropy	0.0324	0.0216	0.0003	0.0332	0.0318
x5	Moment	0.0817	0.0319	0.0004	0.0826	0.0809
x6	Compactness	10.19	7.83	0.1057	10.4	9.98
x7	Avg. foreground	-25008.8	7458.7	100.58	-24811.61	-25205.97
x8	Dev. Foreground	26.47	36.5	0.492	27.43	25.5
x9	Mean contrast	1048.12	1046.05	14.1	1075.7	1020.4
x10	Boundary gradient	175.94	467.26	6.3	188.3	163.59

	Feature	**Class MC, n=609**				
		Mean	SD	Std. error mean	upper 95% mean	lower 95% mean
x1	Eccentricity	0.1224	0.1391	0.0056	0.1335	0.1114
x2	Spread	0.1747	0.0475	0.0019	0.1785	0.1709
x3	Mean entropy	0.1569	0.0576	0.0023	0.1615	0.1523
x4	Dev. Entropy	0.0219	0.0195	0.0007	0.0235	0.0204
x5	Moment	0.079	0.0296	0.0012	0.0814	0.0767
x6	Compactness	11.173	9.295	0.3766	11.912	10.433
x7	Avg. foreground	-25460.3	4128.71	167.3	-25131.76	-25788.89
x8	Dev. Foreground	22.22	21.18	0.8586	23.91	20.54
x9	Mean contrast	1266	709.77	28.761	1322.58	1209.61
x10	Boundary gradient	170.5	465.952	18.881	207.584	133.42

Table 5 gives the class-wise statistics. It is observed that the means of all the features have approximately the same values for both the classes. However as observed in the histograms, the data are not symmetrically distributed. Thus studying just the mean holds little significance in this context. Variables x7 and x9 show very high standard deviation from the mean.

It is difficult to visualize this data in 10-dimensional space. Instead, for visual representation of these classes, we plot them in 2-dimensional space. The input data was reduced to two principal components (PCs), which account for 98% of the variance in the input data. The Principal Component Analysis (PCA) was used here to transform the given dataset into a two-dimensional vector which contains all the information contained in the original dataset. Here, PCA was only used to visualize class seperability in a two-dimensional space and not for any other data analysis.

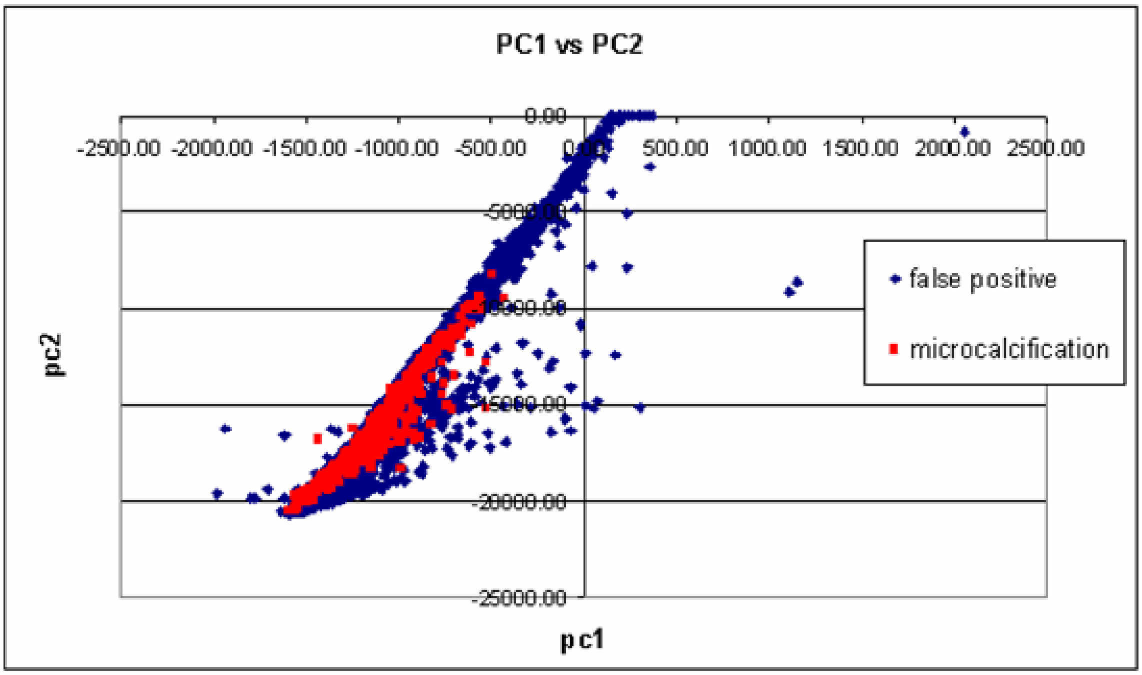


Figure 19 Plot Of PC-1 Vs PC-2

The above scatter plot shows completely overlapping points for both the classes. This indicates that the features are statistically not representative of the classes that they belong to.

The next step is to study the relationships between the input features and the outcome. The main interpretation of logistic regression results is to find the significant predictors of the outcome. A logistic fit of each predictor vs. the outcome was performed.

Table 6 Logistic Fit Of Outcome By Individual Predictors

Variable	Feature	Parameter estimates (β)	Std. Error	Chi-square	Prob > ChiSq
x1	Eccentricity	-0.034	0.229	0.02	0.882
x2	Spread	-0.5759	0.5818	0.98	0.3222
x3	Mean entropy	8.1535	0.6845	141.88	<0.0001
x4	Dev. Of entropy	29.26	2.506	136.35	<0.0001
x5	Moment	2.74	1.37	4	0.0455
x6	Compactness	-0.01	0.004	6.54	0.0105
x7	Avg. Foreground	0.00000902	0.0000061	2.16	0.142
x8	Dev. Foreground	0.0051	0.0018	7.9	0.0049
x9	Mean contrast	-0.0001536	0.0000328	21.97	<0.0001
x10	Boundary gradient	0.0000272	0.0000997	0.07	0.7851

In Table 6, β>0 for variable x3 (mean entropy). Since the coefficient for mean entropy is positive, the log odds (and therefore the probability) of MC increases with mean entropy. On the other hand, the β values for x7 and x10 are close to zero. This would imply that the strength of association for the features 'average foreground' and 'boundary gradient' with the outcome is very poor.

Interpretation of β: The parameter estimate (β) gives the increase in log odds of the outcome, for one unit increase in x i.e. e^{β} represents the change in odds of the outcome, by increasing x by 1 unit. Given below is the interpretation of β:

(a) If β=0, the odds and probability are the same at all x levels ($e^{\beta}=1$)

(b) If β>0, the odds and probability increase as x increases ($e^{\beta}>1$)

(c) If β<0, the odds and probability decrease as x increases ($e^{\beta}<1$)

The overall significance of the variables is tested using the Model Chi-square, which is derived from the likelihood of observing the actual data under the assumption that the model that has been fitted is accurate. The difference in log likelihoods for the model with the predictor and without the predictor is distributed as a chi-square with degrees of freedom equal to the number of predictors. Thus, chi-square tests are used to test if the predictors are significant or not. If we assume a significance level of 0.05, any value of likelihood less than 0.05 would be significant. In Table 6 above*, x3 (mean entropy), x4 (dev. of entropy), x5 (moment), x6 (compactness), x8 (dev. foreground) and x9 (mean contrast)* have values < 0.05 indicating that these features are significant *(individually)* in our model. These features are from both the spatial and morphological domains.

5.2 Feature Selection using SFS

The above univariate analysis gives an interpretation of individual features and their individual relationships with the outcome. The SFS gives the best feature subset as explained in Section 4.5.2. The results of SFS with only the predictors included are as shown in Table 7. This is the *main effect model*.

Table 7 Forward Selection Results For Data: Main Effect Model

Parameter	DF	Estimate	Standard error	Chi-square	Pr > chisq
Intercept	1	1.9213	0.5857	10.7593	0.001
x1	1	-0.5201	0.6946	0.5605	0.4541
x2	1	-15.7646	4.2701	13.6301	0.0002
x3	1	-14.4453	1.1339	162.2841	<.0001
x6	1	0.0462	0.0188	6.0598	0.0138
x7	1	-0.00002	7.59E-06	9.1834	0.0024
x9	1	0.000106	0.000043	6.1915	0.0128

It is observed that the SFS procedure has selected the following features as a good subset:

a) eccentricity
b) spread
c) compactness
d) mean entropy
e) average foreground
f) mean contrast

Features (a) and (b) are margin descriptors; (c) is a morphological feature while features (d) to (f) are spatial domain features. Features in both the spatial and morphological domain have been selected, indicating that these features are significant and important in improving the discriminatory power of the model.

SAS uses the *c*-statistic to determine the discriminating power of the logistic model. The c-statistic is nothing but the Area under the ROC curve, which is close to one for a model that discriminates perfectly. The c-value for the main effect model was 0.693. The ROC curve for the main effect model with the predictors in Table 7 is as shown in Figure 20.

