SEMI-AUTOMATED SEARCH FOR ABNORMALITIES IN MAMMOGRAPHIC X-RAY IMAGES

A thesis submitted to the College of
Graduate Studies and Research in partial
fulfillment of the requirements for the
degree of Master of Science

Department of Physics and Engineering Physics
University of Saskatchewan

By

Michael Barnett

PERMISSION TO USE

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Physics and Engineering Physics

University of Saskatchewan

Saskatoon, Saskatchewan S7N 5E2

ABSTRACT

Breast cancer is the most commonly diagnosed cancer among Canadian women; x-ray mammography is the leading screening technique for early detection. This work introduces a semi-automated technique for analyzing mammographic x-ray images to measure their degree of suspiciousness for containing abnormalities. The designed system applies the discrete wavelet transform to parse the images and extracts statistical features that characterize an image's content, such as the mean intensity and the skewness of the intensity. A naïve Bayesian classifier uses these features to classify the images, achieving sensitivities as high as 99.5% for a data set containing 1714 images. To generate confidence levels, multiple classifiers are combined in three possible ways: a sequential series of classifiers, a vote-taking scheme of classifiers, and a network of classifiers tuned to detect particular types of abnormalities. The third method offers sensitivities of 99.85% or higher with specificities above 60%, making it an ideal candidate for pre-screening images. Two confidence level measures are developed: first, a real confidence level measures the true probability that an image was suspicious; and second, a normalized confidence level assumes that normal and suspicious images were equally likely to occur. The second confidence measure allows for more flexibility and could be combined with other factors, such as patient age and family history, to give a better true confidence level than assuming a uniform incidence rate. The system achieves sensitivities exceeding those in other current approaches while maintaining reasonable specificity, especially for the sequential series of classifiers and for the network of tuned classifiers.

ACKNOWLEDGEMENTS

I would like to thank Dr. Edward Kendall, my supervisor, for his continual advice and guidance during my studies, and for directing me towards such a rewarding research topic.

Thank you to the members of my supervisory committee for their comments, suggestions and feedback: Dr. Ron Bolton, Dr. Doug Degenstein, Dr. Rainer Dick, Dr. Mark Eramian, Dr. Rob Pywell, and my external examiner, Dr. Chris Soteros.

I would also like to thank the Mammographic Images Analysis Society and to the Digital Database for Screening Mammography: the existence of such image data sets with confirmed diagnoses made it possible to test the system developed in this work.

Finally, I would like to thank the National Science and Engineering Research Council of Canada for providing me with a Canadian Graduate Scholarship – Master's to carry out this work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CAD – Computer Aided Detection
DDSM – Digital Database for Screening Mammography
FNF – False Negative Fraction
FPF – False Positive Fraction
MIAS – Mammographic Images Analysis Society
TNF – True Negative Fraction
TPF – True Positive Fraction

CHAPTER 1 - INTRODUCTION

1.1 Thesis outline

This thesis describes the development of an automated detection algorithm for determining the likelihood that an x-ray mammogram image features cancer or other abnormalities. The algorithm will use wavelet analysis in conjunction with a novel concerted-effort set of Bayesian classifiers to classify the images as being either normal (free of abnormalities) or suspicious (showing signs of abnormalities, including cancer). This technique gives a measure of confidence in the classification of an image, providing a quantitative measure for determining which images merit further downstream analysis.

Chapter one introduces the motivations for this work and the objectives that this work should achieve. The issue of breast cancer screening is discussed with emphasis on the limitations of current techniques and the alternatives currently being developed. The physical process of x-ray absorption imaging is discussed along with its challenges and the mechanisms that lead to the observed appearances of various tissues and abnormalities in the images. Finally, the algorithm developed in this research is outlined, along with the aspects that make it distinct from other approaches in current literature.

Chapter two discusses the wavelet transform, used to parse each screening image into a form suitable for analysis; the distinctions between previous uses of wavelet analysis in other works and its use in this work are discussed. The use of the two-dimensional discrete wavelet transform as an image decomposition tool is introduced.

Chapter three discusses pattern recognition techniques in current literature, including the naïve Bayesian classifier implemented in this work. A review of current techniques in computer-aided breast cancer detection is given, including a more detailed look at three approaches sharing similarities to the current work. Finally, the concerted-effort set of Bayesian classifiers, a novel extension to current methods, is introduced.

Chapter four discusses the full methodology for classifying an image: each image undergoes pre-processing to reduce artifacts, is decomposed using a two-dimensional discrete wavelet transform, has a set of scalar features extracted from the output of the transform, and is classified as normal or suspicious based on the values of these features. The concerted effort of several classifiers gives a statistical measure of the likelihood that the image is normal or suspicious, and, along with such data as a patient's age and family history, can contribute to the decision of whether to further examine a particular image or patient.

Chapter five shows the results of the developed system, tested on the MIAS database of digitized mammography images [51]. The full system is also retested on the larger DDSM database of images [23] to assess the system's flexibility and its performance on a large data set.

Chapter six discusses the significance of the results, and chapter seven contains the conclusions that may be drawn from this work.

1.2 Motivation: breast cancer and current screening methods

Breast cancer is the most commonly diagnosed form of cancer in women and the second-leading cause of cancer-related death behind lung cancer [39]. In Canada in 2004, there were 21200 newly diagnosed cases of breast cancer and 5200 deaths from breast cancer. In the same year, among women, there were 9800 new cases of lung cancer and 8200 deaths from lung cancer, the second most commonly diagnosed cancer among Canadian women.

Breast cancer is much more prevalent among older women: among newly diagnosed patients, 21% are younger than 50, 49% are between 50 and 69, and 30% are 70 or older [39]. Because of the increased risk of developing breast cancer with age, Health Canada has recommended that women over 50 receive a screening mammogram every two years; other western countries have similar screening policies in place. X-ray mammography is the primary method for early detection of breast cancer and is capable of detecting signs of cancer too subtle or small to be detected by either self-examination or routine physical examination by a physician.

1.2.1 Types of breast cancer

The term breast cancer refers to a variety of cancers of the breast, some more common or dangerous than others. The individual cancer types are named for the tissue in which they occur and for whether they extend into neighbouring tissues [37]. Invasive cancers have progressed into neighbouring tissues beyond the tumour's site of origin and are more dangerous than in situ cancers, which remain contained within their original tissue. Invasive cancers are more likely to metastasize, or have cells break off

and initiate cancers in other parts of the body. The most common locations for breast cancer are in the ducts and lobules of the mammary glands; the four cancers occurring here are invasive ductile carcinoma, ductile carcinoma in situ, invasive lobular carcinoma, and lobular carcinoma in situ. Less common cancers of the breast include: tumours in the connective tissue of the breast called sarcomas; cancer affecting the nipple and aureole called Paget's disease of the breast; and large, bulky tumours called Phylloides tumours. A common benign abnormality is an adenoma, an abnormal growth of cells in the interior wall of a duct or internal passage of the body. Adenomas have a low risk of developing into malignant cancer.

