DISCRETE COSINE TRANSFORM BASED FEATURE EXTRACTION FOR COMPUTER AIDED DETECTION OF SUSPICIOUS X-RAY MAMMOGRAM IMAGES

MATTHEW T. FLYNN

Discrete Cosine Transform Based Feature Extraction for Computer Aided Detection of Suspicious X-Ray Mammogram Images

by

© Matthew T. Flynn
B.Sc. Honours

A thesis submitted to the
School of Graduate Studies
in partial fulfillment of the
requirements for the degree of
Master of Science in Medicine.

Division of BioMedical Sciences
Faculty of Medicine
Memorial University of Newfoundland
St. John's, Newfoundland Labrador

February 28, 2011

Abstract

One of the best ways to decrease breast cancer mortality is through early detection. X-ray mammography is widely used to screen women with an increased risk of breast cancer. Computer aided detection (CAD) programs have been developed in an effort to boost efficiency and accuracy, but studies have shown that the CAD programs currently in use are not particularly effective.

In this project, a new CAD algorithm was developed. The two main components of the method were the use of whole image classification and a novel feature extraction step using the discrete cosine transform. The features were generated from moments of the mean of square sections centered on the origin of the transform. Feature vectors were then run through k-nearest neighbour and naive Bayesian classifiers.

It was found that the discrete cosine transform could be used to manually filter suspicious characteristics from images. Features extracted from the images were found to change dramatically when a mass was introduced into the image. Using a k-nearest neighbour classifier, sensitivities as high as 98% with a specificity of 66% was achieved. With a naive Bayesian classifier, sensitivities as high as 100% were achieved with a specificity of 64%.

Acknowledgements

I would like to convey my deepest gratitude to my supervisor, Dr. Edward Kendall for his advice and insight. Without his guidance, this project would not have been possible.

Thanks as well to the members of my supervisory committee, Dr. Jules Doré and Dr. Nancy Wadden, and to my examiners, Dr. Mark Eramian and Dr. Michael Noseworthy. Their suggestions and careful attention to detail in their reviews of this thesis have been invaluable.

Finally, I would like to thank my parents, Patrick and Molly and my sister Sarah for their constant support and encouragement.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The first chapter of this thesis consists of background information related to the mammogram computer aided detection (CAD) system developed in this project. Section 1.1 states the problem being investigated and the general approach taken to solve it. Section 1.2 discusses basic breast anatomy and cancer. Section 1.3 explains how mammography works, how images are interpreted, and how computers can be used to aid in the detection of cancer. Section 1.4 provides some mathematical background on the discrete cosine transform and the moments about the mean and what they represent. Section 1.5 introduces the two classifiers used, k-nearest neighbour and naive Bayesian, as well as the cross validation process of training and testing them.

1.1 Objective and approach

The object of this work is to develop and test a computer aided detection method for x-ray mammography with a high level of sensitivity and specificity. Whole image classification will be used in an effort to avoid the n false positives per image problem from which many CAD programs suffer. The novelty of this approach is in the choice of feature vectors. Discrete cosine transforms will be used in conjunction with a unique sampling method to generate feature vectors that can be used to distinguish normal from abnormal images.

1.2 Breast cancer

Breast cancer is a serious problem for women in Canada. It is the most common form of cancer diagnosed in women, with one in nine women expected to be diagnosed with some form of breast cancer in their lifetime. It is second only to lung cancer in cancer related deaths. With increasing rates of breast cancer diagnosis, mortality has been in decline since the mid 1980s [1, 2].

The decrease in breast cancer associated mortality may be attributable to several factors falling under the broad categories of treatment and detection. Both have seen significant advances over the past 30 years, due to the vast amount of research in the field. The most important prognostic factor in breast cancer is the stage at which it is diagnosed. Cancers caught when they are still in the early stages have a much higher probability of being cured than cancers that have metastasized. Screening programs are widely used to detect potential breast cancers. In any screening process, the objective is not to obtain an accurate diagnosis, but rather to separate the individuals who may possibly have cancer from those who do not from within a large population. In order for a screening program to be effective, it must have a high throughput and must offer reasonably accurate results.

Breast cancer screening involves mammography and occasionally clinical breast examination. Many women perform breast self exams, however there is a growing body of evidence that suggests that there is no benefit to this [3, 4]. Clinical breast exams also suffer from low accuracy since many cancers may not be large or close enough to the surface to be detected. MRI is more sensitive than x-ray [5], but is expensive and time consuming and therefore not well suited to a screening program. X-ray mammography allows the detection of masses that are not yet palpable, as well as calcifications that may indicate malignancy. This can be done with good accuracy, high throughput, and relatively low cost, making it ideal as a screening method.

An understanding of breast anatomy is essential to the successful interpretation of x-ray mammograms. In women, the breasts develop during puberty under the influence of estrogen. Male breasts generally develop some subareolar ducts during adolescence, but seldom form lobules [6], partially explaining why breast cancer is so rare in men. Figure 1.1 shows the most important structures found in the female breast that can be seen on a mammogram.

Each breast contains a complex mammary gland composed of many simple mammary glands. The glandular tissue of the breast is made up of ducts and lobes. Each breast contains a median of 23 lactiferous ducts which drain to the nipple [8]. The ducts are connected to lobes made up of many branching lobules which contain the milk secreting exocrine cells. The remaining stromal tissue consists mostly of connective tissue (Cooper's ligaments) and fat (subcutaneous and retromammary adipose).

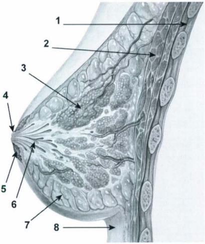


Figure 1.1: Breast anatomy: 1. Chest wall, 2. pectoralis major, 3. lobules, 4. nipple, 5. areola, 6. ducts, 7. fatty tissue, 8. skin. Image from [7].

Blood is supplied to the breasts through the internal thoracic, lateral thoracic, posterior intercostal, subscapular, and thoracodorsal arteries and drains through the axillary, intercostal, and internal thoracic veins [9]. Blood vessels are often visible in mammograms. They are sometimes calcified, but this is a benign finding.

Breasts are bilateral and are generally located between the clavicle and eighth rib, and between the sternum and the midaxillary line, superficial to the pectoralis major. When imaging the breast with mammography, care must be taken to reduce the impact of the neighbouring chest wall and to include tissue around the lateral margin of the pectoralis major muscle [6].

The term breast cancer refers to a large group of cancers originating in the breast with many different histopathological types. In mammography, the primary concerns are ductal and lobular carcinomas. Not only are these the most common forms of breast cancer, they occur deep in the glandular tissue of the breast, where physical changes are generally difficult to detect through visual inspection or palpation in the early stages. 55% of breast cancer diagnoses are for invasive ductal carcinoma, 13% are ductal carcinoma in-situ and 5% are invasive lobular carcinomas [10].

In addition to the histopathology of breast cancers, they are further subdivided according to the presence or absence of three receptors. Estrogen receptor positive cells grow in the presence of estrogen and may be treated with drugs that either reduce estrogen concentrations or antagonize estrogen receptors such as Tamoxifen [11]. Progesterone receptors block transcription unless progesterone is present. Progesterone receptor positive cancers grow faster in the presence of progesterone [11]. Human epidermal growth factor receptor 2 (HER2) is a receptor tyrosine kinase influencing cell growth and differentiation. When it is over expressed, it can lead to increased cell proliferation. HER2 positive cancers are often treated with the drug Herceptin [12]. While mammography is able to detect the presence of such cancers, biopsies must be performed to actually characterize them.

The two most important risk factors of developing breast cancer are female sex and age. Women are roughly 100 times more likely to develop breast cancer than men. As women age, the chances of developing breast cancer steadily increase. Annual risk of breast cancer increases with each year of age (table 1.1).

There are a number of factors that can lead to a greater risk of developing breast

Table 1.1: Risk of breast cancer per year by age decade [13]. As a woman ages, her chances of developing breast cancer each year increases.

Age Range	Annual Risk	Annual Risk (%)
20-29	0.05/1000	0.005
30-39	0.4/1000	0.044
40-49	1.5/1000	0.146
50-59	2.7/1000	0.273
60-69	3.8/1000	0.382

cancer. Most notably, a personal or a family history of breast or ovarian cancer generally indicates an increased chance of developing breast cancer [14]. Several genes have been identified that make an individual much more susceptible to breast cancer. In particular, BRCA1 and BRCA2, two tumor suppression genes have been shown to greatly increase the risk of breast cancer in women carrying a mutation in either of these genes. Lifetime risk of developing breast cancer for women carrying BRCA mutations is 82 % [15]. Women with a high risk of breast cancer may have additional testing outside of normal screening mammography. This may include more frequent mammograms, starting at a younger age in addition to MRI and ultrasound exams.

Breast cancer is staged with the TNM (tumor, node, mass) system. T, the size of the primary tumor, is graded as T0 (no tumor), TX (evidence of a tumor, but cannot be found), Tis (tumor is *in situ*), or T1-T4, with T4 being the largest. Lymph node involvement is graded as NX (nodes cannot be measured), or N0-N3. Finally, M0 indicates the cancer has not metastasized, M1 means it has, and MX means it is unknown if it has metastasized or not [16]. These factors are combined to determine a stage for the cancer. Stage 0 indicates a non invasive cancer in which the tumor is highly localized and there is no evidence of it invading nearby tissue. Stage I is an invasive cancer of up to two cm in diameter without lymph node involvement. Stage II indicates lymph node involvement and/or tumor size of up to five cm. Stage III indicates that the cancer has invaded further but has not yet metastasized. Finally stage IV indicates the cancer has metastasized to distant organs, most commonly bone, liver, brain, or lung [17]. Five year survival of these stages for ductal and lobular carcinomas are 92% for stage 0, 87% for stage I, 75% for stage II, 46% for stage III, and 13% for stage IV [18]. The earlier these cancers are discovered, the higher the chance of survival. This is the rational for using mammographic screening.

1.3 Mammography

X-ray mammography has long been the subject of fierce debate as to its efficacy. Different organizations have different recommendations for screening intervals and these recommendations periodically change. Recommendations generally vary from once a year starting in the late 30s, to not at all. Proponents of frequent and early mammograms claim that this is the only way to catch cancers while they are still curable and that the only reason not to do this is economic. While opponents question the value of mammography at all, citing radiation exposure, false positives, and unnecessary interventions in cases that might never become symptomatic. While the validity of either of these arguments is not the subject of this research, the Canadian Breast Cancer Foundation recommends regular screening between the ages of 50 and 69 [19].

The most important consideration in determining the best possible screening interval for mammography is the mean sojourn time that a cancer will remain curable [6]. This is the amount of time expected to be between the point when a cancer becomes large enough to be detected with mammography and when it is too advanced to be cured. In order for screening to be effective it must be repeated at intervals shorter than the mean sojourn time. Unfortunately, this varies widely. Breast cancers usually grow faster in women younger than 50. This is why when mammography is recommended for individuals under age 50, it may be at yearly intervals, where as individuals over 50 may be once every two years [20].

A mammography unit consists of three vital components. The x-ray tube generates low energy x-rays that are filtered according to the density and thickness of the breast to provide x-ray photons in the region of the spectrum that will provide good contrast in the resulting image. Secondly, there must be plates, which are transparent to x-rays, to compress the breast and keep it still during imaging. Thirdly, there must be a detector to capture the image. It can either be a film, or a digital device.

X-rays are a form of electromagnetic radiation with wavelengths in the range of 10^{-8}m - 10^{-11}m. They can be characterized as high energy (hard) or low energy (soft) x-rays. In mammography, softer x-rays are used because their limited penetration results in higher contrast in soft tissue. At the appropriate wavelength range, photons are absorbed and scattered by some breast tissues and not by others.

The x-ray tube is a vacuum tube with and cathode at one end and a rotating anode at the other. Electrons are liberated from the cathode through thermionic emission and accelerated across a potential difference towards the anode [21]. At the anode, the electrons generate x-ray photons through bremsstrahlung and x-ray fluorescence. Bremsstrahlung (braking radiation) occurs when an electron passes close to an atomic nucleus. The electric field decelerates the electron and the lost kinetic energy is converted to a photon as shown in figure 1.2 [22]. Bremsstrahlung forms the continuous spectrum, indicated in figure 1.4. X-ray fluorescence occurs when a high energy photon knocks an electron from an inner orbital of an atom in the anode. The hole created in the inner orbital is quickly filled by an electron from an outer orbital and a photon is released as it moves to the lower energy state as shown in figure 1.3 [21]. X-ray fluorescence is a quantum process and creates a discrete emission spectrum characteristic of the anode material indicated in figure 1.4.

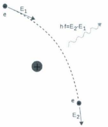


Figure 1.2: As the kinetic energy of an electron is lost through interaction with an atomic nucleus, an X-ray photon is emitted. This is called Bremsstrahlung. Image from [23].

The anode is angled so that the emitted radiation is directed towards a window. Filters can be placed in the window to absorb radiation that is not in the desired range, to reduce exposure. Denser and/or thicker breasts require more energetic x-rays to penetrate the breast.

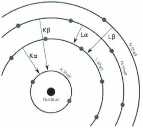


Figure 1.3: When an electron is lost from an inner orbital, higher energy electrons transition down to fill the hole and emit a photon with the energy difference. This is called fluorescence. Image from [24].

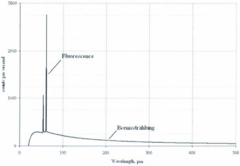


Figure 1.4: A typical X-ray spectrum. Bremsstrahlung causes the broad smooth part of the spectrum, whereas fluorescence causes sharp peaks.

X-ray photons interact with the breast tissue primarily in two ways. In Compton scattering, some of the photon energy is absorbed by an electron, causing it to recoil, and the rest is released as a degraded photon at an angle $\theta = arccos\left(1 - \frac{m_e c(\lambda' - \lambda)}{h}\right)$ to the original photon [25]. Since the photons are scattered at an angle to the incident photon, Compton scattering can cause blurring of the image. In the photoelectric effect the photon is completely absorbed by an electron causing it to be ejected from the atom [26]. This contributes to the sharpness of the image, since an absorbed photon will not reach the target. While both processes play a role in photon attenuation, the photoelectric effect dominates.

In order to obtain the best images possible, the breasts must be compressed during the imaging process. Typically the breasts are compressed with a force of 10-20 lbs (44-89 N) up to 45 lbs (200 N)[6]. Not only does this hold the breast still and out from the chest wall while the image is being taken, but also increases the surface area and decreases the thickness. Since x-ray mammography produces projection images, increased surface area results in less overlap of tissues in the breast making it easier to determine where one structure ends and another begins. The decrease in effective thickness of the breast also means lower beam energies can be used.

There are two main types of mammographic x-ray detectors currently in use; film, and digital. In film based machines, the image is captured on a photosensitive film and developed so they can be viewed in light boxes. These images may subsequently be scanned into a computer for more convenient storage and analysis. Digital mammogram machines are more modern and expensive. They capture the image with a CCD sensor and store images directly on a computer. Digital mammograms use, on average, 22% less radiation [27], and have greater contrast. Digital mammograms have been shown to significantly increase screening accuracy in women with dense breasts and those under age 50 [28].

In a typical mammographic exam, two images are taken for each breast. The mediolateral oblique (MLO) view is through the side of the breast angled parallel to the pectoralis major, so that breast tissue extending lateral to the muscle and into the axilla can be viewed without obstruction [6]. Craniocaudal (CC) views are straight down through the top of the breast. The two views are taken because some features are visible in one view and not another. It also provides the radiologist with a way of visualizing the 3D structure of the breast from the projected images, and localizing any abnormalities found.

The accuracy of the test can be broken up into sensitivity and specificity. In order to define these two terms, we must first define four more. A true positive is when a test correctly identifies a positive result. A false positive is when a test incorrectly returns a positive result. A true negative is when a test correctly identifies a negative result. A false negative is when a test incorrectly returns a negative result. Sensitivity is a measure of how well a test can detect positives; that is, individuals who test positive for some condition. It is defined as the number of true positives divided by the actual number of positives $\frac{tp}{tp+fn}$. Specificity is a measure of how well a test can detect individuals who are negative for some condition. It is defined as the number of true negatives divided by the total number of actually negative individuals $\frac{tn}{tn+fp}$. In a perfect test, sensitivity and specificity will both be 100%. Less than 100% sensitivity can mean that some individuals will have delayed treatment and therefore poorer prognosis. Less than 100% specificity can mean that some individuals will be given unnecessary testing and/or treatment. While neither of these situations are ideal, in the case of mammography, more importance is placed on high sensitivity than high specificity, as low sensitivity can directly lead to increased mortality.

1.3.1 Mammogram interpretation

No matter the skill of the radiologist, there will always be cancers missed that, in retrospect, were visible on the mammogram. The accuracy rate varies widely. In a study of 209 radiologists between January 1, 1995 and December 31, 2000, each reading an average of 6011 mammograms, had a mean sensitivity of 77% with a range of 29% to 97%. Specificity ranged from 71% to 99% with an average of 90% [29]. Higher specificity has been shown to correlate with more experience. There is a significant increase in specificity with more than 25 years of experience vs less than 10, and interpreting more than 2500 mammograms per year vs less than 750 [29]. Positive predictive value has been measured at 34% higher for radiologists reading over 2000 mammograms per year compared to those reading between 480 and 699 [30].

Mammograms are typically read by one or two radiologists. There are many different strategies used for double readings. Anywhere from all images to only a very small portion may be double read. The second reader may or may not be aware to the outcome of the first reading. Action may be taken based on a consensus between radiologists, or on the recommendation of either radiologist alone. Double readings have been shown to significantly increase accuracy [31, 32, 33]. The problem with double readings is that they can also significantly increase costs which may, in some cases, make mammography less accessible [6].

Mammography is an excellent tool for detecting abnormalities. 80-85% of cancers can be seen with x-ray [34]. It is not, however, used in making a definitive diagnosis.

Specificity for distinguishing between benign and malignant lesions is typically around 60%[34].

Mammograms are generally arranged for viewing by placing left and right images next to each other so that they appear as mirror images. This allows the radiologist to easily assess for asymmetries. If previous images are available they will usually be placed above the current ones so that the radiologist can assess for any changes in the breasts since the last screen. Then the radiologist can search for masses, calcifications, or architectural distortions that may indicate cancer.

Masses can be categorized according to their shape and margins. Shapes can be round, oval, lobulated, or irregular. Their margins can be circumscribed, obscured, micro-lobulated, ill-defined, or spiculated, as shown in figure 1.5. They can be very obvious or quite subtle, as shown in figure 1.6.

Architectural distortions are caused by desmoplastic reactions (abnormal growth of fibrous or connective tissue). While they are sometimes the result of harmless process such as scar tissue, they are often a sign of malignancy [38]. They have the appearance of tissue being pulled in toward a central point as seen in figure 1.7.

Most breasts have at least one calcification. Calcifications can be formed from cellular secretions or necrosis. They can be found in breast tissue or on the skin. There are several different types of calcifications, some are usually benign and some indicate malignancy. The BIRADS groups calcifications into the categories: 'typically benign', 'intermediate concern', 'higher probability of malignancy'. Size, shape, and distribution can provide clues as to the origin of a calcification. In general, large,

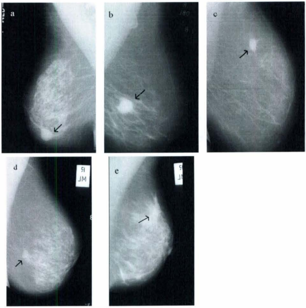


Figure 1.5: Cancerous masses can have many different appearances. A sample of five types of cancerous masses is shown here. a) Circumscribed, b) obscured, c) micro-lobulated, d) ill-defined, and e) spiculated. Images from [35].

rounded calcifications, like lucent calcifications, are benign and small irregular shaped clusters, like pleomorphic or fine linear calcifications, are malignant. Figure 1.8 shows some different types of calcifications.