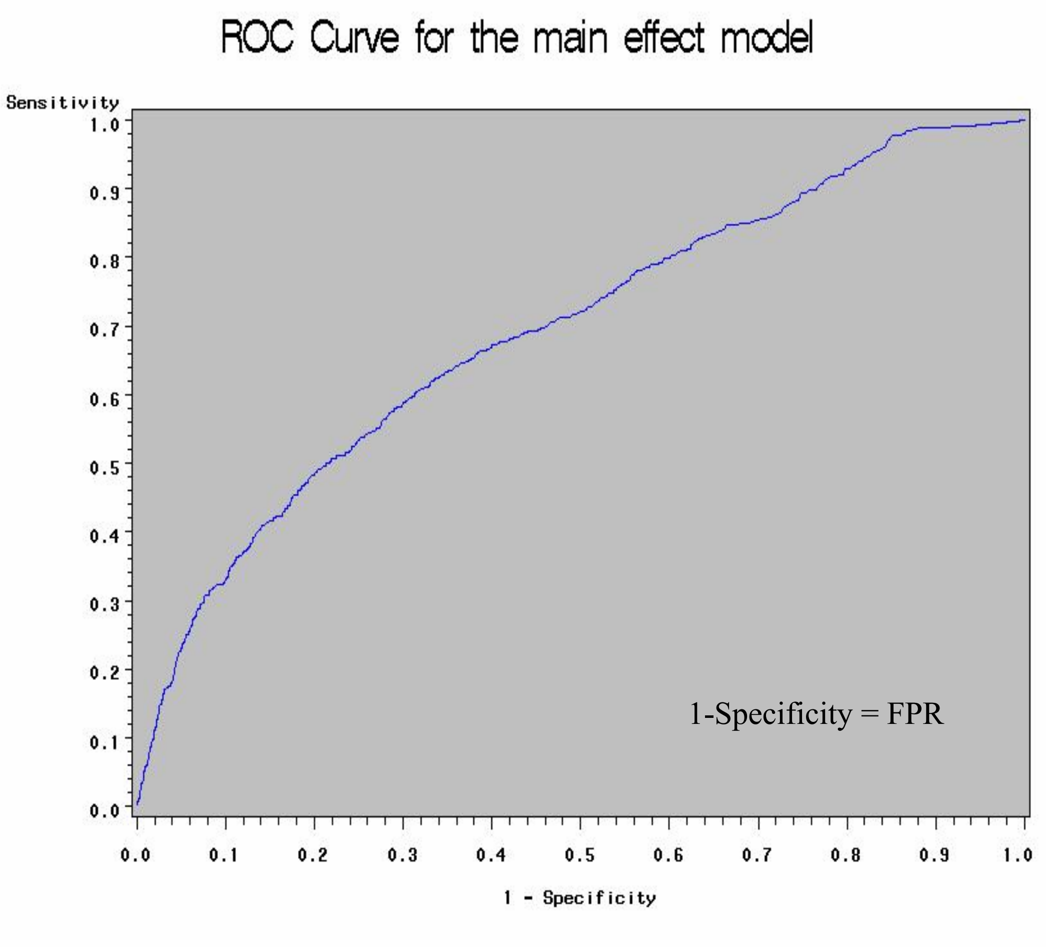


Figure 20 ROC Curve For Main Effect Model, C=0.693

Once the main effect model has been constructed, two-way interactions are studied to assess the predictive effect of two independent variables on the outcome. All possible interactions between the ten input variables were added into the model and a SFS procedure was performed. Results of SFS with all the two-way interactions included into the main effect model are as shown in Table 8.

Table 8 Forward Selection Results For Data: Interaction Effect Model

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	0.7675	0.7844	0.9574	0.3279
x1	1	-0.4961	1.8911	0.0688	0.7931
x3	1	-20.5268	1.8069	129.0604	<.0001
x1*x3	1	16.5311	4.4748	13.6475	0.0002
x7	1	0.000011	0.000028	0.1568	0.6921
x1*x7	1	0.000162	0.000059	7.5173	0.0061
x8	1	0.00385	0.00493	0.6082	0.4355
x9	1	-0.00243	0.000443	30.1192	<.0001
x3*x9	1	0.00991	0.00147	45.7245	<.0001
x7*x9	1	-9.61E-08	1.72E-08	31.4287	<.0001
x8*x9	1	-0.00001	2.71E-06	26.6355	<.0001

This model has chosen the following features and their interactions as the optimal subset:

a) eccentricity

b) mean entropy

c) average foreground

d) dev. foreground

e) mean contrast

f) (eccentricity*mean entropy)

g) (eccentricity*avg. foreground)

h) (mean entropy*mean contrast)

i) (avg. foreground*mean contrast)

j) (dev. foreground*mean contrast)

It is observed here that spatial domain features dominate in significance. Except for eccentricity, the remaining features are all in spatial domain. New indices based on these two-way interactions can be considered to improve the discriminatory power of the features.

The *c*-value of this model is 0.77, which is a significant improvement over that of the main effect model. The ROC curve for the model with the predictors of Table 8 is shown in Figure 21.

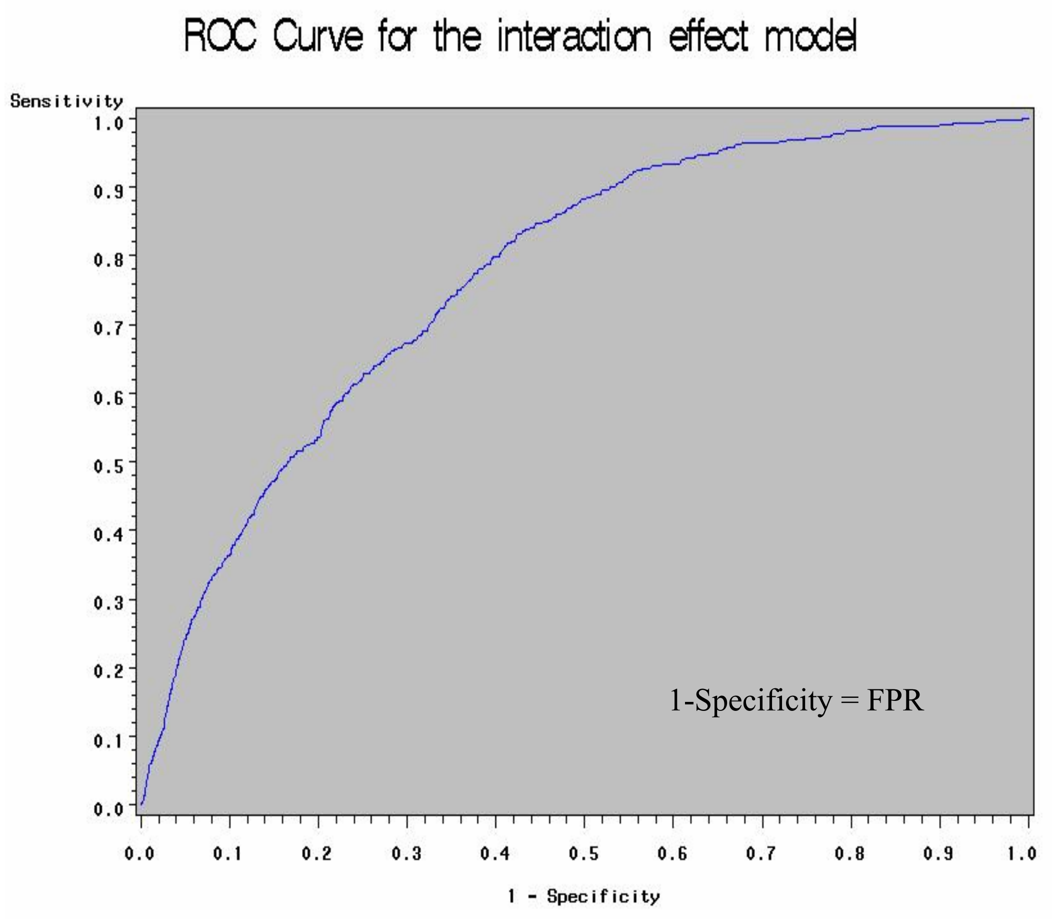


Figure 21 ROC Curve For Interaction Effect Model, C=0.77

This completes the data pre-processing section where we analyzed in detail the feature statistics, the relationships between the predictors and the outcome and performed feature selection (main effect and interaction effects) using SFS.

In summary, it is seen that the features taken individually are not very good predictors of the outcome. Feature statistics and univariate analysis are evidence of this observation. A feature selection procedure helps select the best subset of features that could increase the discriminatory power of our model. Features in both the spatial domain and morphological domain were selected during the feature selection process. However, the SFS procedure used the Logistic regression as the induction algorithm. This may not guarantee the best performance results from NN and SVM. Further classification is performed with all the ten features and the features from the SFS and compared.

5.3 Classification

The parameters used with the NN classifier were presented in the Materials and Methods section. Despite the theoretical advantages that SVMs possess over NNs, SVM requires a certain amount of model selection. The kernel parameter is one of the most important design choices for the SVM since it implicitly defines the structure of the high dimensional feature space where a maximal margin hyperplane will be found. Thus the choice of the SVM kernel is crucial. We study two popular kernels (polynomial and RBF) with various parameters, to see which one best suits our case. These kernels were tested on the 12 training images and on 10 unseen images. The training dataset consisted of 3167 cases, 553 of which were MCs and 2614 cases of FPs.

Table 9 shows the results on training and testing data for various kernels.

Table 9 Choice Of SVM Kernel*

Kernel		Training			Testing		
		Accuracy	Confusion matrix		Accuracy	Confusion matrix	
RBF kernel, g=7	c =1000	0.85	0.12	0.88	0.93	0.16	0.84
(equal weights for classes)			0	1		0.05	0.95
	c =100	0.89	0.39	0.61	0.91	0.25	0.75
			0	1		0.08	0.92
	c =10	0.88	0.35	0.65	0.89	0.34	0.66
			0.01	0.99		0.1	0.9
	c =5	0.88	0.32	0.68	0.9	0.3	0.7
			0.01	0.99		0.09	0.91
RBF kernel	g =9	0.88	0.81	0.19	0.82	0.43	0.57
(with c=50 for class 1 and			0.11	0.89		0.18	0.82
c=10 for class -1)	**g =7**	**0.87**	**0.8**	**0.2**	**0.81**	**0.48**	**0.52**
			0.12	**0.88**		**0.18**	**0.82**
	g =5	0.86	0.76	0.24	0.81	0.45	0.55
			0.12	0.88		0.18	0.82
Polynomial kernel	d =7	0.8	0.5	0.5	0.81	0.54	0.46
(with c=50 for class 1 and			0.14	0.87		0.19	0.82
c=10 for class -1)	d =3	0.8	0.49	0.51	0.81	0.55	0.45
			0.13	0.87		0.19	0.81

** All numerical values rounded to two decimal places*

The confusion matrix interpretation is as follows:

	+1	-1
+1	TPR	FNR
-1	FPR	TNR

It is very important to note the usage of the term ‘False Positive’. In this work, FP is also one of the classes i.e. refers to a false positive MC signal. However the FP that is evaluated in the confusion matrix above refers to (in our case) a FP MC signal misclassified as an MC.

From Table 9, it is observed that sensitivity (TPR) increases drastically by introducing different weights for the classes. This is because our data is highly unbalanced and higher penalty for a positive case would give equal importance to this under-represented class.