1.3 Physics of x-ray mammography

X-ray mammography images show the inverse of the attenuation or absorption rates of x-ray photons passing through breast tissue. Since different atomic species have different rates of interaction with x-ray photons of different energies, imaging can be done by choosing photon energies where the difference in absorption rates between two materials of interest is maximized. In the case of mammography, low energy x-rays with energies around 20 to 30 keV are used to distinguish between different types and densities of soft tissue.

An x-ray image is formed by passing a collimated beam of photons through a target onto an imaging device, either a film plate or a digital detector [15]. The beam is generated by a Bremsstrahlung process: a beam of electrons emitted from a heated cathode are accelerated towards an anode, typically made of molybdenum or rhodium. When the electrons strike the anode and rapidly decelerate, photons are produced with a

continuous spectrum of energies corresponding to a Bremsstrahlung spectrum. The emitted photons are restricted by shielding to exit the beam source only along a single, narrow path, forming a beam that passes into the target.

Regions of the target which absorb photons strongly will result in reduced photon counts at the corresponding location on the imaging plane; in the case of film imaging, this will mean that the location on the film plate will be less exposed and will thus appear brighter once the film is developed. The photons interact almost exclusively with electrons in the target at x-ray energies; hence, the resulting image is a function of the electron density within the sample, projected into a two-dimensional plane. This makes x-ray imaging sensitive both to changes in density between different tissues and to the presence of different elements with different numbers of electrons, such as the calcium in bone.

1.3.1 Absorption processes at clinical beam energies

X-ray mammography images are typically taken with photon beam energies between 20 and 30 keV. At these energies, photons interact with a sample by three possible mechanisms: coherent scattering, also called diffraction; incoherent scattering, also called Compton scattering; and the photoelectric effect. Coherent scatter occurs from structures with a regular long-range structure, such as crystal; however, because atoms in breast tissue do not show long-range order, diffraction is only a minor process in the generation of clinical mammography images. Compton scattering and the photoelectric effect do contribute significantly to the beam attenuation; each is discussed in turn below.

In Compton scattering, a photon interacts with a target electron, scattering the electron at an angle of up to 90° with respect to the incident photon's trajectory, and a lower energy photon is emitted at a different angle. Because a photon may be emitted from the interaction site at any angle, strikes may occur on the imaging surface at any point, artificially raising the measured photon flux at that point. The distribution of scattered photons is forward-peaked, so many of the photons are deflected only slightly from their original path, creating a blurring effect in the imaging plane. These scattered photons make Compton scattering a major contributor to noise in an x-ray image.

At low energies where the photon energy is significantly smaller than the electron rest mass energy of 511 keV, the cross section for Compton scattering is well-approximated by the Thomson scattering formula. The soft x-ray regime where mammographic x-ray images are taken falls into this low energy region, so the total cross section for Compton scattering from a single electron, $\sigma_{Compton}$, is approximately independent of beam energy and atomic number, and has the simple form [26]:

$$\sigma_{Compton} = \frac{8}{3}\pi\left(\frac{1}{4\pi\varepsilon_0}\frac{e^2}{mc^2}\right)^2 = \frac{8}{3}\pi r_e^{\,2}, \qquad (1.1)$$

where e is the electron charge, m is the electron mass, c is the speed of light and r_e = 2.82 x 10^{-15} m is the classical electron radius. At higher energies, the Klein-Nishina formula describes the dependence of the cross section on beam energy: the cross section scales with the inverse of the beam energy for photon energies comparable to or larger than the electron rest mass energy. Since Compton scattering depends most strongly on the density of scattering centers (electrons) while the photoelectric effect depends on the

cube of the atomic number, Compton scattering is not as sensitive as the photoelectric effect for differentiating between different materials in a target.

The photoelectric effect is the absorption of a photon by a bound electron; the photon energy not used in liberating the electron from its bound state is converted into kinetic energy for the electron. The hole left by the liberated electron in its atomic orbital will be filled, either by a free electron or by another electron in the same atom transitioning down from a higher orbital. As the photon energy becomes greater than the binding energy for a lower orbital, it is possible for the photon to interact with and liberate electrons from that orbital, increasing the number of potential interaction centres and thus the total cross section for the photoelectric effect. The full cross section for the photoelectric effect is highly complex and lacks an analytical form for anything but the simplest atoms, though its general trends have a simple form: the cross section scales with the cube of the target's atomic number and with the inverse of the photon's energy.

Since the photoelectric effect is suppressed at high beam energies while Compton scattering is not, Compton scattering becomes the dominant interaction process at beam energies significantly above 30 – 100 keV, with the exact crossover energy increasing as the target's atomic number increases. Figure 1.1 shows the crossover point between the two processes as a function of the incident beam energy and the target atom's atomic number [34]. A low beam energy where the photoelectric effect is dominant is used in x-ray mammography to exploit the photoelectric effect's strong dependence on atomic number, which allows it to differentiate between different materials in a sample. Higher beam energies are used to image thicker targets, since the absorption rates are lower at higher energies: lower absorption rates allow enough

photons to reach the imaging plane and form an image while minimizing the amount of radiation energy deposited into a patient or sample. In either case, image contrast may also be generated by density variations between materials of similar atomic compositions, since denser materials have a larger number of potential interaction centres per unit volume.