The first mammogram a woman receives provides a frame of reference that can be compared with subsequent mammograms. By comparing current images with those of the patient from two or more years ago, it is easier to appreciate any changes that have

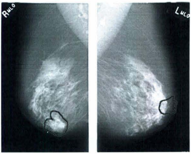


Figure 1.6: Cancerous masses are sometimes easy to spot, but often are not. The first image shows a very obvious mass, the second shows a subtle one [36, 37].

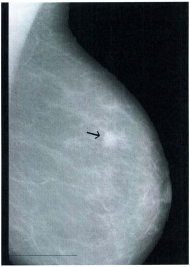


Figure 1.7: An architectural distortion is shown in this image. This is sometimes an indicator of malignancy [35].

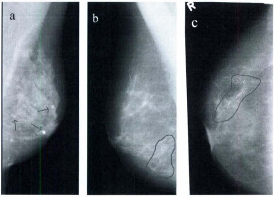


Figure 1.8: Calcifications can appear in numerous forms as the result of different processes. Certain types of classifications are strongly correlated with cancer whereas others are not. A sample of breast calcifications: a) Lucent, b) pleomorphic, and c) fine linear [35].

occurred in the breast in the intervening time. Generally, as time goes on, breasts become less radiopaque. Any changes should be carefully assessed to determine if they are suspicious. Using older scans makes it possible to see if any areas of tissue are growing and potentially cancerous; this has been shown to significantly increase reading accuracy [39].

1.3.2 Computer aided detection

Computer aided detection is a tool a radiologist can use to assist in the detection of abnormal pathology in medical images. It is often used either to replace or augment double readings. Computers are able to rapidly analyze the large amount of data in a medical image and draw the radiologist's attention to any images/regions that may require attention. They are able to do this by utilizing machine learning and pattern recognition to match an image being tested to known patterns of disease.

One of the most common applications of CAD is in mammography. In order for any computer aided detection method to be used on a mammogram, it must be in a digital format, whether a scanned film or digital mammogram. Nearly all CAD programs do region of interest (ROI) analysis, where images are divided into sections and these sections are tested for abnormalities. When abnormalities are detected, they are marked for the radiologist's review. The problem with this method is that many more objects are marked than there are suspicious areas [40]. Specificities are often cited as a number of false positives per case. Having such a high number of areas that require a radiologist's attention may slow the process of reading a mammogram in addition to biasing the radiologist to unnecessarily call back more patients than they ordinarily would.

A common problem with CAD is a slight increase in sensitivity, but significant decrease in specificity. Research by Joshua Fenton has shown a significant decrease in specificity from 90.2% to 87.2% with use of CAD programs and no significant increase in sensitivity [41]. There is a wide variation in the reported performance of CAD software. The actual performance of a CAD program rests largely on the radiologist using the program. It is also important not to make the assumption that the CAD programs reviewed are representative of all CAD programs. David Gur et al. showed a statistically significant difference in the accuracy of two commercially available CAD systems [40].

Michael Barnett developed a method of whole image classification using the discrete wavelet transform [42]. Using this method, he was able to achieve near perfect sensitivity with greater than 60% specificity. Whole image classification could potentially decrease the physician bias seen in ROI analysis. When a whole image is flagged for further review, the radiologist must still asses the image and determine for themselves if and why it needs review.

1.4 Mathematics

1.4.1 Discrete cosine transform

The cosine transform is a modification of the Fourier transform, in which only the cosine term is considered. It decomposes a wave into its component cosine waves and plots the amplitudes of these waves in frequency space. Unlike the Fourier transform, the discrete cosine transform (DCT) is real valued, and hence outputs only magnitude information and not phase.

The DCT is used when there are a finite number of data points in a set. There are eight variants of the discrete cosine transform. For the purpose of this work, only the four most common (I, II, III, and IV) were used (equation 1.1 [43]).

$$DCT\ I : \nu_s = \sqrt{\frac{2}{n-1}} \left(\frac{u_1}{2} + \sum_{r=2}^{n-1} u_r cos \left(\frac{\pi(r-1)(s-1)}{n-1} \right) + (-1)^{s-1} \frac{u_n}{2} \right) \tag{1.1}$$

$$DCT\ II : \nu_s = \frac{1}{\sqrt{n}} \sum_{r=1}^{n} u_r cos \left(\frac{\pi}{n} \left(r - \frac{1}{2} \right) (s-1) \right), \ s = 1, 2, ..., n \tag{1.2}$$

$$DCT\ III : \nu_s = \frac{1}{\sqrt{n}} \left(u_1 + 2 \sum_{r=2}^{n} u_r cos \left(\frac{\pi}{n} (r-1) \left(s - \frac{1}{2} \right) \right) \right) \tag{1.3}$$

$$DCT\ IV : \nu_s \;=\; \sqrt{\frac{2}{n}}\left(\sum_{r=1}^{n} u_r cos\left(\frac{\pi}{n}\left(r-\frac{1}{2}\right)\left(s-\frac{1}{2}\right)\right)\right) \tag{1.4}$$

DCT II is the most commonly used variant. Unless otherwise specified, the term DCT refers to DCT II. The values of ν_s are coefficients of a cosine basis function and are plotted in frequency space. Each wave in the basis function is orthogonal, forming a linearly independent set [44]. The first coefficient in the DCT II transform ($s = 1$) will always be $1/\sqrt{n}\ \sum_{r=1}^{n} u_r$. It is an average of sorts and is called the DC coefficient. The other coefficients ($s > 1$) are called the AC coefficients [45, 44]. In the transform, the DC coefficient represents a constant value, where as the AC coefficients correspond to the waves in the basis function.

The DCT can easily be extended into higher dimensions. The 2D-DCT II transform is defined in equation 1.5 [43]. This allows us to take the transform of a two dimensional array of data, such as an image. The basis functions in the two dimensional case can be derived simply by superimposing horizontally and vertically oriented basis functions on a grid. In the two dimensional transform, the coefficients are mapped to 2D frequency space where each point corresponds to one of the 2D basis functions.

$$\nu_{s_1,s_2} = \sum_{r_1=0}^{n_1-1}\sum_{r_2=0}^{n_2-1} u_{r_1,r_2}\cos\left[\frac{\pi}{n_1}\left(r_1+\frac{1}{2}\right)s_1\right]\cos\left[\frac{\pi}{n_2}\left(r_2+\frac{1}{2}\right)s_2\right] \tag{1.5}$$

A simple way of calculating the 2D DCT is to do one dimensional transforms for each row, then repeat for each column. In practice, since equation 1.5 is symmetric and separable, matrix multiplication can be used to rapidly calculate the transformation. The transformation matrix of an $N \times N$ image can be calculated from equation

1.6. Then the transform is simply $T = AfA$ [46, 44].

$$A_{ij} = \begin{cases} \frac{1}{\sqrt{N}} & j = 0 \\ \sqrt{\frac{2}{N}} \cos \frac{\pi(2j+1)i}{2N} & j \neq 0 \end{cases} \tag{1.6}$$

The DCT separates the components of an image according to frequency. High frequency characteristics, such as small points with sharp edges are mapped to the outer regions of the transform, while low frequency features such as large objects with smooth edges are mapped to the inner portion of the transform. The DCT therefore separates the different types of objects in an image.

The DCT exhibits a high degree of energy compaction. This means that a large proportion of the information stored in a signal is represented in the lower frequency coefficients of the transform [44]. It is this property that makes DCT a popular choice for data compression, and is a motivating factor in the feature extraction method described in chapter 2.

Like the Fourier transform, the cosine transform is a lossless operation. The transform contains the same amount of data as the original signal and for each transform there exists an inverse transform that can be used to reconstruct the original signal. The inverse of DCT I is itself, the inverse of DCT II is DCT III and vice versa, the inverse of DCT IV is itself [43, 44]. That is, two consecutive DCT I or DCT IV transforms result in the original signal, and DCT II followed by DCT III or vice versa results in the original signal. As with the transform, the inverse transform can also be calculated with matrix multiplication using the equation $f = A^{-1}TA^{-1}$, where A^{-1} denotes the inverse matrix of the transformation matrix. Since A is orthogonal, its inverse is simply its transpose.

1.4.2 Central moments

There are four statistical measures used in feature generation in this work: mean, variance, skewness, and kurtosis. They are derived from the central moments about the mean. The central moments about the mean are calculated from equation 1.7 [47]. Where μ_n is the n^{th} central moment, $\langle x \rangle$ is the expectation value, $\mu (= \mu'_1)$ is the mean, and $P(x)$ is the probability density function of x.

$$\begin{aligned} \mu_n &= \langle (x - \langle x \rangle)^n \rangle && (1.7) \\ &= \int_{-\infty}^{\infty} (x - \mu)^n P(x) dx && (1.8) \end{aligned}$$

Mean is simply the total of a list of numbers divided by the number of items in the list (equation 1.9).

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{1.9}$$

The second moment μ_2 is called variance. The positive square root of variance is the standard deviation, σ (equation 1.10). Standard deviation measures how much a point in the data set can be expected to deviate from the mean. Low standard deviation indicates that data is clustered about the mean, while high standard deviation means the opposite.

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2} \tag{1.10}$$

Skewness is defined as the third *standardized* central moment. The n^{th} standardized central moment is simply the n^{th} central moment divided by the standard deviation to the power of n (equation 1.11) [47]. It measures the symmetry of a distribution. A symmetric distribution will have a skewness of zero. Distributions skewed to the left of the mean will have negative skewness and distributions skewed to the right will have positive skewness as shown in figure 1.9.

$$S = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^3}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2)^{3/2}} \tag{1.11}$$

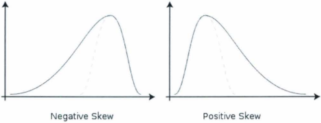


Figure 1.9: Examples of curves with positive and negative skewness. The distribution on the left has a longer tail on the left of the mean so it has a negative skewness. The distribution on the right has a longer tail on the right of the mean so it has a positive skewness. Image from [48].

Kurtosis is defined as the fourth standardized central moment (equation 1.12) [47]. Kurtosis is the measure of how much the standard deviation is due to infrequent large deviations and how much is due to small frequent deviations. Distributions with high kurtosis tend to have sharp peaks and long tails, while distributions with low kurtosis are more rounded and have short tails as shown in figure 1.10.

$$K = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^4}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2)^2} - 3 \tag{1.12}$$

1.5 Classification

Machine learning is a branch of computer science where algorithms are used that allow a computer to extract relevant data from patterns and use that data to make

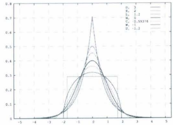


Figure 1.10: Kurtosis of some common functions. Sharper points have greater kurtosis. Image from [49].

intelligent decisions. It is especially useful in situations where patterns may be very complex and not feasible for a human to develop instructions for every possible situation, or even recognize the patterns. In the case of classification, the decision would be to what class the data belongs.

Computer learning falls under two broad categories: supervised and unsupervised learning. Unsupervised learning is used when there is no class data available for a data set. In this case objects are partitioned so as to best cluster the data. Supervised learning is used in situations where there is some sample data available with appropriate decisions that can be used as a training set.

Classifiers often operate in two phases. The training phase is where the relationship between certain features and outcomes is determined and optimized. This is often a long and computationally intensive process. The operating phase is when the training data is put to use to classify an object. This is usually much quicker.

Possibly the most important component of a classification routine is the feature vector. The feature vector is a set of scalar quantities that describe an object. The choice of a feature vector is vital to the success of a machine learning algorithm. The algorithms work by comparing the feature vector of a test object with those of objects already classified. If the data in the feature vector is not appropriate for the classification task, it will fail.

Usually, the initial choice of a feature vector is not the best one. Some features may not contribute to the classification task or might be made redundant by other features. Attempting to classify with these features can not only significantly increase computation time, but can make classifications less accurate. To mitigate this problem, a feature reduction step should take place. A good feature reduction process will result in faster learning due to less data, higher accuracy, and better generalization to other data sets [50]. There are two approaches of choosing a feature vector from all available features, top-down and bottom-up. The top-down approach takes a vector of all features and removes them one-by-one, testing the classification accuracy at each step. The bottom-up approach does the opposite. It starts with an empty vector and adds features to it one by one [50].

Classifiers can be either soft or hard. A hard classifier classifies an object without giving a probability. The assumption is made that an object that meets a certain criteria always belongs to a particular class. Soft classifiers give a probability of their classification. The assumption made is that sometimes objects with similar features may belong to different classes [51].

1.5.1 K-nearest neighbour classifier

K-nearest neighbour classification is performed by finding the K nearest neighbours in the feature space defined by the feature vector. Each neighbour votes on the classification of the unknown object. Each vote may be counted equally, or more priority may be given to votes of the closest neighbours. Closeness of neighbours in n-space is usually calculated from the n-dimensional Euclidean distance metric [52]. For example, votes may be weighted by $1/d$ where d is the distance to the object in feature space. The optimal choice of K and the weighting function, if any, depend on the data set used. The choice of a feature vector is especially important when small K is used due to the effects of features unrelated to the classification task at hand.

K-nearest neighbour is classified as a lazy learner, and as such, there is no initial training phase. Lazy learners have two drawbacks. First, they require more storage space, since all of the training objects must be available each time the classifier runs. Secondly, they require more calculation at the time of classification, since they have to retrain for each object classified [53].

1.5.2 Bayesian classifier

Bayesian Statistics

Bayes' theorem states that for two related events, A and B, the probability of A, given B, is dependent on the probabilities of just A, just B, and B given A according to equation 1.13. Where $P(A|B)$ is the posterior probability, $P(B|A)$ is the likelihood, $P(A)$ is the prior probability, and $P(B)$ is the evidence [54]. For our purposes, posterior probability is the probability the object belongs to a class based on its feature vector. The likelihood describes the chances that an object in a class could produce a particular feature vector. The prior probability is the probability that any object belongs to a particular class, and requires a priori knowledge about the distribution of data. The evidence normalizes the probabilities so that they sum to one.

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \tag{1.13}$$

Naive Bayesian classifier

The naive Bayesian classifier is a type of supervised learning classifier using Bayesian inference and the (often incorrect) assumption that features are independent of one another. Despite this assumption the classifier works well on many types of data [55]. If features were assumed to be related to one another then objects would need to be classified in n dimensional space, instead of in 1 dimensional space n times with the independence assumption. Since many times more points are needed to adequately cover n dimensions than 1 dimension, an advantage of the independence assumption is that a relatively small amount of training data is required to train the classifier. This leads to a drastic reduction in the complexity of the classification process.

Where n features are used, the probability is written as $p(C|F_1, F_2, ..., F_n)$, where C is the classification and F_n is the nth feature in the feature vector. Without the independence assumption the probability is defined by equation 1.14, with the independence assumption it can be simplified to equation 1.15 [53].

$$P(C|F_1, ..., F_n) = P(C)P(F_1|C)P(F_2|C, F_1)...P(F_n|C, F_1, F_2, ..., F_n - 1) \quad (1.14)$$

$$P(C|F_1, ..., F_n) = \frac{1}{Z}p(C)\prod_{i=1}^{n} p(F_i|C) \quad (1.15)$$

The most common implementation of the naive Bayesian classifier bins feature space into histograms for each classification. When a feature for an image is tested, it is first determined to which bin in each histogram the data point would fit, then the number of training data for that bin in each histogram is used to make a classification. The naive Bayesian classifier is an eager learner, so all that needs to be taken from the training phase is the histogram for each classification. Since it only needs to be trained once, it has a fast operational phase.

1.5.3 Cross-validation

Cross validation is used to estimate how a machine learning algorithm will perform when faced with unfamiliar data. It is intended to reduce error associated with one of the pit falls of machine learning, where a hypothesis is formed based on the same data used to test it [51].

In K-fold cross validation the data is randomly divided into K partitions. Data in one partition are used to test, and the remaining partitions are used to train. This means that training data needs to be calculated K times as each partition gets tested.

Leave one out cross validation is the limiting case of K-fold where K is the number of data points in the set. One image is tested against all other available images. This is repeated for each image. This allows the maximum possible use of a set of data for both training and testing. It is particularly useful when either the data set is not particularly large or the classification problem is particularly complex. It is also the most computationally intensive means of cross validation since the system needs to be retrained for each image tested.

Chapter 2

Materials and Methods

This chapter discusses the design of the computer aided detection system. The overall approach taken consisted of four sequential steps. Preprocessing, in which images are prepared by removing useless information and standardizing size, resolution, and bit depth. Transformation, in which images are cosine transformed to arrange image data by spatial frequency in two dimensions. Feature extraction, in which feature sets consisting of a small group of values are calculated from the transform for each image. Finally classification, in which machine learning was used to classify images by comparing them with images of known pathology.

Section 2.1 describes how the computer programs were written for this project and how the data was stored. Section 2.2 discusses the mammogram images collected from two sources. Section 2.3 explains adjustments made to the raw images. Section 2.4 describes how the DCT was applied to images and a proof of concept experiment to test its usefulness. Section 2.5 discusses how features were extracted from the

DCT. Finally, section 2.6 explains how the k-nearest neighbour and naive Bayesian classifiers were implemented to classify images.

2.1 Programming environment

Since this work involves complex calculations on very large datasets, it was necessary to automate the process by developing computer programs. All calculations were carried out on a desktop computer with an Intel core 2 duo E7200 processor and 2 GB RAM. Programs were developed for this work in two different programming environments - Mathematica, and C++. The bulk of the work was done in Mathematica for its large number of built in functions and the speed at which scripts can be developed. C++ was used in situations where running programs in Mathematica would be too time consuming. Programs compiled in C++ tend to run much faster than those in Mathematica, but generally take longer to develop.

2.1.1 Database

All the data generated from the various programs was stored in a relational database using MySQL server 5.1. This enabled rapid and efficient storage and retrieval of data. In order to handle some of the large datasets used, the database server had to be configured to allow for large data packets. The line *max_allowed_packet = 100M* was added to the my.ini configuration settings under *[mysqld]*. This allows packets of up to 100 megabytes, and prevents stack overflow errors when working with large bitmap images. Most queries were performed using the DatabaseLink toolkit for

Mathematica. It was also necessary to increase the size of the Mathematica Java heap in order to accommodate the large amount of data being passed back and forth from the database.

The database contains eight tables: *mias*, *mias_features*, *mias_knn*, *mias_bayes*, *ddsm*, *ddsm_features*, *ddsm_knn*, and *ddsm_bayes*. These tables will be discussed as they are used in the sections to follow.

2.2 Data collection

Two tables, *mias* and *ddsm* were created to locally store all the raw image data. The data stored in the table is as follows:

id: A unique identifier for each image. Images are in right left pairs with right images being odd numbers and left being even.

character (*mias* only): The density of the breast as determined by a radiologist. *f* (fatty), *g* (fatty-glandular), or *d* (dense-glandular).

class (*mias* only): The class of the abnormality as determined by a radiologist. *calc* (calcification), *circ* (well-defined/ circumscribed masses), *spic* (spiculated masses), *misc* (other, ill-defined masses), *arch* (architectural distortion), *asym* (asymmetry), or *norm* (normal).

severity: The severity of an abnormality as determined by a radiologist. *m* (malignant), *b* (benign), or *n* (normal).

image: The binary of each PGM image. Header information is not included since all images are P5 PGMs with 1024 by 1024 resolution and 255 maximum gray value.

img_hash: A md5 hash of the image binary to be used as a check number to verify the database is consistent with the original PGM files.