It is observed that accuracy decreases as value of *c* (error penalty) decreases. A high value of error penalty would force the SVM training to avoid classification errors, thus resulting in a larger search space for the QP optimizer. It is observed that some experiments fail to converge for very large values of *c* (*c* > 1000). An optimum value of *c=10* is chosen. The performance of the RBF kernel largely depends on the value of *g* which is the radius of the RBF kernel. Accuracy decreases with kernel radius. Though accuracy at *g=9* was the highest with a good sensitivity for training data, the TPR on testing data for *g=7* is higher. It can be seen that the polynomial kernel performance on testing data is better (in terms of sensitivity). The FROC for training and testing datasets for the RBF and polynomial kernel is as shown in Figure 22.

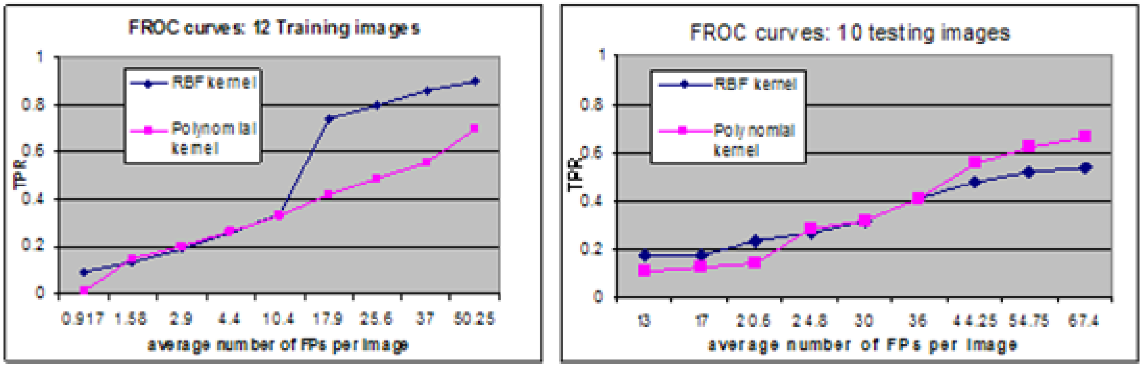


Figure 1 ***FROC Curves For (Left) Training And (Right) Testing Images, C=50 For Class 1 And C=10 For Class -1, RBF Kernel Radius=7; Polynomial Kernel, Degree=3***

Considering the overall performance (on training and testing images), the RBF kernel shows better results than the polynomial kernel. There is a drastic improvement in sensitivity of the RBF kernel for a threshold of 0.65. Sensitivity at this threshold on training images was about 80% while specificity was about 88%. However, the sensitivity of the polynomial kernel on test data was 55% as against 48% for RBF kernel. It is to be noted that the horizontal axis in the above graph gives the average number of FPs/ image and not the FP clusters. The RBF kernel with g=7 was used for further classification.

Once the initial kernel selection was performed, classification and evaluation was done. This step of analysis was broken into experiments which are summarized below:

(a) **Experiment #1**: Ten features with NN and SVM using single test

(b) **Experiment #2**: Ten features with NN and SVM using LOO CV

(c) **Experiment #3**: Ten features with NN and SVM using alternate classes for training

(d) **Experiment #4**: Use features from forward selection results, main effect model, with NN and SVM

(e) **Experiment #5**: Use features from forward selection results, interaction effect model, with NN and SVM

Experiment #1

This experiment used 12 images for training and 10 images for testing. The SVM with RBF kernel (radius=7, error penalty c=50 class 1 and 10 for class -1) was used. The NN used the SBP algorithm with 2 hidden layers, 13 units each. The convergence criterion was set to SSE of 15. All NNs were trained till there was no significant change in the SSE. The performances of the NN and SVM on training and testing images are given in Figures 23 and 24.

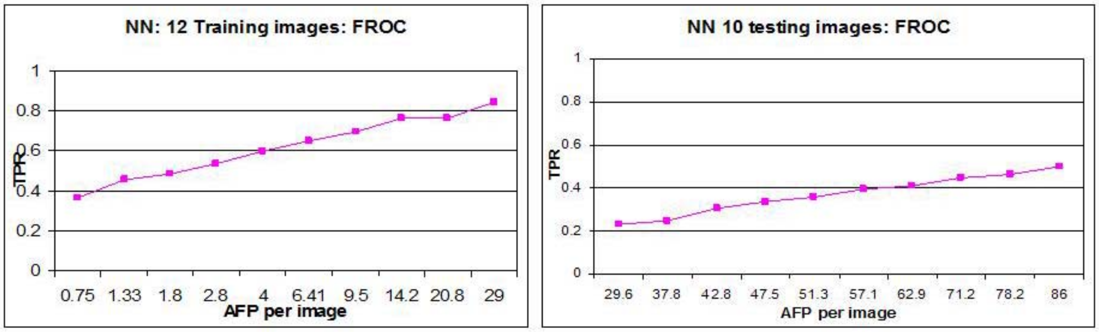


Figure 23 FROC Curve For NN Using All Ten Features- Single Test

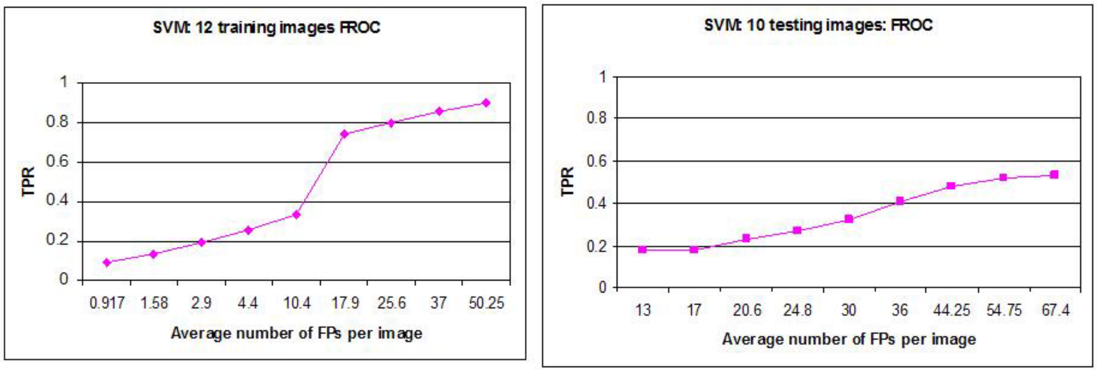


Figure 24 FROC Curve For SVM Using All Ten Features- Single Test

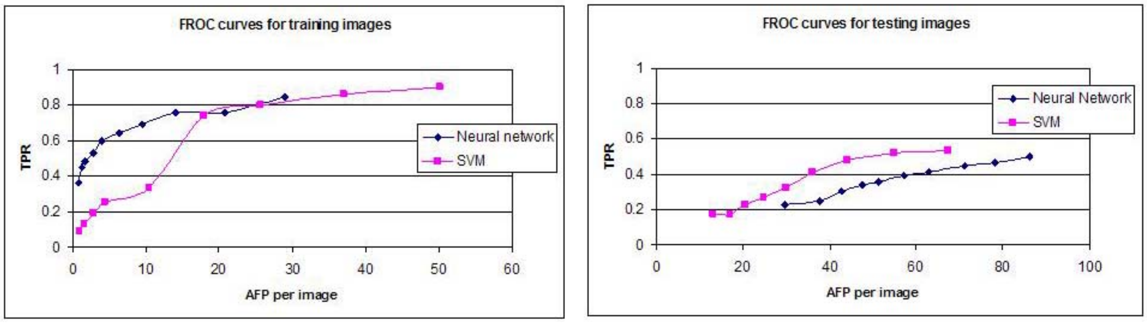


Figure 25 Comparison Of FROC Curves For NN And SVM, (Left) Training Images, (Right) Testing Images

Table 10 Accuracy And Confusion Matrix For Experiment #1

Algorithm	Training			Testing		
	Accuracy	*Confusion Matrix*		*Accuracy*	*Confusion Matrix*	
NN	0.86	0.84	0.16	0.7	0.5	0.5
		0.13	0.87		0.3	0.7
SVM	**0.87**	**0.8**	**0.2**	**0.81**	**0.48**	**0.52**
		0.12	**0.88**		**0.18**	**0.82**

It can be observed from the above graphs that though the SVM shows drastic improvement in sensitivity for a specific threshold (0.7). The accuracy of both the algorithms on training data is comparable. However, the SVM clearly outperforms the NNs performance on testing (unseen) data. The overall accuracy on unseen data is 81% for SVM and 70% for NN with respective specificities of 0.82 and 0.7. Thus, the average number of FPs per image is much lesser for the SVM compared to the NN.

Experiment #2

This experiment used 12 images and performed CV on these images. The 10 testing images were kept completely independent of training dataset to evaluate the classifiers' true generalization capability. The NN was trained using the LOO CV. Here, training is performed by dividing the 12 images into several training and testing sets. In each pass, 11 images are used for training and 1 image for testing. At the end of the process, each image would have been used at least once for testing. LOO is generally used to find the parameters of the classifiers which result in least generalization error. However, with the NN, the model was trained each time with the LOO training datasets.

The SVM on the other hand did not require training with each LOO training set. Here, the LOO CV was performed as a "grid-search" where pairs of (C, γ) are tried and the one with the best CV accuracy is picked. Trying exponentially growing sequences of C and γ is a practical method to identify good parameters (LIBSVM manual) (for instance, C = $2^{-5},2^{-3},....,2^{15}$, $\gamma=2^{-15},2^{-13},...,2^{3}$). Parameter selection was performed using various values of C and γ. The best (C, γ) pair was $(2^7,2^3)$ with the CV rate of 75.56%. Thus c=128, g=8 were used for this experiment.

The performance of these algorithms on training and testing data is as shown in Figures 26, 27 and 28.