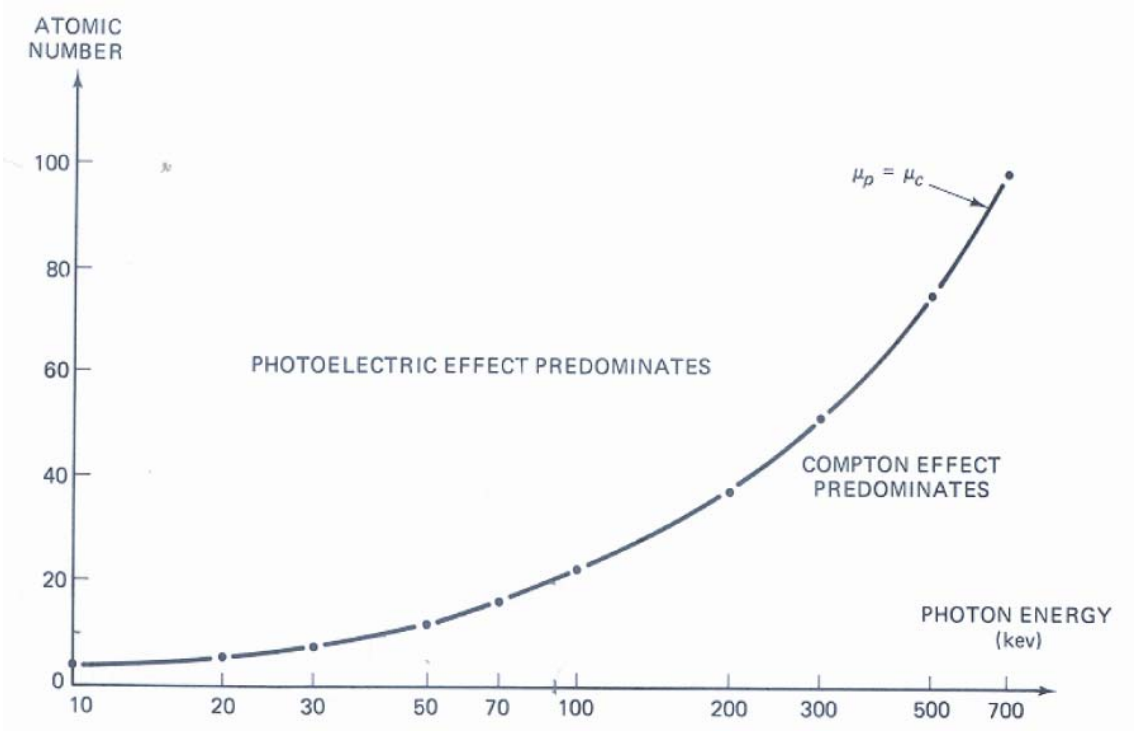
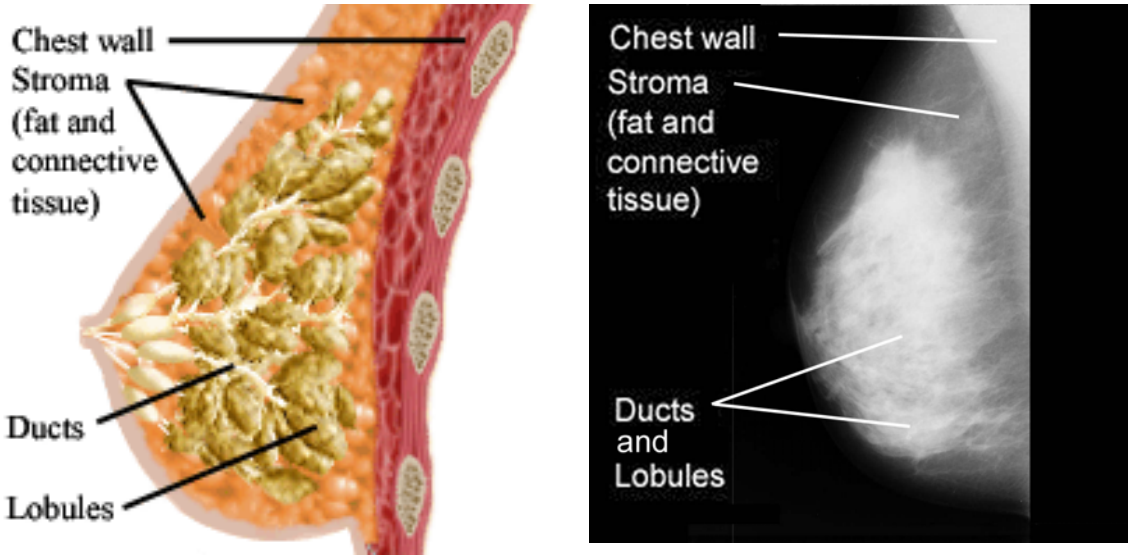


Figure 1.1 - Dominant regimes for the photoelectric effect and Compton scattering as a function of photon energy and atomic number [34]

1.3.2 Appearance of structures in mammographic images

The major tissues of the breast are shown in Figure 1.2 a), and a typical x-ray mammography image for a healthy patient is shown in Figure 1.2 b). These images show the structures typical to all mammographic images: the chest wall, glandular tissue, and stromal tissue. The chest wall, which includes the chest muscles, is relatively thick and dense and appears as a uniformly bright region at the top of Figure 1.2 b).

Glandular tissue consists of the milk-producing lobules and the milk-transporting ducts of the mammary glands; it appears as a relatively bright region which radiates back from the nipple in the x-ray image. Stromal tissue is composed of connective tissue and fatty adipose tissue and forms the relatively dark remainder of the tissue region in the x-ray image.



(a) **(b)**

Figure 1.2 – Structural features of healthy breast and corresponding appearance in x-ray mammography image (Image (a) taken from [42])

Two types of abnormalities may appear in mammographic images to indicate the presence of cancer: clusters of small calcium deposits called microcalcifications, and regions of unusual tissue growth called masses.

Microcalcifications are small deposits composed primarily of calcium compounds such as tricalcium phosphate and calcium hydroxyapatite [43]. Calcium has a relatively large atomic number (Z = 20) compared with more abundant biological elements such as carbon (Z = 6), hydrogen (Z = 1), oxygen (Z = 8) and nitrogen (Z = 7).

At mammographic imaging energies, the photoelectric effect is the dominant process and scales with the cube of the atomic number; hence, microcalcifications, like bones in other x-ray images, have absorption coefficients more than ten times higher than the more common materials in the breast tissue and create the brightest spots on the final x-ray image. The small, sharp appearance of microcalcifications in an image, combined with their relative brightness, makes them easier than masses to discern, although calcifications can still be obscured by dense, bright regions of glandular tissue.

Masses are regions of unusual tissue and may be benign (not at risk of metastasizing to other tissues) or malignant (actively worsening and at an increased risk of metastasis). Masses appear on a mammography image if their density is significantly higher than the density of the surrounding tissue. An increase in material density directly corresponds to an increase in electron density and thus an increase in photon absorption. This difference in density causes masses to appear slightly brighter than surrounding tissue on an x-ray image. Masses which are similar in density to the surrounding tissue may not be directly visible, but their existence may be inferred from the distortions they create in the structure of the surrounding tissue; this is known as architectural distortion, and is far more difficult to detect than directly visible masses.

1.3.3 Digital x-ray image resolution limitations

Mammographic x-ray images are most readily quantitatively analysed when they are in digital form, either by converting film images or by taking the original images with a digital detector. Several processes limit the resolution of the resulting pixel

images and occur during the image formation process, the image recording process and the digitization process.