Data was obtained from two publicly available mammographic image databases. The digital database for screening mammography (DDSM) [36, 37] and the mammographic image analysis society (MIAS) database [35]. These two databases are extremely useful for development and testing of computer aided detection schemes. Combined, there are thousands of images available for download. All images have been carefully examined by radiologists specializing in mammography, and their biopsy confirmed findings documented with each image. Patients were then followed for several years to ensure that no suspicious features had been missed. Both databases include "ground truth" data that describes both the type and location of abnormalities present in the images

The two databases contain images saved in different formats, with different resolutions, collected on different machines, by different technicians. This variety provides extra challenges for any CAD program. While similar images with slight differences in intensity or scaling may not even be noticeable to a human reader, the changes in the raw data will be significant. For example, masses in a 50 micron image may appear 16 times larger than masses in a 200 micron image if the pixel resolution is not taken into account. By testing the program on images from different sources, it can be ensured that it will be robust enough to handle the variations in images from multiple screening sites.

All images used in this research are stored in PGM format. This is an image format that stores data in a very easy to access and read pattern. The P5 variant of PGM was used, which uses binary encoding of the gray map. Comment lines begin with a hash sign (#) and end with a new line. The first three non comment lines of the file contain information about the image; the remainder contains the binary for the image. The first line contains two characters which identify the type of file it is. In this case those characters are P5. The second line contains two numbers that define the width and height of the image. The third line contains one number which defines the maximum gray value for a pixel. The rest of the file contains one byte intensity values for each pixel in the matrix that makes up the image. Files are saved with a .pgm extension.

2.2.1 MIAS database

The MIAS database consists of 322 mediolateral oblique images (161 left, right pairs) with ground truth data for each one. Images are characterized according to density, class of abnormality, and the severity of the abnormality present. There are 112 dense, 104 glandular, and 106 fatty images. Among these images, there are 209 normal, and 113 suspicious (61 benign and 52 malignant). The suspicious images are further broken down into 23 images containing calcifications, 23 containing circumscribed masses, 19 containing spiculated masses, 14 with ill-defined masses, 19 containing architectural distortions, and 15 pairs with asymmetric densities.

All images have been scaled to 200 micron resolution, cropped to no larger than 1024 px by 1024 px, and padded with black to no smaller than 1024 px by 1024 px. The images are saved in eight bit portable gray map format with pixel intensity ranging from 0 to 255. The reduced size and contrast is roughly at the limit of where small calcifications are still able to be resolved as seen in figure 2.1. The advantage of having all images the same shape is that they can all be fed into the same algorithm without having to account for different aspect ratios. 1024 pixel edges is convenient because it is a power of two, which simplifies the partitioning of images into chunks. 1024 pixels square and one byte per pixel greatly reduces the computation necessary to process each image as they contain only around one thirtieth the data of an original image.

2.2.2 DDSM

The DDSM database is much larger than the MIAS. It contains 2620 cases. The cases are partitioned into 43 volumes. 12 normal, 15 cancer, 14 benign, and 2 benign without callback. Each case contains between six and ten files: four lossless jpeg mammogram images (left and right craniocaudal and mediolateral oblique), one overview pgm image of the case, one ics file containing information about study date, patient age, density, date of digitization, type of scanner used, image resolution, sizes, and bit depths, and between zero and four overlay images indicating the locations and

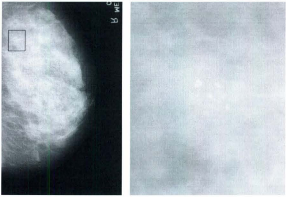


Figure 2.1: The small calcifications in this image are still visible when downsampled to 1024 pixels square. This shows that an acceptable amount of detail is retained at this resolution.

sizes of any abnormalities. Images in the DDSM database were much higher quality than the MIAS images. Resolutions were in the range of 42 to 50 microns per pixel, and bit depth was in the range of 12 to 16 bits per pixel (4096 to 65536 gray values).

Normal volume 1 and cancer volume 1 were used as training data. These two volumes have a combined total of 360 images, with 285 normal and 75 suspicious (5 benign and 70 malignant).

This database uses images stored in a proprietary variant of the jpeg image format. Since there is no documentation available to describe the method used to encode and compress the images, software provided with the database had to be relied upon to convert the data to a more readable form. The decompression program provided with the database was used to decompress and convert the images into PGM (P5) format. The problem with these programs, however, is that they have poor compatibility with most currently available operating systems.

The software was written for the SunOS 5.6 operating system by Michael Heath in 2000 (`http://marathon.csee.usf.edu/Mammography/software/heathusf\_v1.1.0.html`). The most compatible operating system available, Sun OpenSolaris 2009.06, with legacy C compiler packages installed was used to compile and run the software in a tcsh shell. This was run as a virtual machine on a Windows XP system using Sun Virtual Box 3.1. The program was modified to swap bytes from big-endian to little-endian format to make the output files compatible with windows. These images were much too large to be able to perform the required calculations in a reasonable amount of time, so they were scaled to match the MIAS images. The program was altered to output images in eight bit gray scale with a resolution of 200 microns per pixel.

Once images were down sampled to 200 microns per pixel they were cropped in either dimension that exceeds 1024 pixels. Since images are usually centred on the films, lines or rows are removed equally from both sides. This proved effective in all but a few images where the breast exceeded the borders of the image. These images were discarded. Dimensions with less than 1024 pixels are padded with zeros (black) on both sides of that dimension to make up 1024 pixels. At this point, all images are 1024 pixels square.

The DDSM has three main advantages over the MIAS database. It has more images, it has higher resolution images, and it has craniocaudal images in addition to mediolateral oblique. Its only downside is that it is not nearly as user friendly.

2.3 Preprocessing

Preprocessed images from the two image databases were added to the *mias* and *ddsm* tables. Two columns were added to each table.

proc_image: The binary of each preprocessed PGM image.

omit: Whether or not to use the image. Either 1 (omit) or 0 (do not omit).

The first step in the image classification process is preprocessing. In this step the goal is to produce a set of images that only contain the breast tissue. Other image variations unrelated to pathology such as orientation tags, resolution, intensity range, and aspect ratio should be removed or standardized. Differences from one image to another that are unrelated to the final classification should be removed so as to not influence the classifiers.

The most complicated step in the preprocessing stage is removal of extraneous objects, such as orientation tags, from the images. In order for this step to proceed autonomously, the computer must recognize different discrete objects in the image, be able to determine which one is the breast, and remove everything else.

In the first step, intensity values are binned into a histogram with bins for each of the 256 intensity values (figure 2.2. The distribution of the histogram for mammogram images is roughly bimodal with a sharp peak for very low values corresponding to the background, and a broad peak for the brighter foreground. Using this histogram, a cutoff point was chosen, above which, values were considered foreground, and below which, values were considered background. Dense breasts produce very bright mammogram images with a sharp relatively easy to define border between the background and foreground. Fatty breasts yield dim images with a much fuzzier border between the background and foreground.

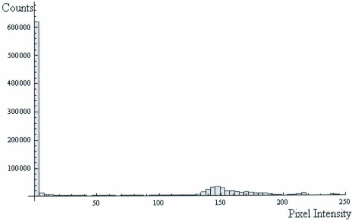


Figure 2.2: Intensity histogram for each pixel of a typical mammogram. The sharp peak at and around zero intensity is due to the dark background.

Four algorithms were used to find threshold values and the results were visually inspected to find which resulted in the best borders between background and foreground. The algorithms used were Otsu's algorithm, Kapur's method, mean, half mean, and the Kittler-Illingworth minimum error thresholding method [56]. Images of all densities were tested.

With background and foreground pixels separated, background was assigned a value of zero (black) and foreground was assigned a value of one (white). At this point background noise had been eliminated and foreground objects were distinct white objects.

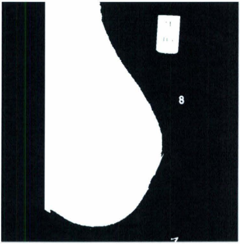


Figure 2.3: A mammogram after the thresholding algorithm was applied. The background is black, foreground objects are white. Objects within the image can now be easily distinguished from one another.

The next step was to separate and quantify the foreground objects. The image was raster scanned to find distinct white segments along each line of the image. When a white pixel was encountered it was assigned a number and all adjacent white pixels on that line were assigned the same number. When a black pixel was encountered the number was incremented and the process continued when the next white pixel was detected.

Once the entire image had been divided into numbered line segments, each pixel in each line segment was tested to determine if it neighboured a pixel of another line segment either on a straight or diagonal line. The segments comprised the nodes of undirected graphs with edges drawn between neighbouring segments. The largest graph was assumed to represent the breast tissue in the image. All the pixels in all the nodes of the largest graph were assigned a value of one and all other pixels were assigned a value of zero. The Hadamard product of the mask and the original image was taken to produce a final image containing only the relevant breast tissue.

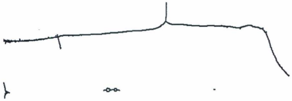


Figure 2.4: A graph representation of the white line segments making up an image. Each white line segment is a node in the graph. Adjacent segments are joined by edges. The four connected graphs correspond to the four objects visible in Figure 2.3. The largest belongs to the breast tissue.

Finally, left facing images were inverted horizontally to produce mirror images so that they all faced the same direction. It was arbitrarily chosen that all images should face right. To determine the orientation of the images, left and right borders of the breast were found by raster scanning each line from left to right, with the first non zero pixel lying on the left border and the last non zero pixel lying on the right border. These left and right borders were then fit to a vertical line. The border with the best fit was on the chest wall side. Images with the chest wall on the right side

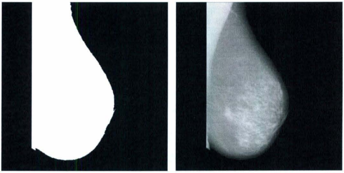


Figure 2.5: The mask shown on the left is the result of removing all but the largest object from the binary image in Figure 2.3. Overlaying the mask with the original image leaves only the region of interest shown on the right. Only the actual breast tissue remains in the image.

were inverted horizontally.

While this process worked extremely well, it was limited by the quality of the input images. Images that were out of frame, had overlapping tags, or other artifacts were discarded. This step was performed manually by displaying and spot checking all images for any abnormalities. Since images only required a brief glance to determine whether or not the preprocessing program was successful, spot checking each one was not a time consuming task. Images that were not properly preprocessed were given a value of 1 in the omit column of their respective database table.

2.3.1 Chest wall

The chest wall is a prominent, non pathological, feature in mammograms that could potentially mislead computer aided detection programs. Early attempts at removing chest wall from mammogram images involved edge detection and smoothing methods to filter the chest wall while leaving the underlying soft tissue, where cancers are sometimes found, relatively undisturbed. It, however, proved to be exceedingly difficult to achieve consistent results on all images without some human input. Even with human input, it was difficult to determine the boundary of the chest wall on some images as shown in figure 2.6. Because of these obstacles, a discrete chest wall removal step was not used.

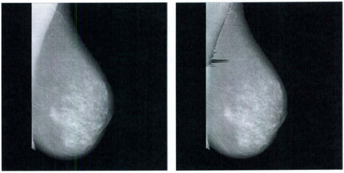


Figure 2.6: The chest wall has been removed, with soft tissue still visible. The blood vessels and other soft tissue overlapping with the chest wall are much easier to see in the second image. Edge detection worked poorly in this example where the border of the chest wall is fuzzy.

2.4 Transformation

Four fields were added to the *mias* and *ddsm* tables.

dct1: The discrete cosine transform I of each preprocessed image represented as a floating point array.

dct2: The discrete cosine transform II of each preprocessed image represented as a floating point array.

dct3: The discrete cosine transform III of each preprocessed image represented as a floating point array.

dct4: The discrete cosine transform IV of each preprocessed image represented as a floating point array.

Preprocessed image data from the database were imported into Mathematica as 1024 by 1024 arrays. 2D Discrete Cosine transforms (I, II, III, and IV) were performed on all preprocessed images using the built in *FourierDCT* function as seen in figure 2.7. The DCT images were then saved to the *mias* database as 32 bit floating point arrays.

2.4.1 Filtering

To test the potential efficacy of extracting useful features from the discrete cosine transform, a proof of concept experiment was performed. This involved writing a Mathematica script that allowed regions of a DCT to be selected and extracted. The extracted region of DCT would then be inverse transformed and the resulting image

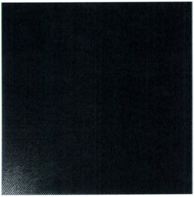


Figure 2.7: The DCT of a mammogram image. Space has been transformed into the frequency domain. The bottom left of the image represents low frequency features. Points along the horizontal and vertical directions represent higher frequency features in those directions.

displayed. Inverse transforming different sections of the DCT image resulted in the original image with some characteristics emphasized and others removed entirely (figure 2.8. The ability to slide through the different components of an image according to frequency spectrum may even make this a useful tool in and of itself.

2.5 Feature extraction

The tables *mias_features* and *ddsm_features* were created to contain all the features generated from the DCT transforms. The data stored in the table is as follows:

id: A unique identifier for each image. Matches images in this table to the ones in *mias* and *ddsm*.

view (*ddsm* only): The view of the mammogram. Either *MLO* or *CC*.

dct1_origin - dct4_kurt10_abs: Each of the 324 features generated in the feature extraction step.

The next step in the process involved the generation of feature vectors for each image. The general strategy here was to partition the DCT into blocks and perform statistical calculations on them, yielding a list of scalar quantities for each image. Ideally, some of these quantities would be useful in the detection of cancer.

It was decided that the data would be partitioned according to distance from the origin so as to separate features according to frequency. There were two methods

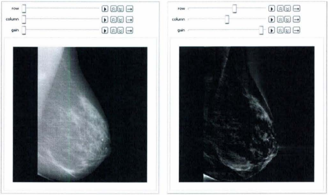


Figure 2.8: Screenshots of a program written to test the effects of removing portions of the DCT image. The sliders at the top are used to remove low frequency rows and columns. The inverse transform image is then displayed. Filtering parts of the DCT for a normal image causes some characteristics of the image to be removed and others to be emphasized.

considered for partitioning the data in the DCT images. The first method tested was to use concentric quarter rings centred on the origin. This seemed like a good choice, since the distance range of points from the origin would be constant in any direction. However there were some technical problems with this method that occurred for very low and very high frequency data. At the low frequency end, narrow rings with a very small radius sampled on a square grid had very disjointed edges and irregular shape. At high frequency, the outermost ring would have to have a curved interior border but a square exterior, making this region radically different from the rest. L shaped blocks were chosen because they sample in a much more uniform way as frequency space is traversed.

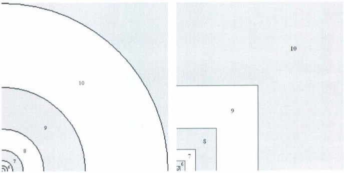


Figure 2.9: Two potential sampling schemes, ring shaped partitions and L shaped partitions. Values of the transformed image would be taken from each numbered partition.

Once the shape of the blocks had been decided, it was necessary to determine an appropriate size for each one. As explained in section 1.4.1, the DCT exhibits a high degree of energy compaction. Since a large portion of the image data is encoded close to the origin in a DCT, it was deemed important to focus more on this area. With this in mind, the widths of blocks were smaller closer to the origin so that more of the features would come from this data rich part of the transform. Blocks were chosen so that each would have twice the width of the proceeding block. This means that in a 1024 by 1024 matrix, there will be ten blocks which span the following frequency ranges. Block 1: 1-2, block 2: 3-4, block 3: 5-8, block 4: 9-16, block 5: 17-32, block 6: 33-64, block 7: 65-128, block 8: 129-256, block 9: 257-512, block 10: 513-1024. These blocks are labeled in figure 2.9. In addition to these blocks, the origin (DC coefficient) was also used.

The partitioning was done for each of the four DCTs. Absolute values of the transforms were also used in addition to the four original transforms. This makes a total of eight transform images for each mammogram image. For each of the ten blocks in each of the eight DCT images, the four statistical quantities, mean, standard deviation, skewness, and kurtosis were calculated giving a total of 320 features. Finally, the intensities of the four DC coefficients were added to the feature sets. Since the DC coefficient is always positive, the absolute value is not necessary. This brings the total number of features calculated for each image to 324.

2.5.1 Feature reduction

A large feature set would make some of the following calculations extremely computationally intensive. This becomes especially problematic when we consider that combinations of features will be used in the classifiers. For example, if one feature is used in a classifier, there will be 324 classifiers to test for each classification scheme. Combining two features, we get 52 362 combinations. There are 5 616 324 combinations of three features, all the way up to $1.51 * 10^{96}$ combinations of 182 features. Testing all combinations of 324 features would be computationally infeasible.

Before any extensive testing can take place, a feature reduction step is necessary to discard redundant or poor performing features. To this end, preliminary tests of single and double feature classifiers were conducted on the Bayesian classifier described in section 1.5.2. It was observed that some features taken from DCT I-IV produced near identical results and largely found the same images suspicious and normal. Since these features appear to be redundant, all but the DCT II based features were discarded, bringing the total feature count down to 81.

These DCTs were chosen for removal from the dataset even though some individual remaining classifiers performed worse than some of the removed classifiers. The justification for this is that the removed features show a high degree of redundancy in the image classification. A feature in DCT II might agree on 95% of their classifications with the corresponding feature in DCT III. This is of little use when combining features as their combination will produce results that are only a marginal improvement over using just one of them alone. A poorly performing feature that classifies images completely different from other features will offer a unique combination classification. This poorly performing feature may be able to pick up the suspicious images that all others missed.

The second trend noted was the poor performance of the features calculated from the absolute value of the transforms. In general, absolute value features did not perform as well as their counterparts. Because of this, absolute value features were also discarded, bringing the total feature number down to a more manageable 41.

41 features can be combined in sets of two in 820 ways, and in sets of three in 10660 ways. This is a large reduction from the original set and makes combining features possible.

Each of the 41 features was assigned a number to make referencing them easier. Feature number one is the DC coefficient of the DCT. The remaining 40 features are numbered as shown in table 2.1. This numbering system remained constant for the entire of the project.

Table 2.1: Feature numbering scheme. Each of the four moments of the mean for each of the ten partitions were assigned a number for use in database storage.

Region	Mean	Standard deviation	Skewness	Kurtosis
1	2	12	22	32
2	3	13	23	33
3	4	14	24	34
4	5	15	25	35
5	6	16	26	36
6	7	17	27	37
7	8	18	28	38
8	9	19	29	39
9	10	20	30	40
10	11	21	31	41

2.5.2 Phantoms

Phantom images were used in order to test the potential efficacy of the features chosen. A phantom image consisted of a normal, preprocessed, image with a mass digitally superimposed into it. Oval circumscribed masses were generated roughly in the centre of the image. Size, eccentricity, and intensity were all variable and gradients were used for the intensities at the edges of the mass in order to simulate the decreased attenuation at the border of an ellipsoid mass. A normal image and a phantom constructed from it are shown in figure 2.10.

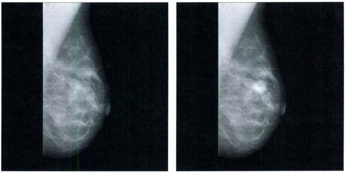


Figure 2.10: Results of the phantom image generation. Left: Normal image. Right: The same image with a round mass digitally superimposed in the centre.

The original image and the phantom with the mass were then run through the DCT and their features extracted. The features were plotted on a graph and compared.

2.6 Classification

There were two types of classification attempted using the feature sets generated for the available images, k-nearest neighbour and naive Bayesian classifiers. The classifiers were tested on four datasets, MIAS (full), DDSM MLO, DDSM CC, and MIAS with benign images removed (MIAS reduced). Images labeled as omit from section 2.3 were left out of both training and testing of the classifiers. Including poor quality data in the testing set would negatively impact the classifications of all other images.

Asymmetric images were not used in any experimentation. Since a classification of asymmetric density really requires comparison of pairs of images, which the k-nearest neighbour and naive Bayesian classifiers do not do, they were omitted for those classifiers.