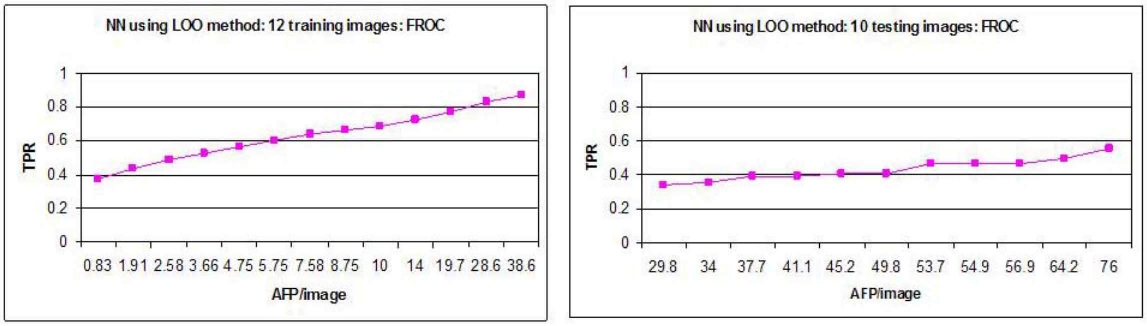


Figure 26 FROC Curve For NN Using All Ten Features- LOO

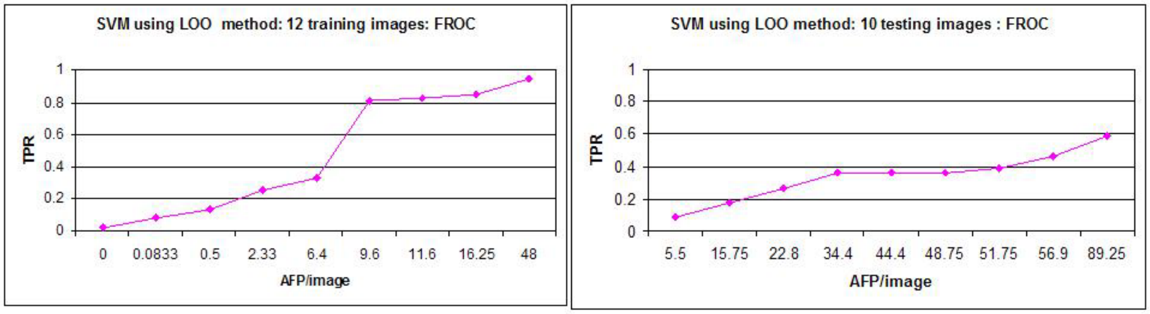


Figure 27 FROC Curve For SVM Using All Ten Features- LOO

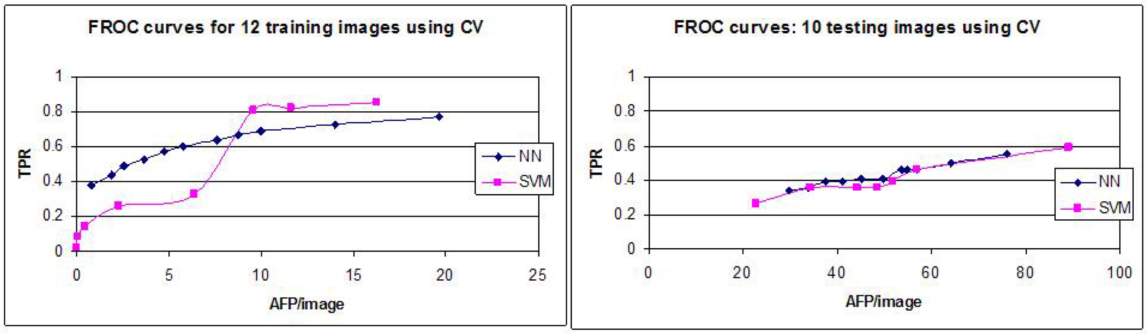


Figure 28 Comparison Of FROC Curves For NN And SVM, (Left) Training Images, (Right) Testing Images

Table 11 Accuracy And Confusion Matrix For Experiment #2

Algorithm	**Training**			**Testing**		
	Accuracy	*Confusion Matrix*		*Accuracy*	*Confusion Matrix*	
NN	0.75	0.47	0.53	0.61	0.38	0.63
		0.19	0.81		0.39	0.61
SVM	**0.93**	**0.81**	**0.19**	**0.76**	**0.46**	**0.54**
		0.04	**0.96**		**0.24**	**0.76**

The SVM outperformed the NN in terms of accuracy and sensitivity on training data. A good value of overall accuracy of the SVM in this experiment shows the importance of performing CV to choose good model parameters. However, the performance of the NN on training data did not improve with this method though generalization performance improved compared to single test method. The performances of both the classifiers on unseen data were comparable.

Experiment #3

The number of MCs in the training set that contained 12 images was 553 and the number of FPs was 2614. Thus the number of FPs exceeded the MCs by almost five times. Initial observations showed that the classifiers were ‘biased’ to the FP class because of the imbalance in the data. Since the dataset is highly unbalanced, it is desired to study the results of using a balanced dataset with equal number of positive and negative cases. 553 FPs were randomly selected from the 2614 cases. The classifiers were presented with a pattern from class MC, and then a pattern from class FP (alternatively). The performance results for three cases is given: (1) the *training set* which contains equal number of MC and FP cases, (2) the *training images* which are the 12 images from which these cases were picked from and (3) *testing images* which are the 10 unseen images. Results of this experiment are summarized below.

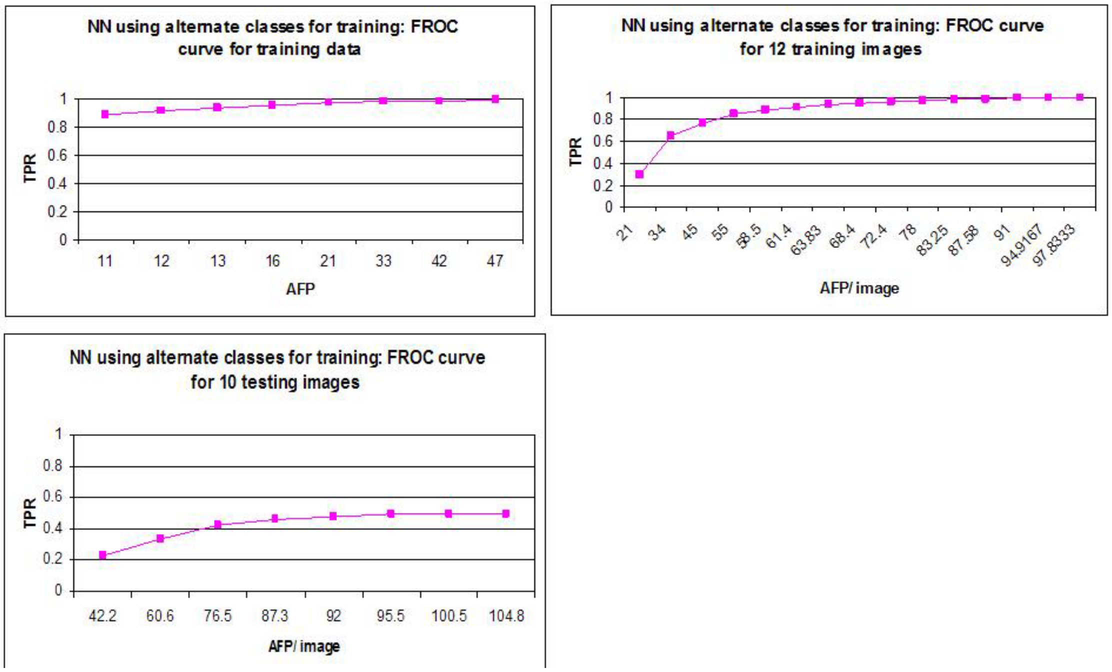


Figure 29   FROC Curve For NN Using All Ten Features- Training With Alternate Classes

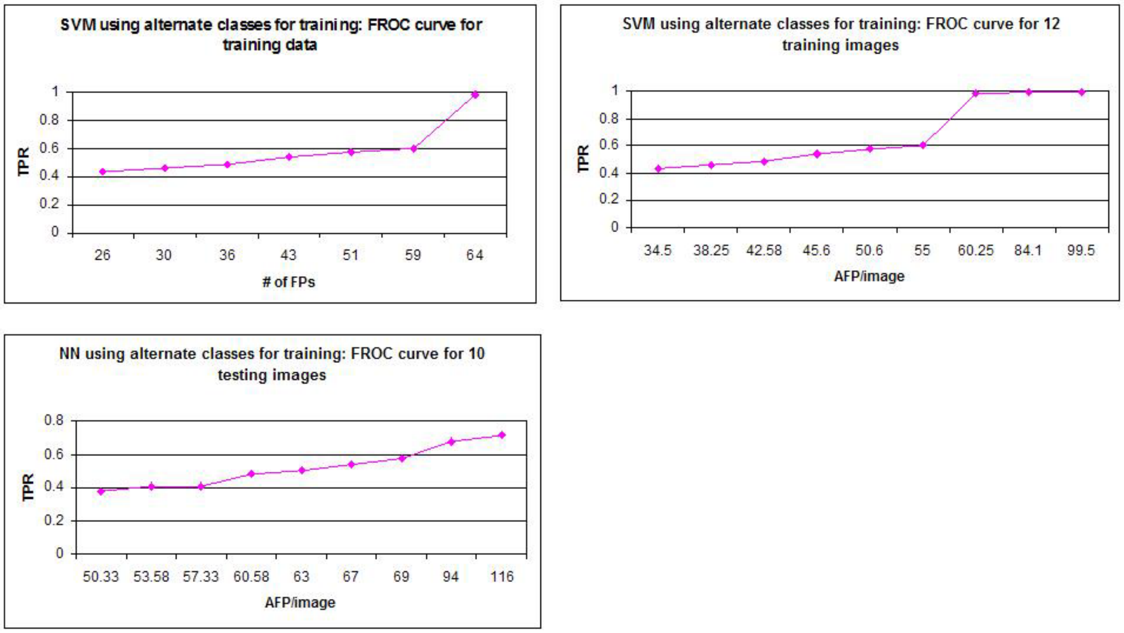


Figure 30 FROC Curve For SVM Using All Ten Features- Training With Alternate Classes