Resolution during the image formation process is limited by the beam shape and by scattering processes. If the x-ray beam is not emitted from a single point, but rather from a source of finite width, then the rays of the beam photons will not radiate from a single point and the resulting image will be blurred. For a source of width W_{source} a distance L_{target} from a point in the target and a distance L_{image} from the imaging plane, the point in the target spreads out to have a finite width W_{point} in the image:

$$W_{point} = W_{source} \frac{L_{image} - L_{target}}{L_{target}}. \tag{1.2}$$

This blurring can be minimized by placing the imaging plane as close to the target as possible, though this limits the magnification of the image. For x-ray mammography, magnification is considered less important than resolution, and so the imaging plane is placed just beneath the target.

A second process limiting resolution during image formation is scattering within the tissue. Photons which pass through the tissue without being absorbed may still be deflected from their initial trajectory by coherent and incoherent scattering. These processes spread out the region where a photon is likely to strike the imaging plane, lowering the image resolution. This effect is intrinsic to the process being studied and cannot be removed; further, scatter is the largest source of resolution loss in x-ray imaging, acting to reduce the sharpness of edges and smear out the appearance of fine structures [15].

The image recording process is performed either on a conventional film plate or on a digital detector. The resolution of a film plate is limited by the size of the film

grains that the photons interact with; typical film grains are approximately 0.1 to 1 microns in diameter [15]. The resolution of a digital detector is set by the element size of the detector. Detector elements are typically similar in size to the elements in a scanner that converts film images into digital form, or approximately 40 to 50 microns.

The digitization process for film images requires selecting a resolution scale corresponding to the size of each pixel. For mammography images, pixels are typically 40 to 50 microns in width; a down-sampled image suitable for computational analysis may have pixels 150 to 200 microns in width, depending on the type of digital scanner used and the desired compromise between image resolution and computational efficiency. The pixel size sets the size scale for the smallest structures that may be resolved on the resulting images; since microcalcifications are the smallest structures of interest in mammographic x-ray images and are typically several hundred microns across or larger, a pixel size of 200 microns or less is sufficiently small to resolve them. Detectable masses range in size from a few millimetres to a few centimetres in diameter and so are also easily detectable with this level of resolution.

1.4 Challenges in x-ray mammography

X-ray mammography consists of an x-ray of the breast tissue under transverse compression. Signs indicative of possible cancer include the presence of clusters of microcalcifications or the presence of a mass, which can be either directly observed or inferred from the distortions it causes in the structure of the surrounding soft tissues. Microcalcifications are clusters of calcium compound deposits left by several biological processes; most notably for cancer screening, they are sometimes, though not always,

secreted by rapidly dividing cells such as cancer cells [43]. Cells in the glandular tissue of the breast are specialized to store milk containing calcium; when these cells become cancerous and rapidly divide, such as in lobular or ductile carcinoma, deposits of calcium may be released during cell division. Microcalcifications alone show a high correlation with breast cancer, but their appearance alone does not guarantee the presence of cancer, nor does their absence negate the possibility of cancer. Because of this uncertainty, any current automated methods which use only the presence of microcalcifications to mark suspicious images are necessarily limited in their capacity to correctly identify all images showing abnormalities.

Interpretation of mammograms is a difficult process for several reasons. Because the images are formed from x-rays passing through the tissue, contrast in the image depends on differences in the absorption rates and densities of different materials within the tissue; however, these differences are much more subtle between the types of soft tissues in the breast than they are between bone and soft tissue in a chest x-ray, for example. The indications of cancer in an image are often also very subtle, such as small distortions in the structure of the ductile tissue of the breast or the presence of small clusters of microcalcifications.

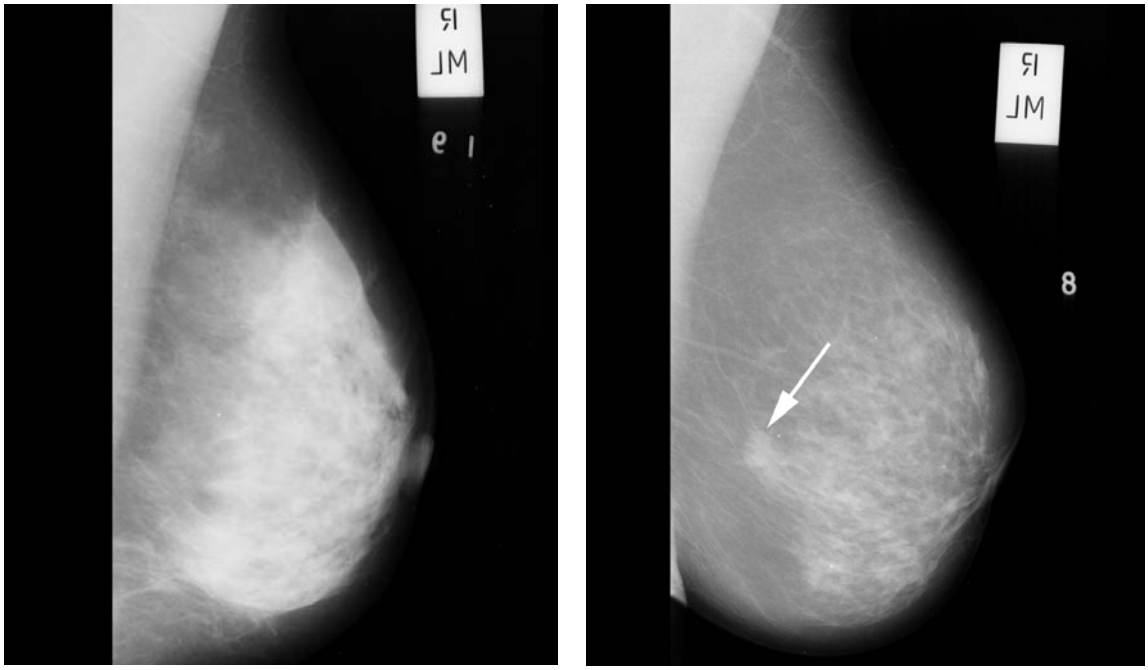


Figure 1.3 – Typical mammograms showing healthy tissue (left) and showing a cancerous mass (right, marked with white arrow)

Figure 1.3 shows two typical mammography images: the left image shows a healthy patient, while the right image shows a cancerous mass.[1] These images are medial-lateral images, as signified by the ML marker on the x-ray film, and are taken with the x-rays passing horizontally through the tissue while the breast is compressed between two vertical plates. The other common image view is cranial-caudal, typically denoted as CC, which is taken with x-rays passing vertically through the tissue while the breast is compressed between two horizontal plates. Breast tissue can vary greatly in appearance between patients depending on the relative amounts of glandular and fatty tissue present, which appear as relatively bright and dim regions, respectively, further adding to the challenge of detecting abnormalities within a particular patient's image.