2.6.1 K-nearest neighbour

The tables *mias_knn* and *ddsm_knn* were created to contain all of the k-nearest neighbour data for each image for all possible combinations of three features. The data stored in the table is as follows:

f1: The first feature number.

f2: The second feature number. 0 indicates the second feature was not used.

f3: The Third feature number. 0 indicates the third feature was not used.

image: The unique identification number of the image; matches the image to the other tables.

set (*mias* only): Which set the image belongs to. Can be either *full* or *reduced*.

view (*ddsm* only): The view of the mammogram. Either *MLO* or *CC*.

p1-p25: The 25 nearest images with respect to the features used.

dist1-dist25: The Euclidean distances of the 25 nearest images referenced in *p1-25* with respect to the features used.

imageseverity: The severity of the image referenced in the *image* column. Can be either *n*, *b*, or *m*. This column is somewhat redundant, but serves to speed up queries.

severity1-severity25: The severity of the images referenced in *p1-25*.

value1-value25: A value assigned to *p1-25* according to the severity of the image. -1 for normal, 1 for suspicious.

pvalue1-pvalue25: *value1-25* multiplied by a weighting factor according to the prior probabilities of the set.

The k-nearest neighbour classifier was tested on the DDSM mediolateral oblique and craniocaudal, as well as both the full and reduced MIAS sets. The first step in the k-nearest neighbour classifier was to generate a table containing the 25 nearest neighbours of each image for each feature vector. 25 neighbours was chosen to allow classification using any number less than or equal to 25 neighbours. Adding more neighbours is unlikely to be beneficial, as the closest neighbours are the most significant, and would increase computational time.

For each feature, differences were calculated with respect to all other images. Since all the features had different typical ranges for their values, the difference calculated was weighted to compensate for this. For example, using difference alone, with a feature equal 1 for one image and 2 for another, the difference is 1. If a feature has a value of 100 for one image and 101 for another, the difference is also 1. However, in the first example, the difference is 100% while in the second example, the difference is only 1%. To compensate for this, the value used in the algorithm is the difference of the two features divided by the sum (equation 2.1). This prevents features that naturally have a large range from dominating the feature vectors.

$$\frac{test\ image\ feature - neighbour\ feature}{test\ image\ feature + neighbour\ feature} \tag{2.1}$$

The 25 closest neighbours to each image were stored in the database. There are two important pieces of information gathered in this step: the ordering of the nearest neighbours, and the distance to each one. Storage in the database permits rapid retrieval and analysis. While it is not necessary, the classification of each neighbour is also stored in the database. This increases the storage requirements, but decreases processing time associated with querying data from more than one table at a time.

Combinations of two and three features were also tested. Percent differences were used again, with the Euclidean distance in two or three dimensional space ($d = \sqrt{x^2 + y^2 + z^2}$) calculated for each pair of images and each feature vector. More than three features at a time was too computationally intensive.

There are a few ways in which the classification algorithm was modified. In addition to the standard vote taking approach, distances to neighbours were also considered, as well as the proportions of normal and suspicious data in the training set (prior probability).

Since k-nearest neighbour classifiers tend to have a slow operational phase, additional columns were added to the database for each neighbour in order to speed classification, at the expense of storage space. Vote columns were added for each neighbour with +1 for positive (suspicious) and -1 for negative (normal). Columns were also added for votes weighted by prior probabilities so as to reduce the effect of having an uneven set of potential neighbours. For example, if there are twice as many normal images as suspicious in the training set, then votes for suspicious will count for twice as much as votes for normal. Normal votes have a fixed value of -1 while suspicious votes have a value of $\frac{number\ of\ normals}{number\ of\ suspicious}$.

The vote taking algorithm was tested for between 1 and 25 calculated neighbours for each test image (leave one out cross validation). Each neighbour gets a vote as to the classification of the test image. For example, if the five nearest neighbours of an image are being used with two of them suspicious and three normal then the image will be classified as normal (three to two).

Next, a weighting on the votes based on the prior probability of an image being suspicious was applied. There are many more normal images in the training data sets, so there was a lower probably of randomly having suspicious neighbours. Because of this, there was more significance placed on a suspicious neighbour. For example, if there are twice as many normal images as suspicious, then suspicious images will get two votes and normal images will get one each. So, in this case, if an image has three normal neighbours and two suspicious, it will be classified as suspicious (four to three). This is especially crucial since we place greater importance on sensitivity and having a much higher chance of a neighbour being normal skews the results toward higher specificity and away from sensitivity.

Finally, distance to each neighbour was considered. If an image has one really close neighbour and the rest are relatively distant, it makes sense to assign more priority to the vote of the closest neighbours and less to the farther ones. To achieve this, each vote is divided by the distance to that point. For example, a point that is 0.2 away will get 5 votes while a point that is 2 away will only get half a vote.

In order to calculate a sensitivity and specificity for each of the datasets for each number of neighbours, the number of true positives and true negatives from each feature vectors was found. Both true positives and true negatives used a similar algorithm. For true positives, with the plain vote taking method and single feature classifiers, first a database query was performed to find all suspicious images and their neighbours. For k neighbours the classification was found by adding the values assigned to the first k neighbours for each image returned in the query together. Since positives are assigned a value of -1 and negatives are assigned a value of +1, adding the votes together for the neighbours results in a value greater than or equal to zero if the consensus is suspicious and less than zero if it is normal. The number of images for each classifier from this set with a positive value is the number of true positives. True negatives were calculated in an almost identical fashion except the database query was for all normal images and their neighbours and the sum of votes should be -1 for a correct classification. This was calculated for each number of neighbours ranging from 1-25.

For classifiers using the prior probability weighted values, the votes were switched out for the weighted votes discussed above and added together as normal. Classifiers using distance weighted votes used instead, *vote/distance* and added as normal. This was done for classifiers using 1, 2, and 3 features. There were twelve variations of this classifier calculated. Unweighted votes, votes weighted by prior probability, weighted by distance, and weighted for both prior and distance for combinations of 1, 2, and 3 features. For each variation, and each number of neighbours, the feature vector with the highest sensitivity was selected and recorded along with its specificity.

2.6.2 Naive Bayesian classifier

The tables *mias_bayes* and *ddsm_bayes* were created to contain all the calculated probabilities for each image for all combinations of one, two or three features. The data stored in the table is as follows:

numbins: The number of bins used in the Bayesian classifier. Ranges from 2-20.

f1: The first feature number.

f2: The second feature number. 0 indicates the second feature was not used.

f3: The third feature number. 0 indicates the third feature was not used.

image: The unique identification number of the image.

posterior_norm: The Bayesian probability that an image is normal with respect to all other images using the same feature set.

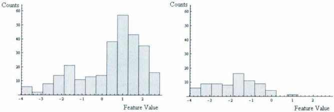
posterior_susp: 1-posterior_norm.

The final classifier tested was the naive Bayesian classifier. The program was designed to run in two stages. First all the probabilities were calculated and stored in a database, and then these were compared with the truth data to find the sensitivity/specificity of each feature. Storing the probabilities before assessing the sensitivity/specificity allows greater flexibility for analysis and refinement of thresholds. Like with the k-nearest neighbour classifier the DDSM mediolateral oblique and craniocaudal were analyzed separately, and both the full MIAS set and the reduced set without benign images were analyzed.

First, the number of normal and suspicious (malignant and benign) images was counted to find the prior probabilities of each classification from equation 2.3. Data was then queried from the appropriate database for the appropriate view. For each feature, two identically binned histograms for normal and suspicious images were setup. The size of the histograms was determined from the entire set of data. The lowest feature value for any of the images was set as the lower bound for the histogram and the highest value was taken for the upper bound. The values in between were divided equally into the desired number of bins. Each image was then placed in its appropriate histogram (figure 2.11).

$$prior_{suspicious} = \frac{number\ suspicious}{total\ images} \tag{2.2}$$

$$prior_{normal} = \frac{number\ normal}{total\ images} \tag{2.3}$$



(a) Normal histogram

(b) Suspicious histogram

Figure 2.11: 13 bin naive Bayesian classifier histograms for a single feature for normal and suspicious images. In this example, lower value features have a greater likelihood of belonging to suspicious images, than higher values. 13 bins.

Once the number of normal and suspicious images in each bin was known, each image was tested using leave one out cross validation. To prevent images from being compared with themselves, the corresponding bin in the appropriate histogram was decremented while an image was being tested. The likelihoods were calculated from equation 2.5. This is the probability that the image is in that bin if it is suspicious or in that bin if it is normal.

$$likelihood_{suspicious} = \frac{number\ of\ counts\ in\ suspicious\ bin}{total\ number\ suspicious} \qquad (2.4)$$

$$likelihood_{normal} = \frac{number\ of\ counts\ in\ normal\ bin}{total\ number\ normal} \qquad (2.5)$$

Next, unnormalized posterior probabilities were calculated. These are the likelihoods multiplied by the prior probabilities for both normal and suspicious probabilities. The normalized suspicious posterior probability was then calculated by dividing the unnormalized suspicious posterior by the sum of the normal and suspicious posterior (equation 2.6. The posterior probability that an image was suspicious was then added to the database.

$$posterior_{suspicious\ (normalized)} = \frac{posterior_{suspicious}}{posterior_{normal} + posterior_{suspicious}} \qquad (2.6)$$

The whole process was repeated for all images in the set, for each feature and using between two and 20 bins. Sturges' rule [57] was used to approximate the best number of bins to use. Using Sturges' formula (equation 2.7 the optimal number of bins should be approximately eight. However, this is only an estimate and may not be accurate for the data sets being used, so it was tested experimentally.

$$k = \lceil \log_2 n + 1 \rceil \qquad (2.7)$$

This process was then expanded to use more than one feature in the feature vector. Combinations of two and three features were used where there were 820 and 10660 possible combinations respectively. For four features there were 101270. Because of the rapidly increasing amount of computation for increasing number of features, three feature vectors were the largest that could be computed in a reasonable amount of time.

For each additional feature added, separate histograms were generated and likelihoods calculated. Using the independence assumption, (section 1.5.2) unnormalized posteriors were calculated by multiplying the likelihoods from each feature with the prior probability (equation 1.15. The posteriors were then normalized in the same way as for one feature.

One problem frequently encountered when doing this calculation occurred when a test image was assigned to an empty bin. If a bin in either a normal or suspicious histogram was empty, and a testing image falls within the range of that bin, it would result in a zero percent likelihood for that feature, and thus a zero percent posterior probability, despite the outcome of other features in the vector. If a bin in both the normal and suspicious histogram was empty, this would result in a zero percent likelihood the image was normal, and a zero percent likelihood the image was suspicious, regardless of the outcome of other features in the vector. To mitigate this problem, one count was added to each bin. This way, images falling in a formerly empty bin still received a low probability of belonging to that classification, but the probability was not so low that it completely dominated probabilities calculated from other features.

With a database populated with probabilities for each image for every possible combination of one, two, and three features, sensitivity and specificity was then calculated. In order to determine if an image was suspicious, a certain threshold probability had to be chosen, above which the image was classified as suspicious and below which, the image was classified normal. The threshold could be adjusted to emphasize sensitivity or specificity or to achieve a balance for the two. For example, setting a threshold of 0% posterior probability the image was suspicious would guarantee 100% sensitivity, but specificity would suffer and be 0%. A threshold of 100% could be chosen with the opposite result. Since the classifications of images in the training set were known and all posterior probabilities were calculated, thresholds were chosen based on this data. For 100% sensitivity, the lowest probability of a suspicious image being suspicious was used as the threshold. Specificity was then incrementally increased by using the second lowest probability of a suspicious image being suspicious as the threshold. This did however lower sensitivity since the suspicious image with the lowest probability of being suspicious was now classified as normal. Because probabilities were calculated and stored first, sensitivity and specificity could be rapidly calculated for any threshold, without having to retrain. Sensitivity and specificity were calculated for various threshold values in order to generate receiver operating characteristic (ROC) data.

To find the optimal number of bins, the single feature classifier was tested for each number between 2 and 20. The number of bins with the highest specificity with 100% sensitivity was used for that image set. In the event that two binning schemes had the same specificity with 100% sensitivity, the specificity of the two feature classifier would be used as a tie breaker. It would be not be feasible to fully test all binning schemes for all datasets since there are 19 binning schemes and four datasets (76 combinations). Testing each one for three features takes 10+ hours. Testing just the single feature classifiers should provide a close guess.

Chapter 3

Results

This chapter contains a summary of the results obtained in the project. Section 3.1 describes the outcome of the preprocessing step. Section 3.2 shows the feasibility of using the DCT to detect cancer. Section 3.3 shows the ability of the chosen features to detect cancer. Section 3.4 presents the accuracy of the k-nearest neighbour classifier and the effects of various parameters. Section 3.5 presents the accuracy of the naive Bayesian classifier and presents the effects of various parameters.

3.1 Preprocessing

The preprocessor was applied to three datasets, MIAS, DDSM mediolateral oblique (MLO), and DDSM craniocaudal (CC). For both the MIAS and DDSM MLO sets, between the various background thresholding algorithms, it was found by visual inspection that half mean performed best for most images. Otsu's method performed slightly better for fatty breasts, but worse for all others. Half mean was chosen as the thresholding algorithm. Half mean was initially used for the DDSM CC set, but that resulted in an unacceptable number of poorly preprocessed images (nearly 20%). All the thresholding methods were tested again on the craniocaudal set and it was found that mean worked best.

Images deemed to be of insufficient quality after preprocessing were excluded from further testing. Table 3.1 shows the distribution of data in each of the sets before preprocessing. Table 3.2 shows the distribution of data in each set after preprocessing. 2% of images were lost from the MIAS set, 6% from DDSM MLO, and 3% from DDSM cc. Since the detection of asymmetric densities requires the comparison of two images, which the classifiers do not do, 15 images classified as asymmetric in the MIAS database were not used. Values in parenthesis show the numbers with asymmetric images removed.

Table 3.1: Total number of normal, benign, and malignant images in each dataset.

	Normal	Benign	Malignant	Total
MIAS	209	61	52	322
DDSM MLO	285	5	70	360
DDSM CC	285	5	70	360

Table 3.2: Number of normal, benign, and malignant images remaining after poor quality images were removed. Numbers in parenthesis indicate numbers after asymmetric images were removed.

	Normal	Benign	Malignant	Total
MIAS	205	60(54)	50(41)	315(300)
DDSM MLO	269	5	65	339
DDSM CC	278	5	68	351

In most cases, the preprocessor was able to successfully remove objects unrelated to breast pathology from the images such as in figure 3.1. However, in some images, either some tags remained or some of the breast tissue was lost. Figure 3.2 shows a poorly preprocessed image.

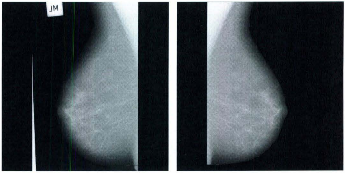


Figure 3.1: An image before and after the preprocessing step. Non-relevant objects have been removed and the image has been flipped to face right.

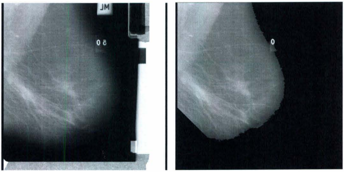


Figure 3.2: An image where preprocessing failed. Part of a tag overlapping with the breast remains, the edge of the breast is rough, and the bottom left corner is clipped. This image was discarded.

3.2 DCT filtering

It was found that appropriate selection of a region in the DCT transform could be used to filter all suspicious masses from an image. Figure 3.3 shows an image with a malignant mass that has been filtered to only show the mass. This was done by simply removing the bottom two rows and three left most columns of the transform, performing the inverse transform, and increasing the gain. This program clearly demonstrated that suspicious objects can indeed be isolated from regions of the DCT.

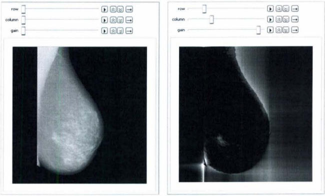


Figure 3.3: Using the program described in section 2.4.1, everything but the mass has been manually removed. In this case it only required removing the two lowest frequency rows and the three lowest frequency columns from the transform.

A second discovery made was an effective method of chest wall removal. It was found that for all the images tested, the chest wall could be cleanly removed from an image by cropping the bottom two lines and leftmost two columns from the transform. The images were first transformed into frequency space using the DCT II algorithm.

The first two lines and first two columns were removed and the image was transformed back using DCT III (figure 3.4). The DCT can be used to accurately distinguish the chest wall from the rest of the breast. This could potentially be exploited to remove the chest wall from an image while leaving the soft tissue of the breast intact. Not only does this method do an excellent job of removing the chest wall, it also means that the chest wall does not need to be removed in a discrete step. The localization of the chest wall in the cosine transform means that due to the inherent nature of the feature extraction step, it will have limited impact on the features.

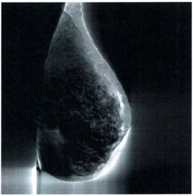
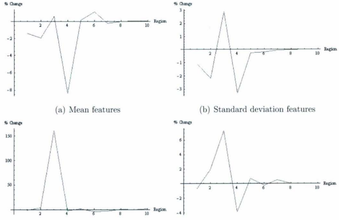


Figure 3.4: The chest wall has been completely removed from an image by removing the two lowest frequency rows and columns from the transform and inverse transforming the result.

3.3 Phantom

As expected, at most frequencies, the features of the phantom were nearly identical to those of the normal image. At the frequencies corresponding to the size and shape of the mass, the features diverged from those of the original image. This shows the potential of using features generated in this manner to classify images with and without masses. The example shown in figure 3.5 plots the percent change for each feature from the phantom shown in figure 2.10.



(a) Mean features

(b) Standard deviation features

(c) Skewness features

(d) Kurtosis features

Figure 3.5: Percent change in feature values upon adding a phantom mass to an image. Note that all plots have different scales.

In this example, the feature numbers (see table 2.1) with the greatest changes were 33 (2%), 14 (2.9%), 23 (3.2%), 34 (7.3%), and 24 (160.3%). It would seem that for this particular mass, feature number 24 would do an excellent job at determining whether or not the image is suspicious.

Since this was only a proof of concept experiment, a more rigorous analysis was not performed. The results demonstrate quite clearly that there is a change in features when an abnormality is introduced into the image. This was enough evidence to move forward with using the feature set to attempt to classify real images.

3.4 K-nearest neighbour

The k-nearest neighbour classifier was tested on four datasets. The full MIAS set, MIAS with benign images removed, DDSM MLO and DDSM CC. Testing single feature classifiers for between 1 and 25 neighbours took on the order of minutes for each set. Testing for all combinations of two features took on the order of hours, and testing for three features took about two days per set. For each of the 12 variants of the classifier, the most sensitive feature or feature set was recorded with its corresponding specificity. Full results can be found in appendix A.

3.4.1 Best features

The best performing features for the k-nearest neighbour classifier were found by tallying all the features from the top performing classifiers for 1-25 neighbours, for each data set, for each of the 12 variants of the classifier. The most frequently occurring features are the best. For MIAS full, the best features were 30 (14.6% of classifiers), 8 (16.7%), and 12 (17.7%). For MIAS reduced, the best features were 23 (16.8%), 33 (21.9%), and 12 (39.4%). For DDSM MLO the best features were

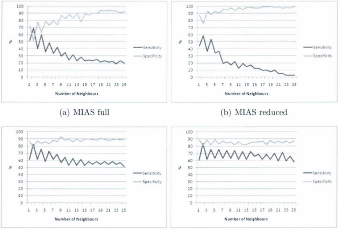
24 (19.1%), 5 (26.7%), and 13 (42.0%). For DDSM CC the best features were 37 (10.8%), 32 (21.9%), and 24 (69.1%).