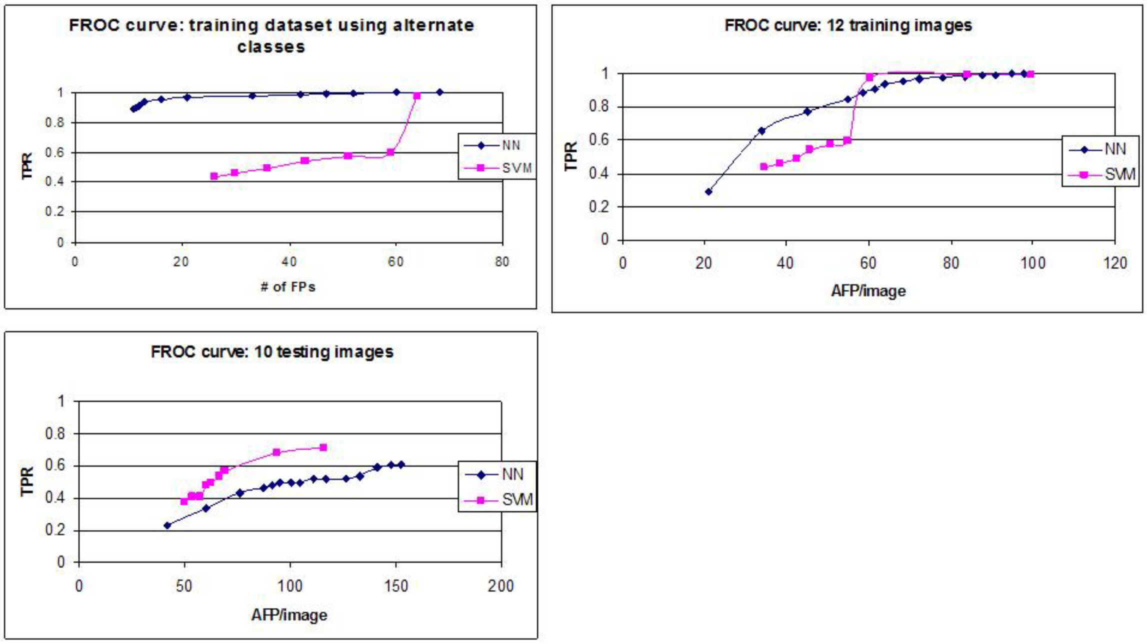


Figure 31 Comparison Of FROC Curves For NN And SVM, (Top Left) Training Dataset, (Top Right) 12 Training Images, (Bottom Left) 10 Testing Images

Table 12 Accuracy And Confusion Matrix For Experiment #3

Algorithm	**Training dataset**			**Training images**			**Testing images**		
	Accuracy	*Confusion Matrix*		*Accuracy*	*Confusion Matrix*		*Accuracy*	*Confusion Matrix*	
NN	0.94	1	0	0.65	0.99	0.01	0.51	0.59	0.41
		0.12	0.88		0.42	0.58		0.49	0.51
SVM	**0.93**	**0.98**	**0.02**	**0.54**	**0.94**	**0.06**	**0.4**	**0.73**	**0.27**
		0.12	**0.88**		**0.55**	**0.45**		**0.61**	**0.39**

From Figure 29, it can be seen that the NN performed extremely well on the training dataset and the training images. The SVM showed sharp increase in sensitivity for the training dataset and images. From the FROC curve on testing images, it is evident that the SVM outperformed the NN in terms of sensitivity. Thus, it is clear that the SVM's capability to generalize is better than the NN. However, it should be noted that this method may not be the most appropriate for training the classifiers since the number of average false positives per image is very high (poor specificity) in spite of high sensitivity. This would mean that a large number of '-1's are being misclassified as '1'. Also the number of FPs in the training dataset is chosen randomly, so the chosen samples may not necessarily be representative samples of the FP class.

The above three experiments are summarized in the FROC graphs of Figure 32.

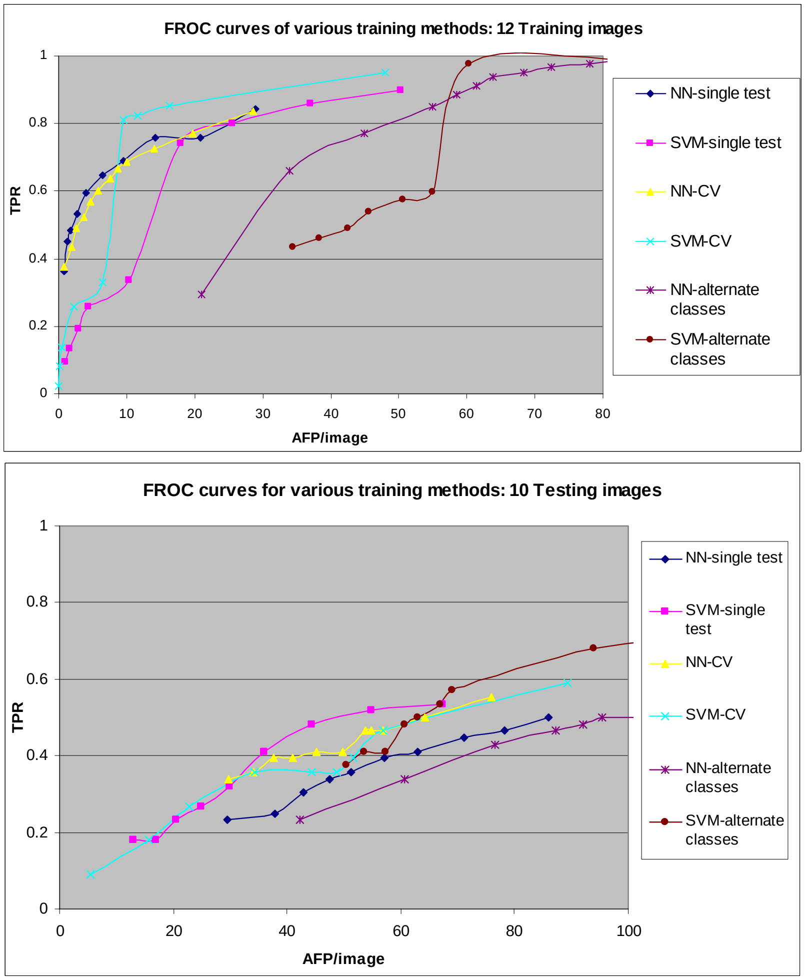


Figure 32 Comparison Of FROC Results From Experiments 1,2 And 3

From the above graphs, we observe that the SVM trained with alternate classes shows the highest sensitivity (98%) for training and testing (73%) images. However, results indicate poor specificity (average number of FPs/ image is high). The SVM with parameters selected from the CV process showed high sensitivity of 81% for training images with low AFP/image value, but showed a sensitivity of about 46% on the testing images. CV improved the performance of NN on unseen data.

The SVM trained with the parameters in experiment 1 had a sensitivity of 80% and specificity of about 88% on training images. On testing images, the sensitivity was 48% and specificity was 81%. This model showed good sensitivity as well as low AFP values on both training and testing images. Overall the SVM outperformed the NN. In Experiments #1 and #3, though the two were comparable on the training data performance, the SVM clearly ruled on the testing (unseen) data.

Experiment #4

This experiment was performed with the feature selection results from the logistic regression. The SFS procedure selected features eccentricity, spread, mean entropy, compactness, average foreground and mean contrast as significant. Only these features were now used in our model to study if there was an improvement in accuracy. Thus the input feature space was six-dimensional. Only the main variables (or effects) were considered first. Training and testing was done using the single test method.

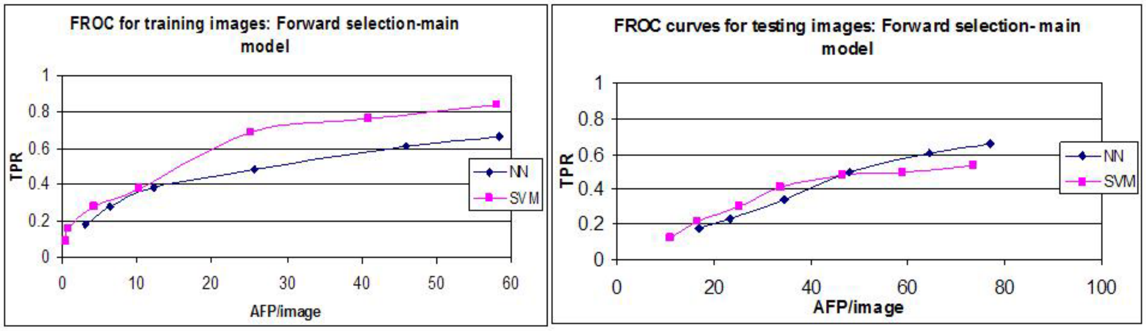


Figure 33 ***FROC Curves For Training And Testing Images Using SFS Main Effect Variables***

Table 13 Accuracy And Confusion Matrix For Experiment #4

Algorithm	**Training**			**Testing**		
	Accuracy	*Confusion Matrix*		*Accuracy*	*Confusion Matrix*	
NN	0.67	0.73	0.27	**0.61**	**0.68**	**0.32**
		0.34	0.66		**0.39**	**0.61**
SVM	**0.85**	**0.69**	**0.31**	0.8	0.48	0.52
		0.12	**0.88**		0.19	0.81

The SVM again outperforms the NN in training and testing performance. Sensitivity of about 69% is seen for specificity of 88% on training images (Table 13). The NN shows slightly better sensitivity (68%) than SVM on testing images in this case. If we consider an AFP value of 50, both the NN and SVM have testing sensitivity of about 50%.

Experiment #5

The interaction effects from the SFS procedure were added into the main effect model. The interactions that were added are discussed on Table 8. New variables were created by multiplying the corresponding variable values.