[1] The text within the images is reversed because the images were reversed during the scanning process that converted the original film plate images into digital form. This reversal does not affect the information content of the images, and so no correction was made. Images are taken from the MIAS database [31].

An additional challenge in the interpretation of mammograms is the small percentage of images that show abnormalities. One typical clinic diagnosed 6.4 cancers per 1000 patients; at two medial-lateral images per patient, this means only one in 300 images showed signs of cancer [4]. This low rate of incidence means that a large fraction of images marked as suspicious are actually normal. Positive findings then require a patient to undergo further procedures to verify or refute the diagnosis. For example, 5-7% of women in the above study were recalled for further tests, though only one in ten of those recalled were actually positive for cancer.

Typical CAD systems deal with the subtlety of cancer signs on images by marking all suspicious regions in images for radiologists to re-examine. This approach may increase the number of cancer cases which are correctly diagnosed, but it also increases the number of images that a radiologist must study in greater detail and may increase the number of healthy patients recalled as being suspect for cancer.

Another approach for CAD design is to pre-screen the images and remove those which are least likely to show abnormal pathology. This will cause a greater fraction of the remaining images to show pathology, potentially reducing the number of false positives generated by any subsequent analysis of the remaining images.

1.5 Terminology of diagnosis rates and consequences of misdiagnosis

The performance of a screening or classification system, such as the use of mammography for detecting breast cancers, can be measured by two parameters: sensitivity and specificity.

Sensitivity is also known as the true positive fraction; it is the fraction of pathologies being screened for which are actually detected. For example, if a large set of patients were screened and 100 of them had breast cancer, a screening procedure which detected 90 of these women's cancers would have a sensitivity of 90%. Because of the serious health consequences of missing a diagnosis of cancer until a later screening, sensitivity is deemed more important than specificity in cancer screening protocols.

Specificity is also known as the true negative fraction; it is the fraction of cases which do not correspond to cancer which are correctly classified as normal. For example, if a set of 1000 patients with 100 cases of cancer were screened, a screening procedure which found 810 of the 900 normal patients to be normal and the other 90 to possibly have cancer would have a specificity of 810/900 or 90%. Although the consequences of a false positive, that is, diagnosing a normal patient as having breast cancer, are less severe than missing a positive diagnosis for cancer, specificity should also be as high as possible. False positives can lead to painful, invasive procedures, such as tissue biopsy, to confirm or refute the diagnosis and can lead to significant anxiety and concern for the patient.

1.6 Current innovations in screening procedures

A new challenge facing physicians as the population ages is to handle the rising volume of mammographic images produced by current screening policies while maintaining a high rate of correct diagnoses. To this end, two major innovations to x-ray mammography have been introduced: computer aided detection (CAD) and double

readings. Other imaging modalities are also being explored for breast cancer screening in addition to x-ray mammography, including magnetic resonance imaging [31,53], diffraction-enhanced imaging [11,28] and phase contrast imaging [1,41]. These modalities attempt to provide better contrast between normal and abnormal tissues, making abnormalities easier to discern in images.

The method of double reading is primarily designed to reduce the number of missed cancer cases by having two radiologists independently interpret each image, conferring when there is a discrepancy between their diagnoses. This method has been shown to boost sensitivity from 74% to 89%, according to one study [27]. The drawback of this method is that throughput is halved for a given number of physicians at a clinic, since each image is read in full detail twice; this limits the number of patients that may be screened and how frequently each patient can be screened over her lifetime.

The second method, CAD systems, uses automated software to mark suspicious regions on each mammogram image. Once a radiologist has studied an original image, a marked up CAD image is consulted to see if there are any additional regions of concern that could change the interpretation of the image. This method is similar to double reading but maintains nearly the same level of throughput as a single reading procedure. The primary drawback of this approach is that CAD systems are conservative and mark far more images as suspicious than actual cancer rates warrant, drawing unnecessary attention to healthy tissue regions and increasing the risk of false positive diagnoses.

An alternative method for automated systems is to identify and remove normal images, leaving the more suspect images for further analysis. This approach relies on a whole-image analysis rather than marking suspicious regions: an image showing no

signs of pathology is marked as normal and is assigned a lower priority for further analysis. For this method to be acceptable in a clinical setting, it must offer an extremely high rate of sensitivity to detect as many of the abnormal images as possible. The method's utility depends on a relatively high specificity, since its goal is to minimize the number of images requiring further analysis.

Either type of automated system faces the challenge of parsing large amounts of data, in the form of a high resolution image, into a single conclusion or set of marked suspicious regions. Typically, a small number of parameters are measured from an image, and the image is characterized based on their values, rather than on the values of every pixel. The process of extracting a small number of parameters, or features, from a large image is known as feature extraction; feature extraction typically involves applying a mathematical transform to reduce the data volume, such as filtering or statistical measurements. Once a small set of features has been measured for an image or region, a classification is made using a pattern recognition tool called a classifier; classifiers use information from images with known classifications to characterize new images. This work will apply the wavelet transform to parse the data into multiple scales and extract scalar features from each view to characterize the image. The wavelet transform has similarities to the Fourier transform and is discussed in detail in Section 4.

1.7 Objectives and novelty of this work

The primary goal of this thesis is to develop a method for detecting a range of breast tissue abnormalities from mammographic x-ray images. The algorithm must classify a given image as being either normal or suspicious and give a confidence level

for this classification; this confidence level will allow a physician to judge the strength of the given image's classification and determine which images merit further study.

To develop the full algorithm, several intermediate objectives must be met:

Objectives:

1. Develop a set of pre-processing steps to isolate the tissue in the images and regularize the appearance of the images to make direct comparisons possible.

2. Apply the discrete wavelet transform to parse an image and generate a set of scalar features based on the output of the transform to characterize each image.

3. Develop an automated tool that can use the generated features to classify the images as normal or suspicious and give a confidence level for this classification.