3.4.2 Majority voting

The first variant of the k-nearest neighbour classifier assigned all votes an identical value. For each data set, the best sensitivity was found when using two neighbours and three features. The highest sensitivity for the MIAS full set was 69.5% with 51.7% specificity. For the reduced MIAS set, the highest was at 58.5% sensitivity and 75.6% specificity. For DDSM MLO the highest was at 82.9% sensitivity and 78.8% specificity. For DDSM CC the highest was at 83.6% sensitivity and 80.9% specificity. This classifier leans towards high specificity and low sensitivity. As the number of neighbours increases, this effect becomes greater, especially for the MIAS sets, as seen in figure 3.6.

3.4.3 Distance weighted votes

Using votes weighted by distance, the best sensitivities were found using three feature classifiers. For the MIAS full set, the highest sensitivity was 51.6% with 72.2% specificity for 1 or 2 neighbours. For the MIAS reduced set, the highest was 43.9% sensitivity and 87.3% specificity for 1 or 2 neighbours. For the DDSM MLO set, the highest was 65.7% sensitivity and 85.1% specificity for 3 neighbours. For the DDSM CC set, the highest was 61.6% sensitivity and 89.6% specificity for 5 neighbours. As with the vote taking method, this classifier tends to have high specificity and low sen-



(a) MIAS full

(b) MIAS reduced

(c) DDSM MLO

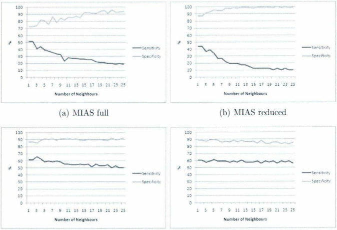
(d) DDSM CC

Figure 3.6: KNN classifier with majority voting sensitivity and specificity by number of neighbours used. Using more neighbours results in higher specificity but lower sensitivity

sitivity. With more neighbours, the sensitivity and specificity diverge further, with the effect more pronounced in the MIAS sets. The distance weighted classifier tends to smooth out the fluctuations seen in the vote taking classifier, as seen in figure 3.7.

3.4.4 Prior adjusted votes

With votes adjusted for the prior probabilities, again, the best classifiers used three features. For MIAS full, the highest sensitivity was 83.2% with 34.6% specificity for 3 neighbours. For MIAS reduced, the highest was 92.7% sensitivity and 39.0% specificity with 6 neighbours. For DDSM MLO the highest was 98.6% sensitivity and 63.2% specificity with 13 neighbours. For DDSM CC the highest was 98.6%



(a) MIAS full

(b) MIAS reduced

(c) DDSM MLO

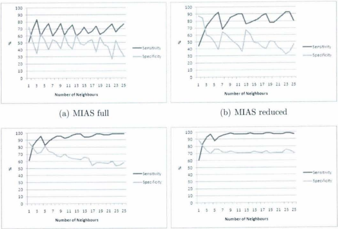
(d) DDSM CC

Figure 3.7: KNN classifier with distance weighted voting sensitivity and specificity by number of neighbours used.

sensitivity and 75.2% specificity with 23 neighbours. Unlike the first two classifiers, this one favored sensitivity over specificity (figure 3.8).

3.4.5 Distance weighted and prior adjusted

When using both distance weighting and prior probabilities, the best classifiers used three features. For MIAS full, the highest sensitivity was 69.5% with 52.5% specificity for 2 neighbours. For MIAS reduced, the highest was 85.4% sensitivity and 55.1% specificity with 24 neighbours. For DDSM MLO, the highest was 98.6% sensitivity and 65.7% specificity with 15 neighbours. For DDSM CC, the highest was 97.3% sensitivity and 75.9% specificity with 12 neighbours. Weighting the prior adjusted



(a) MIAS full

(b) MIAS reduced

(c) DDSM MLO

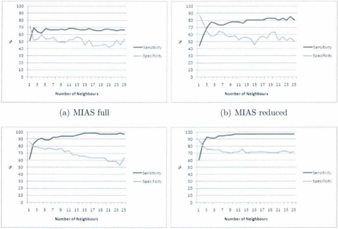
(d) DDSM CC

Figure 3.8: KNN classifier with prior adjusted voting sensitivity and specificity by number of neighbours used.

votes by distance has similar results to the un-weighted prior adjusted classifier, but with most of the fluctuations removed from the curves (figure 3.9).

3.5 Bayesian classification

The naive Bayesian classifier was tested for all four data sets. Feature vectors with between one and three features and histograms with between two and 25 bins were tested. Using various thresholds, sensitivity/specificity levels were adjusted to generate ROC curves. Training/testing took a similar amount of time to the KNN classifier - minutes for all single features, hours for all double features, and days for all triple feature vectors.



(a) MIAS full

(b) MIAS reduced

(c) DDSM MLO

(d) DDSM CC

Figure 3.9: KNN classifier with distance weighted and prior adjusted voting sensitivity and specificity by number of neighbours used.

3.5.1 Best number of bins

The highest accuracy single feature classifiers at 100% sensitivity generally remain the highest accuracy classifiers when the sensitivity is lowered for any number of bins for either the MIAS sets or DDSM MLO. The same is true when using the same number of bins in 2 features sets. Table 3.3 shows the specificity at 100% sensitivity when using different number of bins in the histograms. The best number of bins to use at 100% sensitivity can be expected to remain constant for lower sensitivities as well.

In the cases of both MIAS full and MIAS reduced, there were ties for the best number of bins when using single feature classifiers. In MIAS full, both eight bins

Table 3.3: Naive Bayesian classifier performance by number of bins. Specificity at 100% sensitivity, one feature.

Bins	MIAS full	MIAS reduced	DDSM MLO	DDSM CC
2	0 %	1.4 %	0.7 %	3.2 %
3	0 %	2.9 %	15.6 %	7.6 %
4	3.4 %	21.0 %	41.3 %	38.8 %
5	2.4 %	14.6 %	32.0 %	27.3 %
6	3.4 %	11.7 %	15.6 %	12.2 %
7	3.9 %	12.7 %	43.1 %	31.7 %
8	9.3 %	25.9 %	41.3 %	12.2 %
9	2.0 %	20.0 %	27.5 %	10.8 %
10	5.9 %	15.1 %	44.6 %	18.7 %
11	4.4 %	18.5 %	40.9 %	21.6 %
12	9.3 %	25.9 %	49.1 %	12.2 %
13	5.4 %	21.9 %	46.1 %	20.5 %
14	5.4 %	15.6 %	43.1 %	26.6 %
15	4.4 %	19.0 %	42.0 %	18.7 %
16	4.4 %	24.9 %	46.5 %	27.9 %
17	4.4 %	22.4 %	43.5 %	16.5 %
18	3.9 %	24.4 %	27.5 %	12.9 %
19	5.9 %	24.4 %	47.2 %	15.1 %
20	4.4 %	17.6 %	44.6 %	18.7 %

and 12 bins had a specificity of 9.3% with 100% sensitivity for one feature. With two features and 100% sensitivity, eight bins had a maximum specificity of 13.7%, and 12 bins had a maximum specificity of 16.6%. In MIAS reduced, both eight bins and 12 bins had a sensitivity of 25.9% with 100% sensitivity for one feature. With two features and 100% sensitivity, eight bins had a maximum specificity of 31.7% and 12 bins had a maximum specificity of 43.4%. Therefore, 12 bin histograms were chosen for use in both these sets.

For DDSM MLO, the best number of bins was also 12. The DDSM CC set, however, had a drastic jump in specificity when going from 100% to 98.6% (one false negative) sensitivity. As shown in table 3.4, the best number of bins at 100% sensitivity does not remain the best as the sensitivity drops. For sensitivities lower than 98.6%, the best performing number of bins is 13. With 13 bins at 100% sensitivity, the specificity is 20.5%, whereas for 98.6% sensitivity, specificity jumps to 68.3%. Therefore, for the DDSM CC set, 13 bins was chosen.

Table 3.4: Naive Bayesian classifier performance by number of bins for DDSM CC. Specificity at less than 100% sensitivity for DDSM CC, one feature

Bins	100 % sensitivity	98.6 % sensitivity	97.3 % sensitivity
2	3.2 %	3.2 %	14.0 %
3	7.6 %	59.4 %	59.4 %
4	38.8 %	38.8 %	38.8 %
5	27.3 %	27.3 %	59.8 %
6	12.2 %	59.4 %	59.4 %
7	31.7 %	31.7 %	61.9 %
8	12.2 %	65.8 %	65.8 %
9	10.8 %	59.4 %	59.4 %
10	18.7 %	39.2 %	59.8 %
11	21.6 %	65.1 %	65.1 %
12	12.2 %	47.1 %	61.6 %
13	20.5 %	68.3 %	68.3 %
14	26.6 %	37.8 %	61.9 %
15	18.7 %	32.0 %	59.8 %
16	27.9 %	45.0 %	62.7 %
17	16.5 %	31.3 %	60.9 %
18	12.9 %	52.5 %	61.9 %
19	15.1 %	44.2 %	61.9 %
20	18.7 %	38.8 %	47.8 %

3.5.2 Best performing features

For the MIAS full data set, the best performing features for 100% sensitivity were number 34 for one feature classifiers, 1 and 34 for two feature classifiers, and 1, 22, and 34 for three feature classifiers. For the MIAS reduced data set, the best performing features for 100% sensitivity were number 12 for one feature classifiers, 12 and 22 for two feature classifiers, and 12, 22, and 24 for three feature classifiers. For the DDSM MLO data set, the best performing features for 100% sensitivity were number 5 for one feature classifiers, 5 and 24 for two feature classifiers, and 13, 23, and 24 for three feature classifiers. Finally, for the DDSM CC data set, the best performing features for 100% sensitivity were number 13 for one feature classifiers, 13 and 24 for two feature classifiers, and 5, 13, and 24 for three feature classifiers.

3.5.3 Feature vector size

In most cases, two feature classifiers performed better than one, and three feature classifiers performed better than two. However, there were diminishing gains with each feature added. Table 3.5 compares the best sensitivity/specificity for each of the data sets using one, two, or three features.

Table 3.5: Sensitivities and specificities for the four datasets using one, two, or three features.

		Specificity		
Set	Sensitivity	1 feature	2 features	3 features
MIAS full	100%	9.3%	16.6%	19.0%
MIAS full	94.7%	18.0%	23.9%	25.4%
MIAS full	90.5%	22.0%	33.7%	34.1%
MIAS reduced	100%	25.9%	43.4%	47.8%
MIAS reduced	95.1%	25.9%	43.9%	57.1%
MIAS reduced	90.2%	48.9%	62.9%	61.5%
DDSM MLO	100%	49.1%	53.2%	64.3%
DDSM MLO	95.7%	61.0%	71.4%	75.8%
DDSM MLO	90%	69.5%	81.4%	83.3%
DDSM CC	100%	20.5%	59.0%	62.2%
DDSM CC	94.5%	68.3%	78.1%	79.1%
DDSM CC	90.4%	74.8%	80.6%	81.7%

3.5.4 Overall accuracy

The best classified data sets were the DDSM MLO and DDSM CC. They performed far better than the original MIAS full set. Upon removing the benign images from MIAS full to form *MIAS reduced*, the set performed much better. In fact, specificity nearly tripled at 100% sensitivity. ROC curves are presented for MIAS full (figure 3.10), MIAS reduced (figure 3.11), DDSM MLO (figure 3.12), and DDSM CC (figure 3.13).

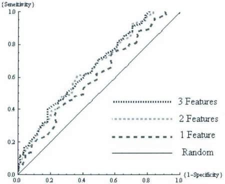


Figure 3.10: ROC curve for the naive Bayesian classifier using one, two, or three features with the MIAS database.

Solid diagonal lines indicate a random classification, red dashed lines represent single feature classifiers, green dashed lines represent double feature classifiers, and blue dotted lines represent triple feature classifiers. In all cases classification was significantly better than random.

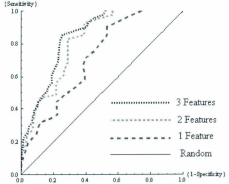


Figure 3.11: ROC curve for the naive Bayesian classifier using one, two, or three features with the MIAS reduced database.

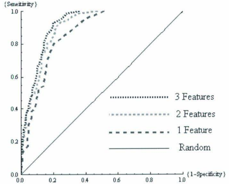


Figure 3.12: ROC curve for the naive Bayesian classifier using one, two, or three features with the DDSM MLO database.

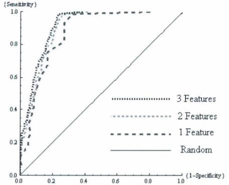


Figure 3.13: ROC curve for the naive Bayesian classifier using one, two, or three features with the DDSM CC database.

Chapter 4

Discussion

This chapter discusses the significance of the results reported in chapter 3. Section 4.1 reviews the effectiveness of the preprocessor. Section 4.2 discusses the results of the KNN classifier and the effects of the four variants of the classifier used. Section 4.3 discusses the results of the naive Bayesian classifier. Finally, section 4.4 compares the two classification algorithms and the different data sets, and summarizes the success of the features used.

4.1 Preprocessor

The preprocessor performed quite well. Only 2% - 6% of images were lost in the data sets used. The images lost were largely poor quality to begin with (figure 3.2) and therefore not appropriate for use in a training set. The preprocessor was ancillary to the main goal of the project and sufficiently accurate, so further improvements were not necessary. In a clinical setting, where images would not be used for training, they would need to be classified regardless of their quality. With some additional fine tuning, the classifier could likely be improved further in order to properly process a greater portion of poorer quality images.

Since the breast is a curved structure, the borders will have a gradual decrease in intensity as the cross section becomes thinner and x-ray attenuation decreases. The nature of the thresholding process creates a sharp border around the periphery. Where intensity decreases gradually, there will be a thin layer of dim foreground that is classified as background. This affects all images, but the effect is especially pronounced in fatty breasts. The region lost, however, is quite thin and any abnormalities this close to the surface should be plainly visible without mammography.

4.2 K nearest neighbour

Despite the long amount of computational time taken to test the classifiers, the operational phase of this classifier should take less than a second per image on the computer setup described in section 2.1. Even though the KNN classifier is a lazy learner, with only approximately 300 images in the training set, it does not take long to evaluate distances between a test image and all the training images. If more images were added to the training set, the computation time would increase.

There is a significant amount of fluctuation in the sensitivity/specificity graphs of the non-distance weighted classifiers. In the majority voting classifier (figure 3.6), all even numbers of neighbours tended to have higher sensitivity and lower specificity than odd numbers of neighbours. The reason for this is that when vote values sum to zero (equal number of suspicious and normal neighbours) the default classification is suspicious. Since vote values can only sum to zero when there is an even number of votes, there will be more images classified as suspicious with an even number of neighbours. Classifying a higher portion of images as suspicious will raise sensitivity and lower specificity. The effect is especially pronounced with lower numbers of neighbours, due to the fact that with less votes, there will be a higher chance of the values totaling zero.

There is similar fluctuation in the prior weighted votes. However in this case, the spikes are farther apart, with a slow rise in sensitivity followed by a sharp drop and a slow drop in specificity followed by a sharp rise. With prior weighting, votes for suspicious are worth more than votes for normal, so if one neighbour is suspicious, then it takes several normal neighbours to flip the classification back to normal. For example, in the MIAS full set, normal votes are worth -1 and suspicious votes are worth 2.16. It therefore takes three normal votes to overpower one suspicious vote. So, if one, two, or three neighbours are being used, it only takes one vote of suspicious to classify the image as suspicious. Four, five, or six neighbour classifiers require at least two suspicious neighbours for a suspicious classification, and so on.

Surprisingly, weighting by neighbour distance, did not have much of an effect on the accuracy of the classifiers. The graphs with distance weighting do, however, have much less fluctuation since each successive neighbour is worth less than the last. Because of this, each successive neighbour has less of a chance of changing a classification assigned by the preceding neighbours.

4.3 Naive Bayesian classifier

The times reported for testing each classifier are not representative of the time it would take to classify a single image. For example, taking roughly two days to test three feature classifiers involves generating all the histograms with between two and 20 bins for all 10660 combinations of three classifiers and testing for approximately 300 images in each set. In the operational phase, where the histograms have already been generated for the feature vector with a fixed number of bins and only one image is tested at a time, classification should take much less than one second on the computer setup described in section 2.1. Increasing the size of the training set would not have a significant impact on the time to test a single image.

100% sensitivity means both that all suspicious images are classified as suspicious and that all images classified as normal are normal. 100% specificity means both that all normal images are classified as normal and that all images classified as suspicious are suspicious. So, in our 100% sensitivity classifiers, all the images classified as normal are actually normal. In our 100% specificity classifiers, all images classified as suspicious are actually suspicious.

One of the advantages of the naive Bayesian classifier is that the detection threshold can be easily tuned to suit the needs of the individual using the program. For example, a radiologist may use the program to double check a small number of images with the highest probabilities of cancer. In this case, a high specificity would be important and a low sensitivity would be acceptable. Conversely, with some further testing, the program could be used as a prescreening step, whereby some portion of the images with a very high probability of being normal would not need to be reviewed by a human at all. In this case, a high (˜100%) sensitivity would be necessary, and lower specificity would be acceptable. Taking the results of the DDSM MLO three feature classifier, for example, this could reduce the radiologists work load by more than half.

With 100% sensitivity, for DDSM MLO with a three feature classifier, there is 64.3% specificity. This means that 173 normal images can be safely removed from the set. In a set consisting of 269 normal images and 70 suspicious images, like DDSM MLO, if one were to choose a single image at random, there would be a 1 in 1.26 chance of picking a normal. Choosing two images at random, there would be a 1 in 1.59 chance of selecting two normal images. Choosing 173 images at random, there would be a 1 in $6.70 * 10^{25}$ chance of selecting all normal images. Since all possible combinations of three features were tested in the classifier, there were 10660 classifiers tested. We could say that the feature set used in the naive Bayesian classifier is $6.3 * 10^{21}$ times better than a random guess.

An odd result found in the data was the classification of the craniocaudal view of image number 336 in the DDSM data set. This image has a very obvious irregular shaped mass with circumscribed margins. Yet for single feature classifiers, using features that work well on other sets, it was consistently assigned a very low probability of being suspicious. In fact, for classifiers that classify other images well, this image often had the lowest probability of being suspicious of any images. This is probably due to there being no images in the training set that closely resemble this one. Machine learning requires that there be training data similar to any test image used. If there are no similar images, then the classifier will not be able to accurately classify the image. Once the sensitivity threshold is dropped, the sensitivity/specificity aligns with that found for the MLO images.

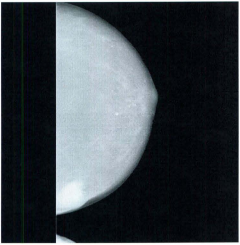


Figure 4.1: An image that was poorly classified by all the classifiers, despite a large obvious mass.[36, 37]

The MIAS full, MIAS reduced, and DDSM MLO all used 12 bins in their naive Bayesian classifiers, where as the DDSM CC used 13. It is natural for the DDSM CC set to work better with more bins since it contains the most images. More images will allow the use of more, smaller bins.

The leave one out cross validation used with the naive Bayesian classifier actually has a slight bias against correct classification built in. Using the MIAS full set as an example, there are 95 suspicious images and 300 images total. The prior probability of an image being suspicious for the entire set is 31.7%. When testing an image, it is removed from the training set. This means that when testing a suspicious image, the training set has 94 suspicious images and 299 images total. This brings the prior probability of the image being suspicious down to 31.4%. When testing a normal image from the same set, the prior probability of the image being suspicious goes up to 31.8%. So, when testing a suspicious image, there is a slightly higher chance of it being classified as normal than when testing a normal image.