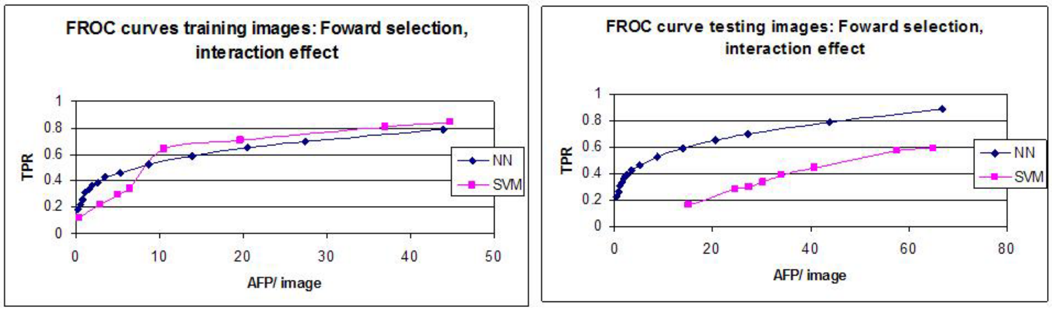


Figure 34 FROC Curves For Training And Testing Images Using SFS Interaction Effect Variables

Table 14 Accuracy And Confusion Matrix For Experiment #5

Algorithm	**Training**			**Testing**		
	Accuracy	*Confusion Matrix*		*Accuracy*	*Confusion Matrix*	
NN	0.73	0.89	0.11	**0.83**	**0.81**	**0.19**
		0.31	0.69		**0.17**	**0.83**
SVM	**0.84**	**0.7**	**0.3**	0.76	0.57	0.43
		0.13	**0.87**		0.24	0.76

It is seen that the testing performance of the SVM is very poor for this case. The NN showed much higher sensitivity on testing data (81%) compared to the SVM (57%) (Table 14).

We now see how these models with feature selection have performed compared to using all the ten features in the model. Figure 35 summarizes this comparison.

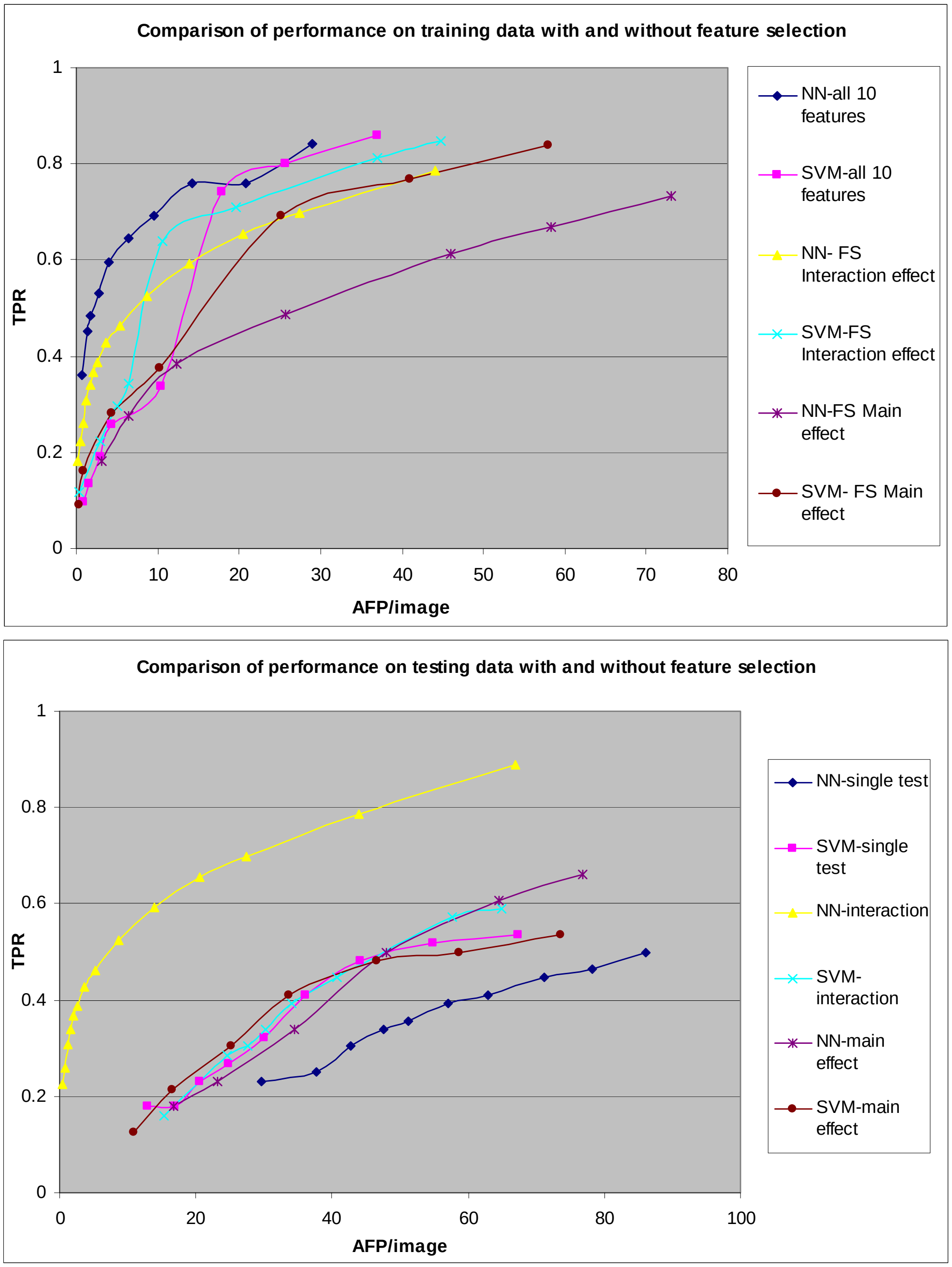


Figure 35 Comparison Of FROC Graphs For Models With And Without Feature Selection

The NN and SVM that used all the features showed a sensitivity of about 78% for an AFP value of 17 on the training images. However, the sensitivity on the testing images was around 33% and 47% respectively for an AFP value of 40 per image. Model performance on training images did not improve with feature selection. However, with feature selection, the generalization capability of the NN classifier increased significantly. For an AFP value of 40 per image, the sensitivity is about 78% for interaction effect variables and 48% for main effect variables, which is a significant improvement over the 33% sensitivity obtained from the NN without feature selection. However, the same feature selection variables did not show any improvement in the generalization performance of the SVM classifier. Performance on training data was worse with feature selection than with all the ten features included.

On unseen data, the NN with feature selection showed great improvement in performance. A reasonable explanation for this would be to go back to the basis of NN algorithms. The feature selection procedure employed used Logistic Regression as the induction algorithm. A logistic regression model is identical to a NN with no hidden units if the logistic (sigmoidal) activation function is used (Bishop, 1995; Hastie T., 2001). In a NN with hidden units, each hidden unit computes a logistic regression (different for each hidden unit) and the output is therefore a weighted sum of logistic regression outputs. The weights (of the NN) or the coefficients (of Logistic Regression) are determined based on the dataset, by maximum likelihood estimation (Dreiseitl Stephan, 2003). However, the decision boundary for a NN can be non-linear, making the NN more flexible compared to logistic regression (Dreiseitl Stephan, 2003). Better results of NN with FS which used logistic regression as the induction algorithm could be attributed to this similarity in mathematical principle.

Figure 36 shows an example of an output image obtained with the SVM using all ten features.

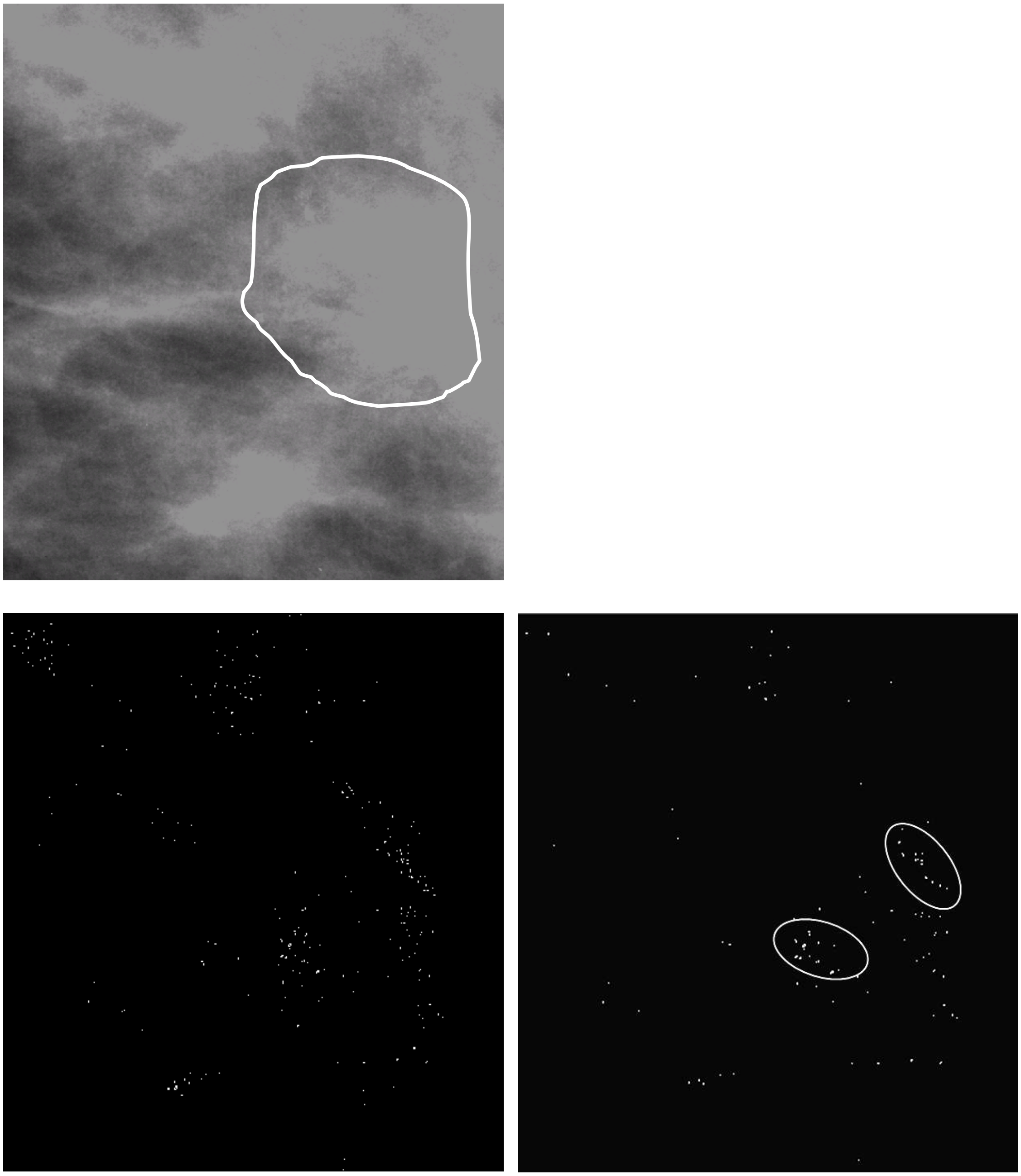


***Figure 36** (Top) Raw Image With Suspicious ROI Outlined, (Bottom Left) Segmented Image, (Bottom Right) Image Showing MC Clusters With Reduced FPs*

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Concluding Remarks

In this work, we presented the use of SVM and NN algorithms for detection of MCs in mammograms. The classifiers were trained through these techniques, to test on every location in the segmented mammogram whether the detected signal was an MC or an FP. Ten features were originally used to represent the two classes.