Although other computer aided detection algorithms currently exist for the detection of breast cancer in mammographic images, this work offers several original innovations. First, the wavelet transform is being used to generate scalar features directly, whereas other works have only applied wavelets as a noise reduction tool or to emphasize certain structures in an image. This technique was applied to small angle x-ray scattering images of breast biopsy tissue by Carissa Erickson [19] and extended to mammographic x-ray images by Krista Chytyk [13], though this is the first rigorous application of this technique to x-ray mammographic images for detecting suspicious images. The second innovation is in the implementation of the pattern recognition tool being applied, a naïve Bayesian classifier, which does not normally provide a confidence level for its classification of a given image; the concerted-effort set of

classifiers being developed in this work offers a mechanism for extracting confidence levels from several Bayesian classifiers working in tandem. Although other ensemble methods for combining multiple classifiers have been proposed in literature [40], those methods typically combine classifiers to improve accuracy alone, not to produce confidence levels from classifiers lacking this information. Thirdly, this work carries out whole-image classification, identifying whether each image merits further study or is likely to be free of abnormalities. Many other approaches identify regions of concern within individual images rather than identifying and removing healthy images and hence potentially boosting throughput. Improved screening efficiency may, among other benefits, allow women to be screened more frequently, increasing the probability of detecting cancers earlier in their development when they are more treatable. While screening frequency is ultimately limited by radiation dose, the current screening rate of once every two years is well within safe standards and is less frequent than the annual screening rate recommended in Canada for patients with a family history of breast cancer [20].

1.8 Approach to be taken in this work

The object of this work is to develop a method for analyzing mammographic images, identifying healthy and suspicious images, and determining a confidence level for this classification. Each image is to be classified as a whole, rather than marking suspicious regions within each image, as a method for removing the images least likely to correspond to cancer from the set of images requiring further analysis. The approach of pre-screening images is uncommon in current literature: it offers a method for

directly increasing the number of patients that may be screened, but it requires extremely high sensitivity, since any missed cancers at the pre-screening stage will not be found later. This removal of images sets an automatic upper limit on the sensitivity of the entire procedure, regardless of the efficacy of later steps.

The images will first be pre-processed to reduce artifacts and noise that would obscure relevant physical differences between the tissue regions of each image. Once the images are as uniform as possible, a multi-level wavelet decomposition will be used to examine the image at multiple scales. Scalar features, such as mean intensity and standard deviation of intensity, will be measured for each image in the decomposition, forming a large set of possible features for a classifier. The use of whole-image parameters from the wavelet decomposition is a novel approach for feature generation.

A naïve Bayesian classifier will classify the images as either normal or abnormal, with the abnormal class containing benign and cancerous masses and calcifications that merit further examination. A feature reduction step will select a small subset of the total set of generated features which are most effective at performing this classification.

To provide confidence levels for the classification, a novel extension to the naïve Bayesian classifier, a concerted-effort set of Bayes classifiers, will be developed. Multiple small sets of features will each form an independent classifier; this may be done by employing different wavelet bases or scalar features before the feature reduction step. The set of classifiers will each individually classify the image as either normal or suspicious. From this point, several methods will be explored for combining

their results and providing confidence levels: a sequential rejection process, a vote-taking process, and a network of tuned classifiers.

The sequential rejection process passes the feature vectors through one classifier at a time, with all classifiers tuned to have maximal sensitivity. If the feature vector is normal after the first classifier, it is binned as normal, with the probability of error equal to the false negative fraction for the classifier. If the feature vector is abnormal, then it is passed onto the second classifier in the sequence. If the second classifier finds the feature vector to be normal, then it is binned as normal, but with a slightly lower probability than if it had been found normal after the first pass. Thus, after each classifier in the sequence, fewer and fewer images will remain, and a larger and larger fraction of the remaining images will be abnormal. The images removed after each step will have a confidence level for their classification, providing a measure for the need to further analyze each image.

The vote-taking process runs each feature vector through all of the classifiers immediately. The confidence level for the resulting classification is computed from the number of classifiers which agreed in their classification and from the sensitivities and specificities of the classifiers. A discrimination threshold may then be selected to remove all images that, for example, have a less than 10% chance of showing abnormalities.

The network of tuned classifiers approach uses individual classifiers that are each tuned to detect different types of abnormalities, such as just masses or just calcifications. Images are first screened for the presence of one type of abnormality: any images found unsuspicious for this abnormality are passed on to other classifiers

that search for other types of abnormalities. By searching for particular types of abnormalities at each step, this approach may offer higher sensitivity than any other process, though its usefulness will depend on the level of specificity it can provide.

For any of the processes to work, the individual classifiers must have low correlations among their misclassified images; that is, they must correctly classify more images together total than each classifier can individually. For example, using the same classifier twice in sequence would have no beneficial effect: since all abnormal images that reached the second classifier would be classified as abnormal again, the second classifier would be redundant. Selecting classifiers with low correlation among their classifications of a set of images will be necessary for any of the proposed methods to be effective.

The full algorithms will be tested on the Mammographic Image Analysis Society (MIAS) database [51]. The database contains over three hundred images, including over 200 normal images and approximately 100 images containing benign and cancerous masses and calcifications. The algorithm is designed to be flexible enough to operate on other image databases and on clinical images without substantial modification, as the feature reduction step will select the most effective features for classification based on the selected imaging view and resolution standard. To test this flexibility as well as the system's performance on a larger scale, the algorithm will be retested on a substantial subset of the Digital Database for Screening Mammography (DDSM) data set consisting of approximately 1000 normal images and 650 suspicious images [23].

CHAPTER 2 – THE WAVELET TRANSFORM

The wavelet transform has similarities to the Fourier transform, but offers several unique advantages. In this work, the wavelet transform operates on x-ray images; hence, it transforms the spatial position information of the original image into the wavenumber domain that encodes the image in terms of spatial frequency components. Whereas the Fourier transform maps a signal completely from the spatial position domain into the wavenumber domain, a wavelet transform maps a signal into a two dimensional position-wavenumber domain. This more complex mapping provides information about a signal's spatial frequency content as a function of position, in contrast to the position-independent output of the Fourier transform. It should be noted, though, that although the Fourier-transformed signal does not explicitly show any spatial dependence, both a Fourier-transformed signal and a wavelet-transformed signal may be mapped back into the spatial position domain by an inverse transformation; this property means that both transforms are conservative and retain all of the information present in the original signal, but differ only in how they represent that information content.

Section 2.1 introduces the continuous wavelet transform as a development out of the Fourier transform. Section 2.2 introduces the discrete wavelet transform in one dimension and some unique properties of wavelet transforms that are relevant to this

work, especially multiresolution. Section 2.3 describes the two-dimensional discrete wavelet transform used in this work to parse the mammographic x-ray images, and Section 2.4 shows a simple example of employing the 2D discrete wavelet transform.