4.4 Overall

The MIAS set was tested with the classifiers before the DDSM images were introduced. The algorithms were adjusted for MIAS to get the most accuracy possible. The best result was a sensitivity of 100% and a specificity of 19.0% using the naive Bayesian classifier. When the DDSM MLO set was introduced, using the exact same algorithm, there was a substantially higher accuracy rate. The best result had a sensitivity of 100% and a specificity of 64.3%. Having more data with which to train can improve the accuracy of a classifier, but the DDSM set was only slightly larger than the MIAS, so that probably did not have a significant impact. Another possibility was that the DDSM images were less subtle than the MIAS images. This would be difficult to prove quantitively, since MIAS does not grade the subtlety of abnormalities in images, and even if it did, the subtlety assigned by one radiologist may not be identical to that assigned by another. Qualitatively, I cannot see a significant difference between MIAS and DDSM images. The third possibility was that the number of calcifications in the MIAS set outnumber those in the DDSM set. Out of 95 suspicious images in MIAS, 23 of them were due to calcifications. Out of 70 suspicious images in the DDSM set, only seven of them were due to calcifications. Removing all images containing calcifications from the MIAS set had little effect on the accuracy.

Finally, in the MIAS set, 60 of the 95 suspicious images were benign, whereas in the DDSM set only 5 of the 70 suspicious images were benign. Furthermore, all five of those benign finding were contralateral to a malignant image. Removing all the benign images from the MIAS set showed a significant gain in accuracy. This gain could be due to two different effects. First of all, having benign images in the training set may blur the difference between normal and suspicious classes. If there is a difference between the feature vectors of benign and malignant images, then this problem becomes three classifications rather than two. This would require more training images to achieve a proper classification. Since the detection of benign images only leads to unnecessary testing and stress for the patient, we are better off not detecting them anyway. The second contributing factor to the increased accuracy is the change in ratio of suspicious to normal images. The MIAS full set had a 2.16:1 normal to suspicious ratio, whereas DDSM MLO had a ratio of 3.84:1. The more the distribution is skewed toward normal or suspicious, the higher the probability of correctly classifying all images. For example, in a set of 10 images, if we know one of them is suspicious, there is a 10% chance of guessing the correct classification of all images. If we know five of them are suspicious, then there is only a 0.4% chance of guessing the correct classification of all images. While the classifiers themselves do not use guessing, it does tip the chance of correct classification in our favor.

From the results of both classifiers, the ones with the highest sensitivities were chosen as the "best". While high sensitivity and low specificity is preferable to low sensitivity and high specificity, the ideal classifier would have a balance between the two. Overall accuracy is not a useful metric since there are usually many more normal images than suspicious and a classifier with a high overall accuracy could miss all suspicious images in a set as accuracy is pushed toward specificity and away from sensitivity. There is no way to define what an acceptable level of sensitivity is for a given specificity. Choosing the classifiers with the highest sensitivity just gives a consistent way of choosing and comparing the better classifiers from a set. As discussed in section 4.3, the ideal mix of sensitivity and specificity depend largely on how the system would be used.

Unless one classifier has both higher sensitivity and specificity than another, it is impossible to say which is better. Because of this, we cannot say that either the KNN or naive Bayesian classifiers was better than the other. Qualitatively, the results were quite evenly matched. The naive Bayesian classifier, however, had some favorable characteristics. As discussed in section 1.5.2 it is an eager learner. This means that it has a more efficient operational phase. Since all that needs to be taken from the training phase are two histograms per feature, storage requirements are low. Since all that is needed to classify an image is a simple equation (equation

1.15) using the histogram data, computational requirements are low. The k-nearest neighbour classifier on the other hand requires the feature values and classifications for all images in the training set to be saved, requiring more storage. Classification of a new image then requires comparison to every image in the training set before votes can be tabulated.

More importantly, is the flexibility of the naive Bayesian classifier. The threshold of what to consider normal and what to consider suspicious can be changed without having to retrain the classifier. This allows the selection of how much emphasis to place on sensitivity and how much to place on specificity. Using this, sensitivities of anywhere between 0-100% can be attained with the specificity appropriately increasing or decreasing. An adjustable KNN classifier could be developed by changing the weighting assigned to suspicious and normal neighbour votes. As seen for the prior adjusted classifier though, this produces some odd behavior in the classifier depending on the number of neighbours used.

There were several features that were consistently used by the best performing classifiers. Of the three-feature classifiers for both naive Bayesian and k-nearest neighbour, for each of the data sets (eight classifiers total), there was a lot of overlap in the features used. Feature 24 occurs in five of the eight classifiers, and features 12 or 13 occur in six classifiers. Among the 24 features used in these eight top classifiers, there were 11 unique features. The consistent high performance of these features for different data sets and different classification algorithms indicates that the high classification rates are due more to the features themselves than to the classifiers or to random chance.

Three quarters of the high performing classifiers come from regions one, two, or three (figure 2.9) in the transform. Of these 24 features, four were mean, six were standard deviation, ten were skewness, and four were kurtosis.

Chapter 5

Conclusions

The objectives of this project were all met. Whole image classification was effectively implemented and a high level of sensitivity and specificity was achieved in the various classifiers.

The use of a discrete cosine transform to separate normal from cancerous breast tissue was tested. It was found that the discrete cosine transform can indeed be used for this purpose. Partitioning the transform in square sections centred on the origin with increasing thickness for regions farther from the origin was tested. Mean, standard deviation, skewness, and kurtosis were calculated from these regions to produce a feature set. It was found that there are indeed differences in these features when comparing a normal image to a cancerous one. K-nearest neighbour and naive Bayesian classifiers were tested using these features and both provided highly accurate classifications.

The methods developed in this project could be used before, or after a radiologists reading. With 100% sensitivity, it could be used as a prescreening tool. At a digital mammography clinic, images could be assessed by the computer instantly, and results like *'no abnormalities present'*, or *'requires further review'* provided to the patient. The radiologist would then have fewer images to read, and be able to dedicate more time to the ones with higher chances of cancer. With less than 100% sensitivity, it could be used to double check the radiologist's results. Similar to a double reading, the radiologist would first read the images without knowing the outcome of the classifier so as to not be biased. Then the classifier would be applied to all the images originally classified as normal by the radiologist. The ones with the highest probability of being suspicious would then be marked for further review. In both of these cases, the radiologist spends more time analyzing suspicious images and less time on the images with a high probability of being normal.

The high sensitivity attained using the DCT generated feature set means that one of the classifiers developed in this project could be used in series with other computer aided detection methods to increase overall accuracy. Using a classifier with near 100% sensitivity, such as the one developed in this project, before applying a second classifier could only boost the accuracy of that classifier. At 100% sensitivity, no suspicious images are lost, while some (64% in the case of naive Bayesian of DDSM MLO) of the normal images can be removed. This gives the second classifier a smaller group of images that are more likely to be suspicious.

It would not be possible to accurately predict which features will perform a given classification task well and which will not. When developing a new feature set, the strategy is usually to make a hypothesis as to which will be most accurate, and then testing that hypothesis. Several of the features in the feature set developed in this project displayed an excellent ability to distinguish normal from cancerous images.

All the classifiers used feature vectors of no more than three features. Despite the small feature vector size, the results were still quite accurate. Using such a small number of features ensured that over training would not be an issue. However, before using such a system in a clinical setting, further testing on larger data sets will be necessary.

5.1 Future directions

The preprocessor algorithm could be modified to generate automated feedback as to whether or not it was successful. While manually reviewing all the images is a quick process, it does somewhat defeat the purpose of an automated system. In a clinical setting the preprocessor would either need to work on all images, or give feedback as to whether or not is was successful. Unsuccessfully preprocessed images would need to be classified as suspicious so that they would be reviewed by a human.

It has been demonstrated both in this thesis and elsewhere [58] that craniocaudal images are equal, if not better than mediolateral oblique for detecting cancers. In cases where both mediolateral oblique and craniocaudal images are available, combining the classifications of both views could boost accuracy. If one view is classified as suspicious and the other normal, then that breast would warrant further attention from a radiologist. This would increase sensitivity, as images with abnormalities that are not visible in one view could be properly classified. Specificity would also increase, since thresholds could be lowered for individual views where detection in only one view would be necessary.

While data is not currently available, it would be interesting to use the DCT features to track the changes in a patient's mammogram over time. Having a baseline feature vector available for an individual would enable the creation of a classifier that could detect the emergence of a new cancer. While mammograms are expected to change gradually with age or with slightly different positioning in a mammogram exam, a large departure from a previously recorded feature vector would likely be caused by a cancerous process.

As long as over training is carefully avoided, adding more features to the feature vector can increase the accuracy of a classifier. By performing an additional feature reduction step, many of the 41 features used could probably be eliminated. With a smaller feature set, it would become possible to test combinations of more features in the vector.

Appendix A

K-Nearest Neighbour Results

Legend:

> **Neighbours:** the number of neighbours used
>
> **f1, f2, f3:** the feature numbers used in a classifier
>
> **sens:** sensitivity
>
> **spec:** specificity

Table A.1: MIAS full non-weighted non-prior

Neighbours	f1	sens	spec	f1	f2	sens	spec	f1	f2	f3	sens	spec
1	1	42.1	73.7	5	34	46.6	64.9	3	10	19	51.6	72.2
2	38	62.1	49.8	9	17	68.4	51.2	8	29	34	69.5	51.7
3	3	34.7	76.6	13	38	36.8	76.6	13	33	35	40	77.6
4	3	48.4	66.8	8	28	51.6	64.4	8	14	29	60	63.4
5	27	30.5	85.4	12	34	35.8	83.9	8	18	29	35.8	79
6	1	41.1	66.8	12	34	44.2	73.7	2	5	39	48.4	73.2
7	27	27.4	84.9	8	29	31.6	81.5	8	14	29	33.7	79.5
8	32	32.6	73.2	9	41	37.9	79.5	8	29	39	42.1	73.7
9	3	22.1	87.3	8	29	27.4	83.4	9	15	41	29.5	87.3
10	15	31.6	79.5	8	29	33.7	80	9	17	41	34.7	82.4
11	27	18.9	91.7	12	29	23.2	91.2	12	35	39	24.2	88.8
12	5	26.3	83.9	5	32	28.4	84.4	13	24	35	31.6	82.9
13	13	14.7	87.3	5	12	20	88.8	11	30	32	23.2	90.1
14	37	21.1	16.6	11	30	24.2	87.8	24	30	35	28.4	77.6
15	1	12.6	93.2	3	5	20	91.7	11	30	32	23.2	89.3
16	37	18.9	84.9	5	23	24.2	92.2	11	30	32	24.3	87.8
17	36	12.6	93.7	11	30	20	92.7	1	3	5	23.2	90.2
18	36	15.8	89.8	5	12	24.2	91.2	3	5	12	25.3	89.8
19	36	14.7	93.7	3	5	16.8	92.2	5	12	21	21.1	94.1
20	30	16.8	92.2	3	5	24.2	90.7	5	12	21	23.2	93.7
21	36	14.7	94.6	5	12	18.9	93.2	5	12	21	21.1	94.1
22	36	14.7	92.2	2	5	22.1	91.7	1	5	32	22.1	93.7
23	36	13.7	95.1	5	12	18.9	93.2	5	12	32	18.9	93.2
24	36	14.7	92.7	5	12	22.1	92.7	1	1	12	23.2	90.7
25	36	13.7	95.6	5	12	21.1	93.7	2	2	12	20	93.2

Table A.2: MIAS reduced non-weighted non-prior

Neighbours	f1	sens	spec	f1	f2	sens	spec	f1	f2	f3	sens	spec
1	12	39	83.9	19	37	41.5	85.4	10	23	41	43.9	87.3
2	12	46.3	77.6	35	39	58.5	73.2	33	34	40	58.5	75.6
3	36	19.5	93.7	24	33	29.3	93.7	12	24	33	36.6	93.2
4	12	31.7	87.8	27	33	43.9	88.8	12	24	33	53.7	88.3
5	3	19.5	94.6	24	33	19.5	92.2	12	24	33	34.1	93.2
6	3	24.4	92.7	22	23	29.3	90.7	12	24	33	36.6	90.2
7	3	17	94.6	22	23	19.5	95.1	23	35	39	19.5	96.1
8	3	19.5	94.2	22	23	26.8	93.7	23	35	39	22	95.1
9	36	14.6	95.1	32	33	12.2	96.1	17	23	33	17.1	97.6
10	36	14.6	99	32	33	17.1	94.1	22	32	33	22	94.1
11	36	9.7	99.5	5	12	9.8	98.5	17	23	33	12.2	98
12	36	9.7	99.5	11	33	12.2	98.5	23	35	36	19.5	95.1
13	36	7.3	99.5	33	41	4.9	99	11	23	39	14.6	98.5
14	36	7.3	2	1	5	9.8	98	11	23	39	17.1	98
15				1	5	7.3	98	23	35	36	12.2	97.6
16				2	5	9.8	98	23	35	36	12.2	97.6
17				21	40	4.9	99.5	23	35	36	9.3	99
18				21	40	4.9	98.5	21	33	36	9.8	99
19				12	23	4.9	99.5	23	35	36	7.3	99.5
20				12	23	4.9	99.5	23	35	36	9.8	98
21				12	23	2.4	99.5	23	35	36	4.9	99
22				33	35	2.4	98	23	35	39	4.9	97.1
23								23	35	39	2.4	98.5
24								23	35	39	2.4	97.6
25								23	35	39	2.4	99.5

Table A.3: DDSM MLO non-weighted non-prior

Neighbours	f1	sens	spec	f1	f2	sens	spec	f1	f2	f3	sens	spec
1	13	48.6	82.5	5	6	57.1	84.4	5	6	33	61.4	87
2	13	65.7	72.1	5	6	80	78.4	5	17	30	82.9	78.8
3	13	37.1	90	24	38	57.1	90.7	5	6	37	61.4	87.4
4	13	54.3	83.6	5	6	71.4	83.6	5	15	23	75.7	84
5	5	40	88.9	5	6	55.7	88.1	5	6	37	58.6	86.2
6	13	54.3	82.5	13	40	64.3	85.5	5	6	37	72.9	83.3
7	5	37.1	89.6	5	6	55.7	88.8	6	23	25	61.4	88.1
8	22	44.3	89.2	5	6	62.9	87.3	6	23	25	68.6	86.2
9	23	34.3	93.7	5	6	51.4	89.2	5	15	36	57.1	93.3
10	13	45.7	84.8	16	24	57.1	89.2	13	19	37	64.3	88.1
11	13	37.1	88.1	13	36	48.6	88.5	24	25	37	52.9	89.2
12	13	50	84.8	13	37	57.1	86.2	5	6	28	62.9	86.2
13	13	35.7	88.5	13	22	50	90	13	35	37	52.9	89.6
14	13	51.4	91.6	13	36	60	87.7	5	13	37	61.4	86.6
15	13	38.6	84.8	13	36	47.1	88.8	5	13	36	52.9	89.9
16	13	55.7	84	13	20	58.6	88.1	13	22	35	58.6	90.3
17	13	38.6	85.9	13	35	51.4	90.3	5	13	36	54.3	89.2
18	13	52.9	84	13	36	54.3	87.4	13	16	36	58.6	88.1
19	13	40	87	13	35	51.4	91.1	24	25	37	54.3	91.4
20	13	54.3	85.5	13	36	54.3	87.7	13	22	35	58.6	89.2
21	36	38.6	90.7	13	35	51.4	91.4	13	16	36	54.3	88.1
22	13	50	84.8	13	37	58.6	85.9	13	22	35	58.6	88.5
23	36	37.1	90.7	13	35	50	91.4	13	22	35	54.3	90
24	13	52.9	85.9	13	36	52.9	88.1	13	22	35	57.1	89.2
25	13	48.6	88.5	13	35	48.6	91.8	4	13	16	51.4	88.8

Table A.4: DDSM CC non-weighted non-prior

Neighbours	f1	sens	spec	f1	f2	sens	spec	f1	f2	f3	sens	spec
1	32	56.2	84.2	32	39	57.5	84.9	24	26	41	60.3	88.8
2	24	78.1	77.7	3	5	79.5	80.6	2	17	32	83.6	80.9
3	24	49.3	87.4	32	37	54.8	88.5	6	19	24	61.6	88.8
4	24	58.9	82	4	26	78.1	84.5	15	24	26	75.3	81.7
5	32	52.1	91.4	6	24	54.9	88.5	6	24	32	63	89.2
6	4	65.8	84.9	6	24	74	86.3	15	24	26	75.3	83.1
7	4	49.3	87.4	6	24	57.5	88.5	6	24	28	63	87.1
8	4	58.9	84.5	3	24	68.5	81.7	24	31	35	74	84.5
9	32	47.9	90.3	3	24	57.5	85.6	6	24	28	63	86
10	4	56.2	80.6	3	24	69.9	83.1	6	24	28	74	82
11	32	52.1	91	24	26	54.8	87.4	6	24	35	61.6	86.3
12	32	53.4	89.9	6	24	64.4	85.3	3	24	28	72.6	81.7
13	22	42.5	88.1	22	33	53.4	89.2	3	24	28	61.6	82.4
14	32	52.1	11.5	24	31	65.8	84.9	2	24	25	72.6	85.3
15	32	46.6	90.3	22	33	54.8	89.6	24	25	34	65.8	86
16	32	49.3	88.8	24	31	63	85.3	24	25	34	68.5	86
17	32	47.9	92.1	24	31	54.8	87.1	24	25	34	61.6	87.1
18	32	50.7	89.9	24	31	61.6	83.8	3	19	24	68.5	82.4
19	32	47.9	90.6	15	24	54.8	87.8	24	25	34	61.6	87.8
20	22	52.1	85.6	6	24	61.6	83.1	6	24	35	69.9	84.5
21	32	43.8	91.4	6	24	54.2	85.6	6	24	35	58.9	87.1
22	22	54.8	87.1	6	24	60.3	83.8	6	24	35	71.2	84.2
23	32	43.8	90.6	6	24	52.1	86	9	24	28	58.9	87.1
24	32	46.6	86.7	6	24	60.3	84.9	6	24	35	65.7	84.5
25	32	46.6	89.6	22	40	52.1	89.9	9	24	28	57.5	87.4

Table A.5: MIAS full distance weighted non-prior

Neighbours	f1	sens	spec	f1	f2	sens	spec	f1	f2	f3	sens	spec
1	1	42.1	73.7	5	34	46.3	64.9	3	10	19	51.6	72.2
2	1	42.1	73.7	5	34	46.3	64.9	3	10	19	51.6	72.2
3	12	38.9	71.7	18	37	40	74.1	12	24	29	41.1	74.6
4	12	38.9	73.2	8	28	40	78	11	16	33	44.2	82.4
5	12	36.8	73.2	12	34	42.1	81	8	28	38	38.9	81
6	12	36.8	74.6	21	41	36.8	84.4	8	14	29	37.9	75.6
7	12	37.9	75.1	5	25	33.7	79.5	11	16	33	35.8	86.8
8	12	37.9	75.1	30	41	34.7	80.4	8	14	29	33.7	78.5
9	12	37.9	76.6	30	41	31.6	81	12	30	41	32.6	84.9
10	12	36.8	77.1	11	32	30.5	85.4	21	30	41	23.5	82
11	12	35.8	77.1	30	41	31.6	82.4	9	15	41	28.4	85.9
12	12	34.7	77.1	30	41	30.5	83.4	21	30	41	27.4	85.4
13	12	35.8	77.6	30	41	27.4	83.4	9	15	41	27.4	87.3
14	12	32.6	78	30	41	25.3	83.9	16	30	41	26.3	85.4
15	12	32.6	78	30	41	26.3	83.9	11	17	18	26.3	92.2
16	12	31.6	79	30	41	25.3	84.4	11	17	18	25.3	92.2
17	12	33.7	79	35	39	22.1	88.3	11	17	18	25.3	91.7
18	12	31.6	79	30	41	22.1	86.8	11	17	18	22.1	91.2
19	12	31.6	79	16	30	21.1	89.3	17	35	36	21.1	93.7
20	12	32.6	80	16	30	22.1	88.8	3	11	22	21.1	95.6
21	12	32.6	80	16	30	21.1	90.7	11	16	22	20	91.2
22	12	32.6	80.5	35	39	21.1	89.3	3	11	22	20	96.1
23	12	31.6	80	3	11	20	94.6	11	16	22	18.9	92.7
24	12	30.5	80.5	3	11	20	94.6	11	16	22	20	93.2
25	12	30.5	80.5	3	11	18.9	94.6	11	16	22	18.9	93.7