Experimental results were obtained using a database of 22 images. A detailed statistical analysis of the dataset was performed prior to classification. It was observed that based on statistics alone, it was difficult to characterize these classes. However, the SVM and NN algorithms, considered to model highly non-linear data, do show interesting results.

The classifiers were trained using different training methods like single test, cross validation and alternate class training. The LOO CV was used with the SVM to perform parameter selection. Accuracy improved from about 87% to 93% on training data for the SVM with parameter search. However, performance on the testing set did not improve significantly. The single test SVM showed good results overall (training and testing). CV improved the performance of the NN on unseen images. With the alternate class training method, the classifiers showed high sensitivities of about 95% and 65% (average for NN and SVM) on training and testing data respectively. Though the sensitivity was high the average number of FPs per image was also high. Also, this method chose random cases of the FP class which may not be the ideally representative samples. Overall, in all the experiments that used all the ten features, the SVM outperformed the NN. Though the algorithms were comparable with results on training set, the SVM performed better on unseen data. The SVM with CV parameter selection showed the best performance with much lesser number of FPs per image.

Feature selection using Stepwise Forward Selection method with logistic regression as the induction algorithm was performed. The most significant features were selected and given to the classifiers. For the SVM, though the models with feature selection showed lesser accuracy on the training data than the models that used all features, the testing sensitivities were comparable. Thus, the models with feature selection achieved the same generalization performance as those without feature selection. This helps us remove irrelevant and redundant features and achieve comparable testing performance with fewer features. In particular, the sensitivity of the NN model on unseen data with interaction effects added was extremely high (around 78% for an AFP/image value of 40) as compared to 33% for the NN model without these interaction terms. The NN with main effect model terms showed a sensitivity of around 42%. The improvement of the NN in accuracy and sensitivity on unseen data can be linked to its mathematical similarity with logistic regression which was used as the inductor during the feature selection process. New variables that incorporate the interactions between significant features could be added into the analysis to improve the discriminatory as well as generalization power of the classifiers.

In summary, the SVM outperformed the NN in almost all cases. The generalization capability of the SVM was clearly noteworthy. The training time taken by SVM was also several magnitudes lesser. Thus we can say that for this complex dataset, the SVM is more suited for our analysis.

6.2 Future Work

The most crucial part of future work would be to cluster the output to enhance clinical utility. This work only involves detection of MC spots and reduction of FP signals. However, these spots are considered suspicious when seen in clusters of four, five or six (Faculty of Medicine, 1999). Clusters have to be identified individually as shown in Figure 36 for each image and threshold.

Feature selection showed an improvement in generalization performance of the NN. However, this was not the case with SVM. Wrappers that use the SVM as the induction algorithm could be used to select features. However, all the methods based on the wrapper approach are tuned for/ by a given learning machine. The filter approach to feature selection could be a better alternative here, since it would provide a generic selection of variables not specific to any learning algorithm. A study based on the comparison of all these feature selection approaches would be worthwhile.

Also a direct medical understanding of the features' effect on the class would make analysis of the results easier. Feature selection could be performed just based on domain knowledge. Future work could include using a bigger database with more representative cases. More number of images and training samples would help establish our results and observations. Comparison with other techniques like decision trees and statistical classifiers could be performed.

REFERENCES

Histogram.from http://www.sytsma.com/tqmtools/hist.html

Interactive mammography analysis web tutorial. (1999). from http://sprojects.mmi.mcgill.ca/mammography/anat.htm

4woman.gov. (March 2002). Mammograms. from http://www.4woman.gov/faq/mammography.htm

Anand. (1999). The backpropagation algorithm. from http://www.speech.sri.com/people/anand/771/html/node37.html

Anttinen I, P. M., Soiva M, Roiha M. (1993). Double reading of mammography screening films: One radiologist or two? *Clin. Radiol., 48*, 414-421.

Arce GR, F. R. (1989). Detail-preserving ranked-order based filters for image processing. *IEEE Trans Acoust., Speech, Signal processing, 37*(1), 83-98.

Association, B. B. (Dec.2002). Computer-aided detection (CAD) in mammography. *Assessment Program*, from http://bcbs.com/tec/vol17/17_17.html

Bamberger RH, S. M. (1992). A filter bank for the directional decomposition of images: Theory and design. *IEEE Trans Signal Processing, 40*, 882-893.

Bauer PH, Q. W. (1991). A 3-D non-linear recursive digital filter for video image processing. *IEEE Pacific Rim Conf. on Commun., Comput., and Signal Processing, 2*, 494-497.

Bishop, C. (1995). *Neural networks for pattern recognition*: Oxford: Oxford University Press.

Brien, D. O. Image entropy. from http://www.astro.cornell.edu/research/projects/compression/entropy.html

Burbidge R, T. M., Buxton B, Holden S. (2001). Drug design by machine learning: Support vector machines for pharmaceutical data analysis. *Comput. Chem., 26*, 5-14.

Castleman, K. (1979). *Digital image processing*: Prentice Hall, Englewood Cliffs, Reading, MA.Center, S. C. (2004). Breast cancer statistics. from http://cancer.stanfordhospital.com/healthInfo/cancerTypes/breast/

Chan HC, D. K., Vyborny CJ, et al. (1990). Improvement in radiologists' detection of clustered microcalcifications on mammograms: The potential of computer-aided diagnosis. *Invest. Radiol., 25*, 1102-1110.

Chan HP, D. K., Galhotra S, Vyborny CJ, MacMahon H, Jokich PM. (1987). Image feature analysis and computer aided analysis in digital radiography: Automated detection of microcalcifications in mammography. *Med Phys, 14*, 538-548.

Chapelle O, H. P., Vapnik VN. (1999). Support vector machines for histogram-based image classification. *IEEE Trans Neural Networks, 10*, 1055-1064.

Cheng HD, L. Y., Freimanis RI. (1998). A novel approach to microcalcification detection using fuzzy logic technique. *IEEE Trans Medical Imaging, 17*(3), 442-450.

Chih-Chung Chang, C.-J. L. (2001). LIBSVM: A library for support vector machines.

Chun-Nan Hsu, H.-J. H., Dietrich S. (Apr 2002). The ANNIGMA- wrapper approach to fast feature selection for neural nets. *IEEE Trans Systems, Man and Cybernetics, 32*(2), 207-212.

Cox, D.R. and Snell, E.J. (1989), *The Analysis of Binary Data*, Second Edition, London: Chapman and Hall.

Cristianini N, S.-T. J. (2000). *An Introduction to Support Vector Machines*: Cambridge University Press.

Daphne Koller, M. S. (1996). *Toward optimal feature selection.* Paper presented at the International Conference on Machine Learning.

Dash M, L. H. (1997). Feature selection for classification. *Intelligent Data Analysis- An International Journal, 1*(3).

Davies DH, D. D. (1992). The automatic computer detection of subtle calcifications in radiographically dense breasts. *Phys. Med. Biol., 37*, 1385-1390.

Ding CH, D. I. (2001). Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics, 17*, 349-358.

Dreiseitl Stephan, O.-M. L. (Feb 2003). Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics, 35*(5-6), 352-359.

Edwards DC, K. M., Nagel R, Nishikawa RM and Papaioannou J. (2000). Using a bayesian neural network to optimally eliminate false- positive microcalcification detections in a CAD scheme. *Proceedings of International Workshop on Digital Mammography, at press*.

Faculty of Medicine, M. (1999). Tutorial 2: Calcifications. from http://sprojects.mmi.mcgill.ca/mammography/calcifications.htm

Gader PD, K. J., Krishnapuram R, Chiang JH, Mohamed MA. (1997). Neural and fuzzy methods in handwriting recognition. *Computer, 30*, 79-86.

Gavrielides, M. (1996). *Shape analysis of mammographic calcification clusters.* Unpublished Masters thesis, University of South Florida, Tampa.

Glatt A, L. H., Arnow T, Shelton D, Ravdin P. (1992). *An application of weighted majority minimum range filters in the detection and sizing of tumors in mammograms.* Paper presented at the SPIE Medical imaging VI: Image processing.

Graps, A. (2004). An introduction to wavelets. from http://www.amara.com/IEEEwave/IEEEwavelet.html

Hagan MT, D. H., Beale MH. (1996). Neural network design. *Boston, Mass: PWS.*

Harris JR, H. S., Hendersen IC, Kinne DW. (1991). *Breast diseases*: JB Lippincott Company, Philadelphia, PA.

Hastie T., T. R., Friedman J. (2001). *The elements of statistical learning: Data mining, inference and prediction*: New York: Springer; 2001.

Hendee WR, B. C., Hendrick E. (1999). Proposition: All mammograms should be double-read. *Med Phys, 26*, 115-118.

Hiep Van Khuu, H.-K. L., Jeng-Liang Tsai. (2003) *Machine learning with neural networks and support vector machines* (Unpublished).

Hosmer, D.W, Jr. and Lemeshow, S. (1989), *Applied Logistic Regression*, New York: John Wiley & Sons, Inc.

Huai Li, R. L. K., Shih- Chung B. Lo. (1997). Fractal modeling and segmentation for the enhancement of microcalcifications in digital mammograms. *IEEE Trans Medical Imaging, 16*(6), 785-798.

Imaginis. (Sept.2004). Breast cancer diagnosis.

Imaginis. (Sept.2004). Breast cancer screening/ prevention.

Imaginis. (Sept.2004). General information on breast cancer. from http://imaginis.com/breasthealth/statistics.asp

Imaginis. (Sept.2004). Tutorial 2: Calcifications.

Imaginis. (Sept.2004). Tutorial: Masses.