2.1 Comparison of wavelet transform to windowed Fourier transform

The wavelet transform may be understood by examining its development out of the Fourier transform and its extension, the windowed Fourier transform [24,25]. The Fourier series represents an arbitrary periodic function *f(x)* by an infinite series of sine waves of various frequencies, amplitudes and phases. For a function of position *f(x)* with a period $L=x_1-x_0$, the Fourier series *f(x)* can be written as:

$$f(x) = a_0 + \sum_{n=1}^{\infty} \left[a_n \cos(k_n x) + b_n \sin(k_n x) \right], \tag{2.1}$$

where a_n and b_n are the Fourier coefficients for each wavenumber $k_n = 2n\pi/L$. The Fourier coefficients are found from the original function *f(x)* as follows:

$$a_n = \frac{2}{L}\int_{x_0}^{x_1} f(x)\cos(k_n x)dx \qquad b_n = \frac{2}{L}\int_{x_0}^{x_1} f(x)\sin(k_n x)dx. \tag{2.2}$$

Note that these coefficients can be interpreted as the correlation coefficients between the input function *f(x)* and the trigonometric functions $\sin(k_n x)$ and $\cos(k_n x)$ over the period *L* of *f(x)*; this interpretation will be useful in discussing the wavelet transform later.

The Fourier transform for non-periodic functions is similar to the Fourier series, but allows the wavenumber *k* to be continuously valued, replacing the summation with an integral and exploiting Euler's identity $e^{-ikx} = \cos(kx) - i\sin(kx)$:

$$\tilde{f}(k) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(x)e^{-ikx}dx. \tag{2.3}$$

Thus, $\tilde{f}(k)$ shows the amplitude of a signal as a function of wavenumber or spatial frequency, whereas *f(x)* shows the amplitude of a signal as a function of time. For some applications, however, it is desirable to measure how the frequency content of a signal varies over space and what its frequency content is at a particular location. To this end, the windowed Fourier transform was developed, which limits the bounds of integration in (2.3) to a finite window size *L*. The transform is then calculated as a function of the window location, so that the output $\tilde{f}(k,\chi)$ depends on both the wavenumber *k* and the center point of the observation window χ:

$$\tilde{f}(k,\chi) = \frac{1}{\sqrt{2\pi}} \int_{\chi-L/2}^{\chi+L/2} f(x) e^{-ikx} dx. \tag{2.4}$$

The choice of the window size *L* presents a limitation to the windowed Fourier transform. To find the spatial frequency content of a signal at a single point in space, the window size would have to be arbitrarily narrow to exclude contributions from the signal at nearby locations. Low spatial frequency components, however, cannot be accurately distinguished over short spatial intervals, due to the uncertainty relation $\Delta k \Delta x > 1/2$; although the uncertainty in wavenumber Δk is the same for all wavenumbers, the relative uncertainty $\Delta k/k$ is greater at low wavenumbers or low spatial frequencies. To achieve a large relative frequency resolution for low frequencies, a larger window is needed, though this reduces the spatial resolution of the transformed signal. The choice of window size is thus application-dependent and must be selected to give the desired compromise between spatial and frequency resolution in the transformed signal.

The wavelet transform works around the problem of window size by varying it dynamically as a function of the frequency range being probed. Rather than using an infinite set of basis functions, like the sine functions of varying wavenumber used in the Fourier transform, the wavelet transform uses a single basis function, then scales and translates it to probe all possible spatial positions and frequencies. A wavelet basis function has a finite extent, effectively defining a window size over which it is convolved with the spatial domain signal to be transformed. The extent of the basis function can be enlarged or reduced by scaling the length of the function in space in a process called dilation. For example, the simplest wavelet basis is the Haar wavelet: it has a value of 1 for x = 0 to 0.5, -1 for x = 0.5 to 1.0, and zero elsewhere, and is commonly used as an edge detector to mark sharp changes in a signal. To measure changes at a lower frequency, the wavelet can be stretched to be nonzero from x = 0 to 10, which both lowers the frequency to which the basis function is most sensitive by a factor of ten and increases the absolute frequency resolution of the transform by the same amount. Similarly, high frequencies can be probed by shortening the wavelet duration to, for example, x = 0 to x = 0.10, which makes it more sensitive to higher frequencies and provides a higher spatial resolution; the loss of absolute frequency resolution at higher frequencies is less noticeable since the relative wavenumber resolution, $\Delta k/k$, remains constant across all frequencies. This automatic scaling of the basis function's window size with frequency provides a constant relative frequency resolution while maximizing the spatial resolution across all frequencies of the signal.

Figures 2.1-2.3 show the output of convolving the same input signal with a wavelet scaled to three different sizes. The wavelet used is the Daubechies 2 wavelet,

and is scaled to have a duration of 768 units (Figure 2.1), 192 units (Figure 2.2), and 32 units (Figure 2.3) for the three different convolutions. The input signal is a piecewise continuous sine wave that changes frequency at 256, 512 and 768 units. The results of each convolution with the input signal are discussed in turn below.