Table A.6: MIAS reduced distance weighted non-prior

Neighbours	f1	sens	spec	f1	f2	sens	spec	f1	f2	f3	sens	spec
1	12	39	83.9	19	37	41.5	85.4	10	23	41	43.9	87.3
2	12	39	83.9	19	37	41.5	85.4	10	23	41	43.9	87.3
3	12	34.1	87.8	18	37	24.1	90.2	12	24	33	36.6	91.7
4	12	34.1	88.8	27	33	31.7	90.7	12	24	33	39	93.2
5	12	34.1	87.3	12	37	26.8	95.6	11	33	40	34.1	95.6
6	12	34.1	88.8	21	22	26.8	91.2	13	22	23	26.8	94.6
7	12	34.1	88.3	16	22	22	94.6	12	24	33	26.8	94.6
8	12	29.3	89.3	16	22	22	95.6	11	23	33	22	98
9	12	29.3	89.8	15	40	22	98	11	33	40	19.5	97.6
10	12	29.3	91.2	16	22	22	96.6	11	23	33	19.5	98.5
11	12	29.3	90.7	16	22	19.5	97.1	11	33	40	19.5	98
12	12	26.9	91.2	16	22	19.5	97.6	11	33	39	17.1	99
13	12	26.8	91.7	16	22	17.1	98	11	33	39	17.1	99
14	12	24.4	91.7	16	22	17.1	98.5	11	18	23	14.6	98.5
15	12	19.5	91.7	16	22	17.1	98.5	23	35	36	12.2	98
16	33	17.1	92.7	16	22	17.1	98.5	11	23	39	12.2	99
17	33	17.1	92.7	16	22	17.1	98.5	11	23	39	12.2	99
18	33	17.1	92.7	16	22	17.1	98.5	11	19	23	12.2	99.5
19	33	17.1	92.7	16	22	17.1	99.5	11	17	23	12.2	99
20	33	17.1	92.7	16	22	17.1	99.5	23	35	36	9.8	98.5
21	33	17.1	92.7	16	22	14.6	99.5	11	17	23	12.2	100
22	33	17.1	92.7	16	22	14.6	99.5	23	35	36	9.8	98.5
23	33	17.1	92.7	16	22	14.6	100	11	13	23	12.2	99.5
24	33	17.1	92.7	16	22	14.6	100	23	35	36	9.9	98.5
25	33	17.1	93.2	16	22	14.6	100	23	35	38	9.9	99

Table A.7: DDSM MLO distance weighted non-prior

Neighbours	f1	sens	spec	f1	f2	sens	spec	f1	f2	f3	sens	spec
1	13	48.6	82.5	5	6	57.1	84.4	5	6	33	61.4	87
2	13	48.6	82.5	5	6	57.1	84.4	5	6	33	61.4	87
3	13	45.7	84.8	24	38	55.7	91.1	5	6	24	65.7	85.1
4	13	45.7	85.5	5	6	58.6	87.7	5	6	21	62.9	88.8
5	13	45.7	84.4	5	6	57.1	88.1	25	33	35	58.6	91.1
6	13	44.3	84.8	5	6	57.1	87.7	25	33	35	60	90.3
7	13	44.3	85.5	5	6	55.7	88.8	25	33	35	58.6	91.1
8	13	44.3	84.8	5	6	57.1	87.7	5	15	20	60	89.2
9	13	42.9	84.8	5	6	54.3	88.1	5	15	19	58.6	91.4
10	13	44.3	84.4	5	6	54.3	88.1	5	15	36	55.7	91.8
11	13	42.9	85.1	5	6	54.3	88.1	5	15	36	55.7	92.2
12	13	40	84.8	5	6	50	88.8	24	25	37	54.3	90.3
13	13	40	85.5	13	37	48.6	88.5	5	15	36	54.3	91.4
14	13	40	85.5	5	6	50	88.1	3	5	15	55.7	89.2
15	13	41.4	85.5	22	24	48.6	90	3	5	15	54.3	88.8
16	13	40	85.5	22	24	47.1	90.3	3	5	15	55.7	90
17	13	41.4	85.9	13	15	48.6	90	5	13	36	51.4	89.6
18	13	42.9	84.8	13	41	47.1	91.8	5	13	36	55.7	89.6
19	13	41.4	86.2	13	41	47.1	92.6	13	35	37	52.9	90.3
20	13	41.4	85.1	22	24	47.1	90.3	3	5	15	52.9	89.2
21	13	42.9	86.2	13	36	47.1	89.6	3	5	15	54.3	88.5
22	13	42.9	86.6	5	15	48.6	91.1	24	25	35	50	92.2
23	13	41.4	86.6	13	36	47.1	90.3	13	22	35	52.9	89.6
24	13	41.4	87	5	15	48.6	91.4	13	22	35	50	90
25	13	41.4	87	13	20	48.6	90.7	17	24	25	50	92.2

Table A.8: DDSM CC distance weighted non-prior

Neighbours	f1	sens	spec	f1	f2	sens	spec	f1	f2	f3	sens	spec
1	32	56.2	84.2	32	39	57.5	84.9	24	26	41	60.3	88.8
2	32	56.2	84.2	32	39	57.5	84.9	24	26	41	60.3	88.8
3	32	53.4	87.1	4	18	54.8	87.4	19	20	24	57.5	87.1
4	32	52.1	88.5	4	26	54.8	87.4	24	25	26	58.9	89.2
5	32	52.1	89.2	32	37	53.4	90.6	6	24	32	61.6	89.6
6	32	52.1	88.8	6	24	53.4	88.5	24	25	26	58.9	88.8
7	32	52.1	89.9	32	37	53.4	91	24	31	35	58.9	86
8	32	50.7	89.2	32	37	50.7	90.6	24	31	35	58.9	87.1
9	32	49.3	89.2	32	37	52.1	89.9	24	26	35	57.5	86.3
10	32	49.3	89.2	32	37	52.1	89.6	16	24	31	58.9	88.5
11	32	49.3	90.3	32	37	52.1	90.6	14	24	31	57.5	86.3
12	32	50.7	89.9	32	37	53.4	90.3	24	31	35	60.3	88.1
13	32	49.3	89.9	32	37	53.4	90.3	16	24	24	57.5	86.7
14	32	50.7	90.3	24	31	50.7	86.7	24	31	31	57.5	86.3
15	32	49.3	90.7	32	37	50.7	90.6	24	25	25	57.5	87.4
16	32	49.3	90.3	22	37	52.1	90.3	3	4	15	58.9	84.1
17	32	47.9	90.3	18	24	53.4	88.4	19	24	25	56.2	88.1
18	32	49.3	90.3	22	37	52.1	90.3	3	4	15	58.9	84.2
19	32	49.3	90.3	22	37	52.1	88.8	15	24	31	57.5	83.8
20	32	50.7	89.9	22	37	53.4	89.9	6	24	35	58.9	86
21	32	49.3	88.8	18	24	50.7	88.5	24	31	38	56.2	86
22	32	50.7	89.2	22	37	50.7	90.6	3	4	15	58.9	83.8
23	32	50.7	88.8	18	24	49.3	90.3	3	4	15	57.5	84.9
24	32	50.7	89.9	22	37	49.3	89.6	3	4	15	58.9	83.4
25	32	50.7	89.6	15	22	50.7	89.9	15	24	31	56.2	85.3

Table A.9: MIAS full non-weighted prior adjusted

Neighbours	f1	sens	spec	f1	f2	sens	spec	f1	f2	f3	sens	spec
1	1	42.1	73.7	5	34	46.3	64.9	3	10	19	51.6	72.2
2	38	62.1	49.8	9	17	68.4	51.2	8	29	34	69.5	51.7
3	8	77.9	32.2	13	27	80	37.1	13	36	38	83.2	34.6
4	3	48.4	66.8	8	28	51.6	64.4	8	14	29	60	63.4
5	8	63.2	45.4	15	35	64.2	46.3	8	14	29	70.5	53.7
6	1	75.8	44.9	7	40	74.7	33.7	8	15	40	77.9	40
7	32	53.7	58.5	9	16	54.7	65.9	8	9	29	60	54.6
8	3	61.1	52.2	8	29	64.2	53.7	8	27	29	67.4	51.7
9	3	68.4	39.5	8	29	72.6	50.7	8	19	29	77.9	41.4
10	32	51.6	56.1	8	29	54.7	61.5	13	30	41	61.1	61
11	32	58.9	50.7	13	38	63.2	50.7	8	19	29	69.5	47.3
12	18	69.5	43.4	8	29	72.6	46.3	8	19	29	75.8	41
13	32	53.7	59	8	29	58.9	60.5	13	30	41	60	62
14	32	63.2	54.6	8	15	63.2	41.5	24	33	39	64.2	48.8
15	32	67.4	46.3	13	38	68.4	52.7	13	19	30	72.6	46.8
16	38	55.8	59.5	13	38	57.8	56.6	24	33	39	63.2	52.2
17	38	63.2	53.7	13	38	64.2	53.2	20	30	39	66.3	54.6
18	38	67.4	47.8	13	38	61.1	48.8	21	24	39	73.7	36.6
19	27	56.8	59	30	41	56.8	56.1	16	30	41	62.1	58
20	4	60	45.4	6	31	62.1	55.6	12	30	41	65.3	47.3
21	4	69.5	42	4	32	69.5	39	29	30	31	71.6	44.4
22	4	74.7	36.1	1	34	75.8	40	8	29	30	76.8	26.8
23	27	58.9	55.6	12	37	64.2	47.8	12	30	40	65.3	53.2
24	24	68.4	41.5	12	37	69.5	44.4	12	36	38	71.6	40.5
25	24	70.5	40.5	24	40	74.7	30.7	19	24	40	76.8	31.2

Table A.10: MIAS reduced non-weighted prior adjusted

Neighbours	f1	sens	spec	f1	f2	sens	spec	f1	f2	f3	sens	spec
1	12	39	83.9	19	37	41.5	85.4	10	23	41	43.9	87.3
2	12	46.3	77.6	35	39	58.5	73.2	33	34	40	58.5	84.6
3	12	63.4	68.3	1	12	70.7	66.3	1	12	37	73.2	60
4	12	65.9	62.4	3	12	80.5	56.1	1	12	20	80.5	57.6
5	24	75.6	38	3	12	85.4	47.8	13	37	40	87.8	49.8
6	25	80.5	40	33	37	87.8	39.5	15	21	39	92.7	39
7	12	65.9	75.1	22	33	63.4	66.8	21	22	38	68.3	64.4
8	12	68.3	64.4	22	33	65.9	64.4	6	18	41	75.6	61
9	12	78	59	22	33	70.7	61	13	34	40	85.4	54.1
10	12	78	51.2	13	38	78	51.2	13	34	40	87.8	49.8
11	12	80.5	42.9	13	38	82.9	44.4	13	34	40	90.2	44.9
12	12	80.5	38.6	1	12	87.8	44.4	15	21	36	90.2	36.1
13	32	63.4	64.4	4	33	68.3	67.8	4	21	33	75.6	67.3
14	32	65.9	61	24	35	68.3	51.2	4	33	40	78	61.5
15	32	68.3	56.1	4	12	78	51.2	15	30	40	80.5	49.3
16	32	68.3	50.2	23	34	80.5	53.7	6	7	31	82.9	48.3
17	33	73.2	52.2	4	12	82.9	47.8	12	32	38	87.8	42.4
18	33	78	47.3	12	37	70.7	43.4	12	37	39	90.2	40
19	37	65.9	59.5	22	40	68.3	56.6	12	37	39	78	51.2
20	37	65.9	58	21	37	73.2	46.3	12	37	39	78	50.2
21	4	68.3	57.6	12	37	82.9	49.8	21	37	38	82.9	42
22	37	70.7	55.6	12	37	82.9	47.3	12	19	39	87.8	38
23	1	75.6	50.2	12	37	82.9	43.9	12	21	39	92.7	32.2
24	24	75.6	34.6	12	39	90.2	33.7	12	34	40	92.7	35.6
25	37	68.3	58	24	40	73.2	44.4	12	37	39	80.5	46.3

Table A.11: DDSM MLO non-weighted prior adjusted

Neighbours	f1	sens	spec	f1	f2	sens	spec	f1	f2	f3	sens	spec
1	13	48.6	82.5	5	6	57.1	84.3	5	6	33	61.4	87
2	13	65.7	72.1	5	6	80	78.4	5	17	30	82.9	78.8
3	24	77.1	68.4	5	6	87.1	73.2	13	16	18	90	71.4
4	24	88.6	63.2	5	35	92.9	69.5	5	12	15	95.7	72.9
5	13	72.9	79.2	13	35	75.7	82.2	5	6	15	82.9	81.8
6	13	80	77	11	24	85.7	78.1	24	27	28	88.6	74.3
7	13	81.4	76.2	11	24	88.6	75.8	11	24	26	92.9	72.5
8	24	87.1	68	4	5	92.9	67.3	4	5	38	95.7	67.7
9	24	90	66.5	4	5	94.3	67.3	11	24	28	95.7	66.2
10	13	78.6	78.1	13	17	87.1	75.5	6	24	28	92.9	70.6
11	4	82.9	61.7	4	5	91.4	68.4	11	24	28	94.3	65.8
12	4	85.7	57.2	4	5	92.9	68	11	24	28	97.1	63.9
13	4	90	54.3	4	5	94.3	67.3	11	24	28	98.6	63.2
14	4	92.9	53.5	4	31	95.7	56.5	11	24	28	98.6	62.5
15	4	91.4	58	5	6	91.4	71.4	6	24	29	94.3	65.8
16	4	91.4	56.9	24	31	91.4	63.9	24	28	30	94.3	64.3
17	4	94.3	56.1	24	29	92.9	68.8	4	29	31	95.7	53.9
18	4	94.3	55.8	4	31	97.1	55.8	4	5	11	98.6	58
19	4	94.3	54.6		31	97.1	54.6	4	7	17	98.6	58
20	4	94.3	56.9	4	31	94.3	54.5	4	31	38	97.1	57.6
21	4	94.3	55	4	38	95.7	57.6	4	31	38	97.1	56.9
22	4	94.3	52.4	4	38	97.1	54.3	4	7	15	98.6	60.2
23	4	97.1	52.4	4	41	97.1	53.9	4	11	28	98.6	53.5
24	4	97.1	52.4	4	41	98.6	53.2	4	28	31	98.6	54.6
25	4	97.1	55	4	38	97.1	55.8	4	7	18	98.6	58

Table A.12: DDSM CC non-weighted prior adjusted

Neighbours	f1	sens	spec	f1	f2	sens	spec	f1	f2	f3	sens	spec
1	32	56.2	84.2	32	39	57.5	48.9	24	26	41	60.3	88.8
2	24	78.1	77.7	3	5	79.5	80.6	2	17	32	83.6	80.9
3	24	84.9	70.9	24	40	90.4	73	24	33	40	93.2	74.1
4	24	89	68.3	24	40	95.9	69.8	20	24	37	97.3	69.8
5	24	74	78.4	24	40	83.6	77.3	5	24	35	87.7	75.9
6	24	80.8	74.8	24	37	93.2	73.4	7	15	24	93.2	75.9
7	24	89	72.7	7	24	94.5	74.1	22	24	38	95.9	71.9
8	4	91.8	71.6	2	24	97.3	71.2	24	38	39	97.3	71.2
9	24	93.2	71.6	24	37	97.3	70.9	1	22	38	98.6	73.2
10	4	90.4	70.9	24	37	95.9	71.2	17	24	32	97.3	71.2
11	4	91.8	70.9	16	24	97.3	70.9	24	28	39	97.3	70.9
12	24	95.9	70.9	24	38	97.3	70.9	24	37	38	97.3	70.9
13	24	95.9	70.9	24	38	97.3	70.9	24	37	38	97.3	70.9
14	24	97.3	70.9	24	39	97.3	70.9	5	24	24	98.6	70.9
15	24	97.3	70.9	24	38	97.3	70.9	24	35	39	97.3	72.7
16	24	97.3	70.9	24	38	97.3	70.9	24	35	40	97.3	71.6
17	24	97.3	70.9	24	38	97.3	70.9	24	38	39	97.3	70.9
18	24	97.3	70.9	24	39	97.3	70.9	1	22	38	98.6	73.4
19	24	97.3	70.9	24	41	97.3	70.9	5	24	25	98.6	69.8
20	24	97.3	70.9	24	38	97.3	70.9	24	40	41	97.3	70.9
21	24	97.3	70.9	24	38	97.3	70.9	24	40	41	97.3	70.9
22	24	97.3	70.9	24	41	97.3	70.9	24	40	41	97.3	70.9
23	24	97.3	70.9	24	41	97.3	70.9	2	3	38	98.6	75.2
24	24	97.3	70.9	24	41	97.3	70.9	2	3	38	98.6	74.1
25	24	97.3	70.9	24	41	97.3	70.9	24	40	41	97.3	70.9

Table A.13: MIAS full distance weighted, prior adjusted

Neighbours	f1	sens	spec	f1	f2	sens	spec	f1	f2	f3	sens	spec
1	1	42.1	73.7	5	34	46.3	64.9	3	10	19	51.6	72.2
2	3	50.5	53.7	27	37	62.1	51.7	3	11	14	69.5	52.2
3	38	49.5	65.9	24	29	58.9	57.1	8	23	34	63.2	53.7
4	38	50.5	62.9	12	24	55.8	53.2	8	14	29	62.1	59.5
5	3	52.6	59	34	37	60	58	8	29	41	68.4	53.7
6	34	52.6	54.1	34	37	57.9	59.5	8	29	41	66.3	54.1
7	24	49.5	60	34	37	64.2	58.5	8	14	29	66.3	56.1
8	34	50.5	54.6	13	41	60	59.5	23	34	37	66.3	50.2
9	1	51.6	64.9	5	6	61.1	60	8	29	41	67.4	49.3
10	38	49.5	64.9	8	29	58.9	55.1	8	29	41	66.3	49.3
11	34	53.7	54.6	30	38	60	58	13	30	38	68.4	52.7
12	34	52.6	53.7	30	38	61.1	56.6	13	30	38	68.4	53.2
13	34	52.6	53.2	34	37	63.2	57.1	8	14	29	67.4	57.1
14	34	52.6	52.7	34	37	61.1	55.1	8	14	29	66.3	54.6
15	34	52.6	53.2	30	38	62.1	54.6	8	30	41	66.3	44.9
16	34	52.6	53.2	24	39	62.1	49.3	8	14	29	68.4	52.7
17	38	51.6	66.3	24	39	62.1	48.8	24	32	39	66.3	43.4
18	38	50.5	64.9	34	37	62.1	55.1	8	30	41	65.3	43.9
19	12	51.6	57.6	34	37	62.1	55.6	24	32	40	65.3	44.9
20	38	50.5	64.9	34	37	63.2	54.6	8	23	29	67.4	46.3
21	34	50.5	53.2	34	37	64.2	54.1	8	23	29	67.4	42
22	38	50.5	65.4	34	37	63.2	52.7	8	23	29	66.3	44.4
23	38	50.5	65.4	30	38	62.1	53.7	23	34	37	65.3	52.7
24	34	50.5	52.7	13	38	52.7	52.7	12	36	38	66.3	45.4
25	34	52.6	52.7	34	37	53.1	53.2	21	30	38	66.3	53.7