Isabelle Guyon, A. E. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157-1182.

Issam El-Naqa, Y. Y., Miles Wernick, Nikolas Galatsanos, Robert Nishikawa. (Dec 2002). A support vector machine approach for detection of microcalcifications. *IEEE Trans Medical Imaging, 21*(12), 1552-1563.

Joachims, T. (1999). Making large-scale SVM learning practical. In C. B. B. Schölkopf, A. Smola (Ed.), *Advances in kernel methods - support vector learning*: MIT Press.

John G, K. R., Pfleger K. (1994). *Irrelevant features and the subset selection problem.* Paper presented at the Machine learning: Proceedings of the Eleventh International Conference, San Francisco, CA.

Karssemeijer, N. (1993). *Recognition of clustered microcalcifications using a random field model, biomedical image processing and biomedical visualization.* Paper presented at the SPIE Proc., San Jose, CA.

Keerthi, S., Shevade, S., Bhattacharyya, C., & Murthy, K. (1999). *Improvements to Platt's SMO algorithm for SVM classifier design.* Paper presented at the Proceedings of the Tenth European Conference on Machine Learning.

Kerlikowske K, G. D., Rubin SM, et al. (1995). Efficacy of screening mammography: A meta-analysis. *JAMA, 273*, 149-154.

Ko SJ, L. Y. (Sept. 1991). Center weighted median filters and their applications to image enhancement. *IEEE Trans Circ. Syst., 38*, 984-993.

Kohavi, P. (1988). Glossary of terms. *Machine Learning, 30*(2-3), 271-274.

Kohavi R, J. G. (Dec 1997). Wrappers for feature selection. *Artificial Intelligence, 97*(1-2), 273-324.

Kohavi, S. (1995). *Feature selection using the wrapper model: Overfitting and dynamic search space topology.* Paper presented at the First International Conference on Knowledge Discovery and Data Mining.

Kong, H. (1998). *Self-organizing tree map and its application in digital image processing.* Unpublished Ph.D. thesis, Univ. of Sydney, Sydney, Australia.

Kregelmeyer WP, P. J., Bourland PD, Hillis A, Riggs MW, Nipper ML. (1994). Computer-aided mammographic screening for spiculated lesions. *Radiology, 191*, 331-337.

Kubat M, H. R., Matwin S. (1998). Detection of oil spills in satellite radar images of sea surface. *Machine Learning, 30*, 195-215.

Lai SM, L. X., Bischof WF. (1989). On techniques for detecting circumscribed masses in mammograms. *IEEE Trans Medical Imaging, 8*(4), 377-386.

Lanyi, M. (1988). *Breast calcifications*: Springer-Verlag, Berlin, Heidelberg.

Lecun Y, J. L., Bottou L, Brunot A, Cortes C, Denker JS, Drucker H, Guyoin I, Muller A, Sackinger E, Simard P and Vapnik V. (1995). *Comparison of learning algorithms for hand written digit recognition.* Paper presented at the ICANN '95.

Lei Yu, H. L. (2003). *Feature selection for high-dimensional data: A fast correlation-based filter solution.* Paper presented at the Proceedings of the Twentieth International Conference on Machine Learning (ICML- 2003), Washington DC.

Liang H, L. Z. (2001). Detection of delayed gastric emptying from electrogastrograms with support vector machine. *IEEE Trans Biomed Eng, 48*, 601-604.

M. Pontil, A. V. (1998). Support vector machines for 3d object recognition. *IEEE Trans Pattern Analysis Machine Intelligence, 20*, 637-646.

Mckenna RJ, S. (1994). The abnormal mammogram radiographic findings, diagnostic options, pathology, and stage of cancer diagnosis. *Cancer (Suppl 1), 74*, 244-255.

MedCalc. (2004). Logistic regression. from http://www.medcalc.be/manual/logistic_regression.php

N.Karssemeijer. (July 1991). *A stochastic model for automated detection of calcifications in digital mammograms.* Paper presented at the 12th Int. Conf. Information Processing Medical Imaging, Wye., U.K.

Nagel Rufus H, N. R. M., Papaioannou John, Doi Kunio. (Aug 1998). Analysis of methods for reducing false positives in the automated detection of clustered microcalcifications in mammograms. *Med Phys, 25*(8), 1502-1506.

Nishikawa RM, D. K., Geiger ML, et al. (1995). Computerized detection of clustered microcalcifications: Evaluation of performance on mammograms from multiple centers. *RadioGraphics, 15*, 445-452.

Osuna E, F. R., Girosi F. (1997). *Training support vector machines: An application to face detection.* Paper presented at the Computer Vision and Pattern Recognition.

Park, D. (2000). Centroid neural network for unsupervised competitive learning. *IEEE Trans. Neural Networks, 11*, 520-528.

Platt, J. C. (1999). Fast training of support vector machines using sequential minimal optimization. In C. B. B. Scholkopf, A. J. Smola (Ed.), *Advances in kernel methods- support vector learning* (pp. 185-208): MIT Press.

Pontil M., V. A. (1998). Object recognition with support vector machines. *IEEE Trans. On Pattern Analysis and Machine Intelligence, 20*, 637-646.

Popli, M. (2001). Pictorial essay: Mammographic features of breast cancer, *Ind J Radiol Imag* (Vol. 11, pp. 175-179).

Qian W, C. L. (July 23-28,1995). *Hybrid m-channel wavelet transform method and application to medical image processing.* Paper presented at the 37th annual meeting of the American Association of Physicists in Medicine, Boston, Mass.

Qian W, C. L., Kallergi M, Clark RA. (1994). Tree-structured nonlinear filters in digital mammography. *IEEE Trans Medical Imaging, 13*, 25-36.

Qian W, L. L., Clarke LP. (1999). Feature extraction for mass detection using digital mammography: Influence of wavelet analysis. *Med Phys, 26*, 402-408.

Qian W, S. D., Sun Xuejun, Clark Robert A. (June 2001). *Multistage statistical order using neural network for false positive reduction in full field digital mammography.* Paper presented at the 2001 IEEE- EURASIP Workshop on Nonlinear Signal and Image Processing, Baltimore, Maryland, USA.

Quinlan, J. (1993). *C4.5: Programs for machine learning*: Morgan Kaufmann.

Rajkumar S, H. L. (Nov. 1999). Screening mammography in women aged 40-49 years. *Medicine, 78(6)*, 415.

Rufus H. Nagel, R. M. N., John Papaioannou, Kunio Doi. (1995). Analysis of methods for reducing false positives in the automated detection of clustered microcalcifications in mammograms. *Med Phys, 25*, 1502-1506.

Shen L, R. R., Leo Desautels JE. (1994). Application of shape analysis to mammographic calcifications. *IEEE Trans Medical Imaging, 13*, 263-274.

Society, A. C. (2004). Cancer prevention & early detection, facts & figures. from http://www.cancer.org/downloads/STT/CPED2004PWSecured.pdf

Songyang Yu, L. G. (2000). A CAD system for the automatic detection of clustered microcalcifications in digitized mammogram films. *IEEE Trans Medical Imaging, 19*(2), 115-126.

Statsoft Inc. (1984-2003). Neural networks. from http://www.statsoft.com/textbook/multilayerb

Stetson PF, S. F., Macovski A. (Aug 1997). Lesion contrast enhancement in medical ultrasound imaging. *IEEE Trans Medical Imaging, 16*, 416-425.

Struble, C. A. *Dimension reduction and feature selection* (Presentation): Dept. of Mathematics, Statistics and Computer Science, Marquette University.

Systems, G. M. (2003). Digital mammography. from http://www.hersource.com/breast/02/a-mammo.cfm

Takehiro Ema, K. D., Robert M. Nishikawa, Yulei Jiang, John Papaioannou. (February 1995). Image feature analysis and computer-aided diagnosis in mammography: Reduction of false-positive clustered microcalcifications using local edge-gradient analysis. *Med Phys, 22*(2), 161-169.

te Brake GM, K. N., Hendriks JH. (1998). Automated detection of breast carcinomas not detected in a screening program. *Radiology, 207*, 465-471.

Ted C. Wang, N. B. K. (Aug 1998). Detection of microcalcifications in digital mammograms using wavelets. *IEEE Trans Medical Imaging, 17*(4).

Tembey, M. (2003). *Computer Aided Diagnosis for mammographic microcalcification clusters.* Unpublished Masters thesis, University of South Florida, Tampa.

Thurfjell E.L., L. K. A., Taube A.A.S. (1994). Benefit of independent double reading in a population-based mammography screening program. *Radiology, 191*, 241-244.

Vapnik, V. (1995). *The nature of statistical learning theory*: Springer Verlag.

Vapnik, V. (1998). *Statistical learning theory*: John Wiley.

Vyborny, C. J. (1994). Can computers help radiologists read mammograms? *Radiology, 191*, 315-317.

W. Morrow, R. P., R. Rangayyan, J. Desautels. (June 1992). Region based contrast enhancement of mammograms. *IEEE Trans Medical Imaging, 11*, 392-406.

Warren Burhenne LJ, W. S., D'Orsi CJ, et al. (2000). The potential contribution of computer-aided detection to the sensitivity of screening mammography. *Radiology, 215*, 554-562.

Woods KS, D. C., Bowyer KW, Solka JL, Priebe CE and Kegelmeyer WP Jr. (1993). Comparative evaluation of pattern recognition techniques for detection of microcalcifications in mammography. *Int. J. Pattern Recog. Artificial Intell., 7*, 1417-1436.

Wouter J, V. H., Karssemeijer Nico. (Nov.2000). Automated classification of clustered microcalcifications into malignant and benign types. *Med Phys, 27*(11), 2600-2608.

Yoshida H, D. K., Nishikawa RM. (1994). *Automatic detection of clustered microcalcifications in digital mammograms using wavelet transform techniques.* Paper presented at the SPIE.

Zhang W, D. K., Giger ML, Nishikawa RM, Schmidt RA. (1996). An improved shift-invariant artificial neural network for computerized detection of clustered microcalcifications in digital mammograms. *Med. Phys., 23*, 595-601.

Zheng B, Q. W., Clarke LP. (June 1994). *Artificial neural network for pattern recognition in mammography.* Paper presented at the Proc. World Congress Neural Networks, San Diego, CA.