Table A.14: MIAS reduced distance weighted, prior adjusted

Neighbours	f1	sens	spec	f1	f2	sens	spec	f1	f2	f3	sens	spec
1	12	39	83.9	19	37	41.5	85.4	10	23	41	43.9	87.3
2	12	46.3	78	35	39	58.5	74.1	33	34	40	58.5	75.6
3	12	53.7	76.1	35	39	65.9	63.9	12	32	33	70.7	62.9
4	12	53.7	75.6	3	12	73.2	64.4	12	32	33	78	57.6
5	12	56.1	74.6	1	12	73.2	62	13	23	41	75.6	58.5
6	12	56.1	74.1	3	12	65.9	65.9	12	24	37	73.2	64.4
7	12	61	74.1	1	12	70.7	60.5	12	24	37	73.2	62.9
8	12	63.4	74.6	1	12	73.2	59.5	12	33	34	75.6	57.6
9	12	63.4	75.1	1	12	73.2	59	13	34	40	78	57.1
10	12	63.4	75.6	1	16	70.7	61	12	23	35	78	58.5
11	12	63.4	74.6	1	16	70.7	61.5	12	21	37	78	52.7
12	12	63.4	72.7	1	16	73.2	61.5	21	24	30	75.6	56.1
13	12	65.9	72.7	25	33	70.7	55.6	12	24	30	80.5	56.1
14	12	65.9	71.7	25	33	70.7	54.6	12	24	30	80.5	54.1
15	12	65.9	71.7	21	33	70.7	59.5	12	24	34	80.5	45.6
16	12	65.9	71.7	1	12	73.2	58.5	1	24	30	80.5	55.1
17	12	63.4	71.2	1	12	73.2	57.1	13	14	40	80.5	58
18	12	61	71.2	4	21	75.6	58.5	4	12	38	82.9	54.6
19	12	58.5	71.2	25	33	73.2	56.1	4	33	40	82.9	62.4
20	12	56.1	70.7	25	33	70.7	56.6	4	33	40	82.9	63.4
21	12	58.5	72.2	25	33	70.7	56.1	12	24	34	80.5	51.7
22	12	53.7	71.2	6	31	70.7	58	4	12	38	82.9	57.1
23	12	56.1	71.7	4	21	73.2	58	15	30	41	80.5	51.7
24	12	56.1	71.2	4	21	73.2	57.1	4	12	38	85.4	55.1
25	12	56.1	70.7	4	21	73.2	58	19	30	41	80.5	51.2

Table A.15: DDSM MLO distance weighted, prior adjusted

Neighbours	f1	sens	spec	f1	f2	sens	spec	f1	f2	f3	sens	spec
1	13	48.6	82.5	5	6	57.1	84.4	5	6	33	61.4	87
2	13	64.2	74.3	5	6	80	79.2	5	16	30	82.9	79.9
3	24	68.6	74.7	5	6	84.3	75.1	6	24	36	88.6	78.1
4	13	67.1	72.5	5	6	85.7	74.3	5	6	12	91.4	77.3
5	13	70	72.5	5	6	85.7	76.6	12	13	38	88.6	75.5
6	13	71.4	73.2	5	6	85.7	75.1	20	24	35	88.6	77.3
7	13	72.9	73.6	15	37	84.3	74	16	24	26	92.9	75.8
8	13	75.7	73.6	13	20	84.7	72.5	6	24	25	92.9	75.1
9	13	75.7	74	5	6	87.1	74.7	5	6	13	94.3	77
10	13	75.7	73.6	24	29	87.1	72.5	7	16	24	94.3	72.1
11	13	75.7	73.6	24	29	90	72.5	24	26	36	94.3	74
12	13	77.1	74	24	29	91.4	72.5	24	27	30	94.3	67.3
13	13	78.6	73.2	4	7	91.4	64.3	6	24	29	95.7	67.7
14	13	78.6	73.2	4	7	91.4	63.6	6	7	25	97.1	65.4
15	13	78.6	74.3	24	29	91.4	72.5	8	11	24	98.6	65.7
16	13	78.6	74.3	4	7	91.4	62.8	8	11	24	98.6	63.9
17	13	78.6	74.3	24	29	91.4	71.7	8	11	24	98.6	63.6
18	13	78.6	74.3	4	16	91.4	69.5	11	24	28	98.6	63.2
19	13	78.6	74.3	24	29	92.9	70.3	11	24	28	97.1	63.2
20	13	78.6	74.3	24	29	92.9	69.1	11	24	28	97.1	63.6
21	13	78.6	74.7	24	29	92.9	69.9	4	23	31	97.1	58.4
22	13	78.6	74.7	4	31	94.3	58	4	27	31	97.1	58.4
23	13	78.6	74.3	24	29	91.4	68	4	31	37	97.1	58.4
24	13	78.6	74.3	4	31	92.9	57.2	4	8	30	98.6	53.2
25	13	78.6	74	4	7	95.7	58.4	9	24	28	97.1	63.2

Table A.16: DDSM CC distance weighted, prior adjusted

Neighbours	f1	sens	spec	f1	f2	sens	spec	f1	f2	f3	sens	spec
1	32	56.2	84.2	32	39	57.5	84.9	24	26	41	60.3	88.8
2	24	71.2	79.9	3	5	76.7	80.6	2	17	32	83.6	80.9
3	24	74	77.3	4	24	83.6	74.1	15	24	26	93.2	75.2
4	24	75.3	77	24	40	84.9	86.3	24	33	40	91.8	75.2
5	24	71.2	77.3	24	26	84.9	75.5	15	24	26	91.8	74.8
6	24	75.3	76.3	24	37	90.4	73.4	15	24	26	94.5	74.8
7	24	78.1	76.6	24	37	90.4	74.5	24	29	30	94.5	71.9
8	24	76.7	76.3	24	37	93.2	73.7	24	30	35	95.9	71.2
9	24	78.1	76.3	24	37	93.2	73.7	24	30	35	95.9	70.5
10	24	78.1	76.3	24	37	94.5	72.7	14	16	24	97.3	71.6
11	24	76.7	76.3	24	29	94.5	70.5	16	24	37	97.3	71.9
12	24	79.5	76.3	24	29	94.5	70.9	24	26	41	97.3	75.9
13	4	79.5	76.3	24	35	95.9	72.3	24	29	36	97.3	70.5
14	24	79.5	75.5	8	24	97.3	71.2	24	34	35	97.3	71.6
15	24	79.5	75.5	24	35	97.3	72.3	24	34	35	97.3	71.2
16	24	79.5	75.5	24	35	97.3	72.3	24	35	38	97.3	71.9
17	4	80.8	74.8	24	29	97.3	70.5	24	35	38	97.3	71.9
18	4	80.8	74.8	24	35	97.3	72.3	24	36	37	97.3	71.6
19	4	80.8	74.5	24	35	97.3	71.6	24	36	37	97.3	70.9
20	4	82.2	74.1	24	35	97.3	71.6	24	36	40	97.3	70.9
21	24	82.2	74.8	24	35	97.3	71.6	24	38	39	97.3	70.9
22	24	82.2	74.8	24	35	97.3	70.9	24	40	41	97.3	72.7
23	4	83.6	74.5	24	38	97.3	71.6	24	40	41	97.3	72.7
24	4	83.6	74.5	24	35	97.3	70.9	24	39	40	97.3	70.9
25	4	83.6	74.5	24	37	97.3	71.6	24	40	41	97.3	71.9

Bibliography

[1] Donald A. Berry, Kathleen A. Cronin, Sylvia K. Plevritis, Dennis G. Fryback, Lauren Clarke, Marvin Zelen, Jeanne S. Mandelblatt, Andrei Y. Yakovlev, J. Dik F. Habbema, , and Eric J. Feuer. Effect of screening and adjuvant therapy on mortality from breast cancer. *The New England Journal of Medicine*, 353(17):1784–1792, 2005.

[2] Canadian Cancer Society/National Cancer Institute of Canada. *Canadian Cancer Statistics*, 2007.

[3] Joan Austoker. Breast self examination. *BMJ*, 326(7379):1–2, 2003.

[4] JP Kösters and PC Gøtzsche. Regular self-examination or clinical examination for early detection of breast cancer. *Cochrane Database of Systematic Reviews*, (2), 2003.

[5] Wendie A. Berg, Lorena Gutierrez, Moriel S. NessAiver, W. Bradford Carter, Mythreyi Bhargavan, Rebecca S. Lewis, and Olga B. Ioffe. Diagnostic accuracy of mammography, clinical examination, US, and MR imaging in preoperative assessment of breast cancer. *Radiology*, 233(3):830–849, 2004.

[6] Daniel B. Kopans. *Breast Imaging*. Lippincott Williams & Wilkins, third edition, 2007.

[7] Patrick J. Lynch. Breast. `http://commons.wikimedia.org/wiki/File:Breast.svg`, 2009. [Online; accessed 15-June-2010].

[8] Jennifer E. Rusby, Elena F. Brachtel, James S. Michaelson, Frederick C. Koerner, and Barbara L. Smith. Breast duct anatomy in the human nipple: three-dimensional patterns and clinical implications. *Breast Cancer Research and Treatment*, 106(2):171–179, 2007.

[9] Jay R. Harris. *Diseases of the breast*. Lippincott Williams & Wilkins, fourth edition, 2010.

[10] Christie R. Eheman, Kate M. Shaw, Aliza Blythe Ryerson, Jacqueline W. Miller, Umed A. Ajanil, and Mary C. White. The changing incidence of in situ and invasive ductal and lobular breast carcinomas: United States, 1999-2004. *Cancer Epidemiology, Biomarkers & Prevention*, 18(6):1763–1769, 2009.

[11] Allan Lipton. Hormonal influences on oncogenesis and growth of breast cancer. In Daniel F. Roses, editor, *Breast Cancer*. Elsevier Inc., second edition, 2005.

[12] Allan Lipton, Laurence Demers, Kim Leitzel, Suhail M. Ali, Rainer Neumann, Christopher P. Price, and Walter P. Carney. Circulating her2/neu: Clinical utility. In Giampietro Gasparini and Daniel F. Hayes, editors, *Biomarkers in Breast Cancer: Molecular Diagnostics for Predicting and Monitoring Therapeutic Effect*. Humana Press Inc., 2006.

[13] National Cancer Institute. Seer cancer statistics review 1975-2007. `http://seer.cancer.gov/csr/1975_2007/browse_csr.php?section=4&page=sect_04_table.18.html`. [Online; accessed 24-June-2010].

[14] Kathleen E. Malone and Kerryn W. Reding. Inherited predisposition: Familial aggregation and high risk genes. In Christopher I. Li, editor, *Breast Cancer Epidemiology*. Springer, 2010.

[15] Mary-Claire King, Joan H. Marks, and Jessica B. Mandell. Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. *Science*, 302(5645):643–646, 2003.

[16] I.O. Ellis, S.J. Schnitt, X. Sastre-Garau, G. Bussolati, F.A. Tavassoli, V. Eusebi, J.L. Peterse, K. Mukai, L. Tabár, J. Jacquemier, C.J. Cornelisse, A.J. Sasco, R. Kaaks, P. Pisani, D.E. Goldgar, P. Devilee, M.J. Cleton-Jansen, A.L. Børresen-Dale, L. van't Veer, and A. Sapino. Tumours of the breast. In Fattaneh A. Tavassoli and Peter Devilee, editors, *World Health Organization Classification of Tumours Pathology & Genetics: Tumours of the Breast and Female Genital Organs*. IARC Press, 2003.

[17] Tara L. Huston and Michael P. Osbourne. Evaluating and staging the patient with breast cancer. In Daniel F. Roses, editor, *Breast Cancer*. Elsevier Inc., second edition, 2005.

[18] National Cancer Institute. *SEER Cancer Statistics Review, 1975-2007.*

[19] Canadian Breast Cancer Foundation. Screening mammography: When should i get a mammogram? `http://www.cbcf.org/breastcancer/bc_early_sc_we.asp`. [Online; accessed 12-November-2010].

[20] Myron Moskowitz. Breast cancer: Age-specific growth rates and screening strategies. *Radiology*, 161(1):37–41, 1986.

[21] David J. Dowsett, Patrick A. Kenny, and R. Eugene Johnston. *The Physics of Diagnostic Imaging*. Hodder Arnold, second edition, 2006.

[22] William R. Hendee and Russell E. Ritenour. *Medical Imaging Physics*. Wiley-Liss Inc., fourth edition, 2002.

[23] Chris Flynn. Bremsstrahlung. `http://commons.wikimedia.org/wiki/File:Bremsstrahlung.svg`, 2001. [Online; accessed 6-May-2010].

[24] Henrik Midtiby. Characteristic radiation. `http://commons.wikimedia.org/wiki/File:CharacteristicRadiation.svg`, 2008. [Online; accessed 6-May-2010].

[25] Winfried Schlke. *Electron dynamics by inelastic X-ray scattering*. Oxford University Press, 2007.

[26] A. A. Bharath. *Introductory Medical Imaging*. Morgan & Claypool, 2009.

[27] R. Edward Hendrick, Etta D. Pisano, Alice Averbukh, Catherine Moran, Eric A. Berns, Martin J. Yaffe, Benjamin Herman, Suddhasatta Acharyya, and Constantine Gatsonis. Comparison of acquisition parameters and breast dose in

digital mammography and screen-film mammography in the American College of Radiology Imaging Network Digital Mammographic Imaging Screening Trial. *American Journal of Roentgenology*, 194(2):362–369, 2010.

[28] Etta D. Pisano, Constantine Gatsonis, Edward Hendrick, Martin Yaffe, Janet K. Baum, Suddhasatta Acharyya, Emily F. Conant, Laurie L. Fajardo, Lawerence Bassett, Carl D'Orsi, Roberta Jong, and Murray Rebner. Diagnostic performance of digital versus film mammography for breast-cancer screening. *The New England Journal of Medicine*, 353(17):1773–1783, 2005.

[29] Rebecca Smith-Bindman, Philip Chu, Diana L. Miglioretti, Chris Quale, Robert D. Rosenberg, Gary Cutter, Berta Geller, Peter Bacchetti, Edward A. Sickles, and Karla Kerlikowske. Physician predictors of mammographic accuracy. *Journal of the National Cancer Institute*, 97(5):358–367, 2005.

[30] Andrew J. Coldman, Diane Major, Gregory P. Doyle, Yulia D'yachkova, Norm Phillips, Jay Onysko, Rene Shumak, Norah E. Smith, and Nancy Wadden. Organized breast screening programs in Canada: effect of radiologist reading volumes on outcomes. *Radiology*, 238(3):809–815, 2006.

[31] I. Anttinen, M. Pamilo, M. Soiva, and M. Roiha. Double reading of mammography screening films - one radiologist or two? *Clinical Radiology*, 48(6):414–421, 1993.

[32] E. L. Thurfjell, K. A. Lernevall, and A. A. Taube. Benefit of independent double reading in a population-based mammography screening program. *Radiology*, 191(1):241–244, 1994.

[33] R.M.L. Warren and S.W. Duffy. Comparison of single reading with double reading of mammograms, and change in effectiveness with experience. *The British Journal of Radiology*, 68(813):958–962, 1995.

[34] Robert J. McKenna. The abnormal mammogram radiographic findings, diagnostic options, pathology, and stage of cancer diagnosis. *Cancer*, 74(S1):244–255, 1994.

[35] J Suckling, C R M Boggis, and I Hutt. The mammographic image analysis society digital mammogram database. In *Exerpta Medica. International Congress*, 1069, pages 375–378, 1994.

[36] Michael Heath, Kevin Bowyer, Daniel Kopans, Richard Moore, and W. Philip Kegelmeyer. The digital database for screening mammography. In M.J. Yaffe, editor, *Proceedings of the Fifth International Workshop on Digital Mammography*, pages 212–218. Medical Physics Publishing, 2001.

[37] Michael Heath, Kevin Bowyer, Daniel Kopans, W. Philip Kegelmeyer, Richard Moore, Kyong Chang, and S. Munish Kumaran. Current status of the digital database for screening mammography. In *Proceedings of the Fourth International Workshop on Digital Mammography*, pages 457–460. Kluwer Academic Publishers, 1998.

[38] S. Obenauer, K. P. Hermann, and E. Grabbe. Applications and literature review of the BI-RADS classification. *European Radiology*, 15(5):1027–1036, 2005.

[39] Antonius A. J. Roelofs, Nico Karssemeijer, Nora Wedekind, Christian Beck, Sander van Woudenberg, Peter R. Snoeren, Jan H. C. L. Hendriks, Marco Rosselli del Turco, Nils Bjurstam, Hans Junkermann, David Beijerinck, Brigitte Sradour, and Carl J. G. Evertsz. Importance of comparison of current and prior mammograms in breast cancer screening. *Radiology*, 242(1):70–77, 2007.

[40] David Gur, Jennifer S. Stalder, Lara A. Hardesty, Bin Zheng, Jules H. Sumkin, Denise M. Chough, Betty E. Shindel, and Howard E. Rockette. Computer-aided detection performance in mammographic examination of masses: Assessment. *Radiology*, 233(2):418–423, 2004.

[41] Joshua J. Fenton, Stephen H. Taplin, Patricia A. Carney, Linn Abraham, Edward A. Sickles, Carl D'Orsi, Eric A. Berns, Gary Cutter, R. Edward Hendrick, William E. Barlow, and Joann G. Elmore. Influence of computer-aided detection on performance of screening mammography. *The New England Journal of Medicine*, 356(14):1399–1409, 2007.

[42] Michael Barnett. Semi-automated search for abnormalities in mammographic x-ray images. Master's thesis, University of Saskatchewan, 2006.

[43] Wilhelm Burger and Mark J. Burge. *Principles of digital image processing: core algorithms*. Springer, 2009.

[44] Syed Ali Khayam. The discrete cosine transform (dct): Theory and application. Technical report, Michigan State University, 2003.

[45] Mark Owen. *Practical Signal Processing*. Cambridge University Press, 2007.

[46] Changhoon Yim. An efficient method for DCT-domain separable symmetric 2-D linear filtering. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(4):517–521, 2004.

[47] Warren J. Ewens and Gregory R. Grant. *Statistical methods in bioinformatics: an introduction*. Springer, second edition, 2005.

[48] Rodolfo Hermans. Skewness statistics. `http://commons.wikimedia.org/wiki/File:Skewness_Statistics.svg`, 2008. [Online; accessed 21-June-2010].

[49] Mark Sweep. Standard symmetric. `http://commons.wikimedia.org/wiki/File:Standard_symmetric_pdfs.png`, 2006. [Online; accessed 21-June-2010].

[50] Mehmed Kantardzic. *Data Mining: concepts, models, methods, and algorithms*. IEEE Press, 2003.

[51] Pierre A. Devijver and Josef Kittler. *Pattern Recognition: A Statistical Approach*. Prentice Hall, 1982.

[52] Vladimir Cherkassky and Filip Mulier. *Learning from Data: concepts, theory, and methods*. John Wiley & Sons, Inc, 1998.

[53] Sergios Theodoridis and Konstantinos Koutroumbas. *Pattern Recognition*. Elsevier Inc, fourth edition, 2009.

[54] E. T. Jaynes. *Probability theory: the logic of science.* Cambridge University Press, 2003.

[55] David J. Hand and Keming Yu. Idiot's Bayes-not so stupid after all? *International Statistical Review*, 69(3):385–398, 2001.

[56] Fan Jiuluna and Xie Winxinb. Minimum error thresholding: A note. *Pattern Recognition Letters*, 18(8):705–709, 1997.

[57] Herbert A. Sturges. The choice of a class interval. *Journal of the American Statistical Association*, 21(153):65–66, 1926.

[58] Seung Ja Kim, Woo Kyung Moon, Nariya Cho, Joo Hee Cha, Sun Mi Kim, and Jung-Gi Im. Computer-aided detection in digital mammography: Comparison of craniocaudal, mediolateral oblique, and mediolateral views. *Radiology*, 241(3):695–701, 2006.