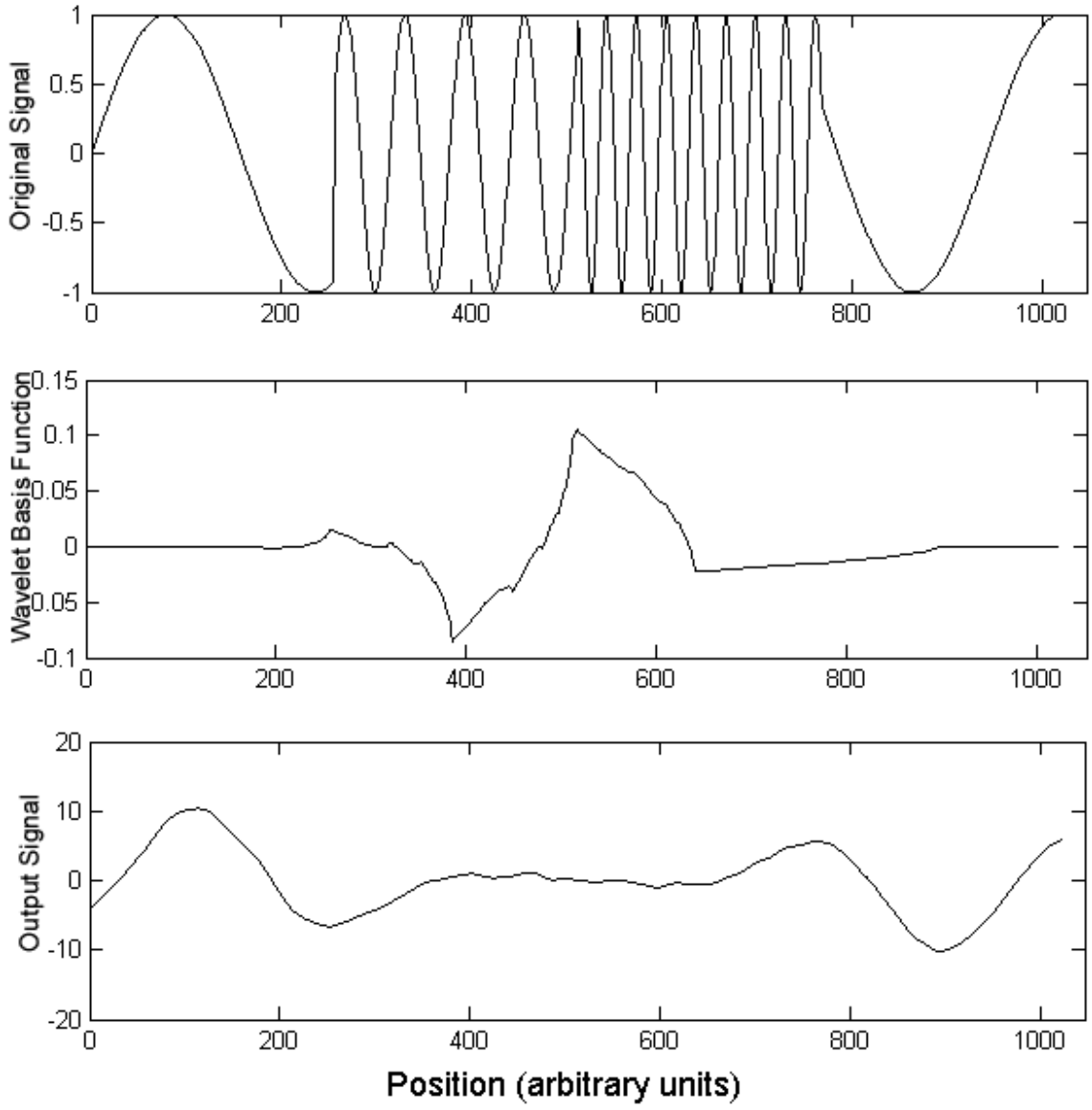


Figure 2.1 – Signal convolved with Daubechies 2 wavelet at a large scale

The first convolution, shown in Figure 2.1, dilates the basis function to have a relatively long duration of 768 units, making it most sensitive to lower frequency

components in the signal. Hence, the first and last quarters of the output signal have the largest coefficients, since these were the corresponding locations in the input signal that were dominated by low frequency components. There is little response to the relatively high frequency central part of the input signal.

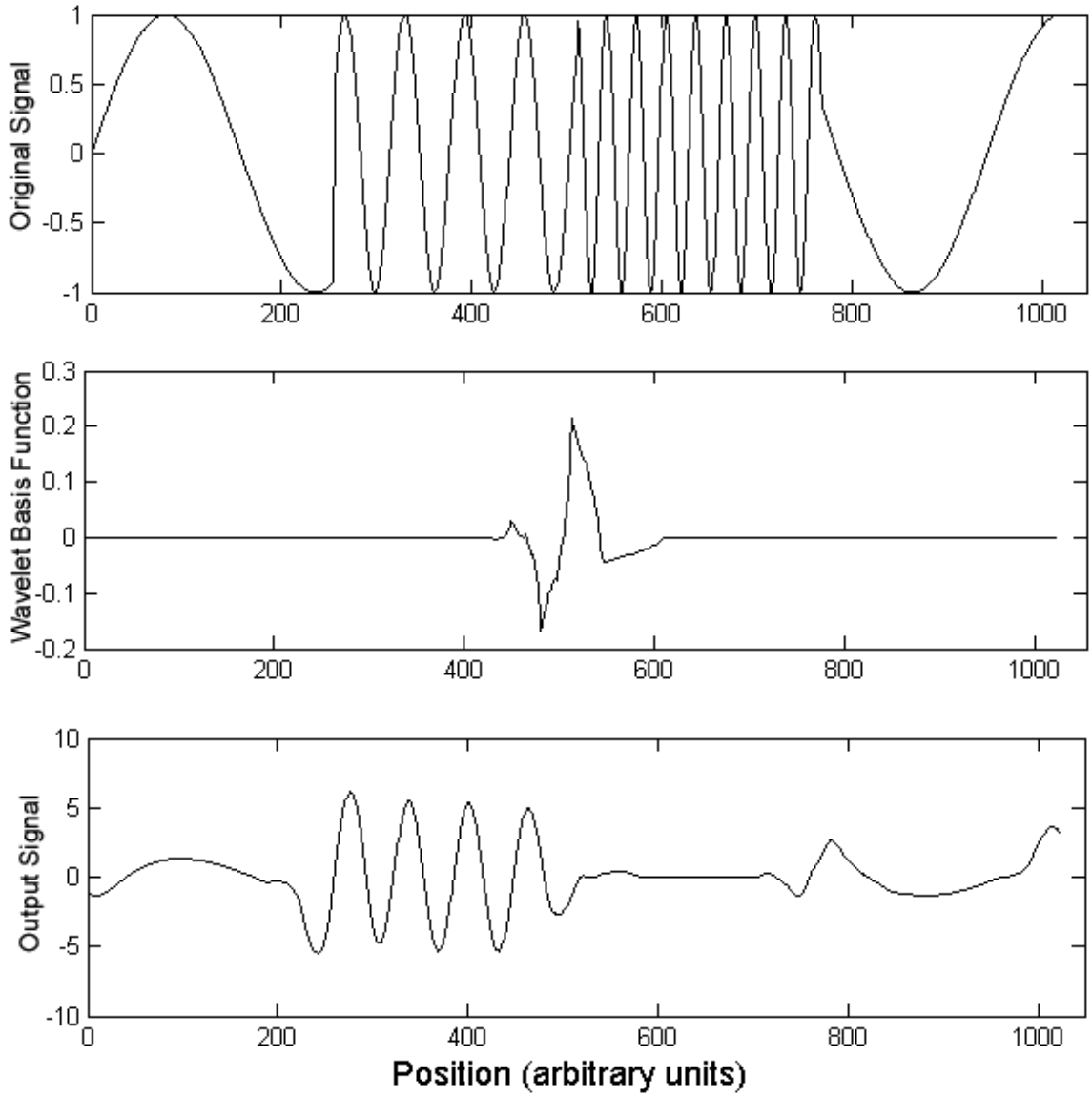


Figure 2.2 – Signal convolved with Daubechies 2 wavelet at a medium scale

The second convolution, shown in Figure 2.2, uses a basis function scaled to a medium duration of 192 units. The output coefficients are greatest for the second

quarter of the signal where the input signal's frequency was moderately high, while the output coefficients are very low in the other regions where the input signal's frequency was significantly higher or lower. Edge effects create the sharp response at around 768 units along the signal where the spatial frequency of the input signal changes abruptly.

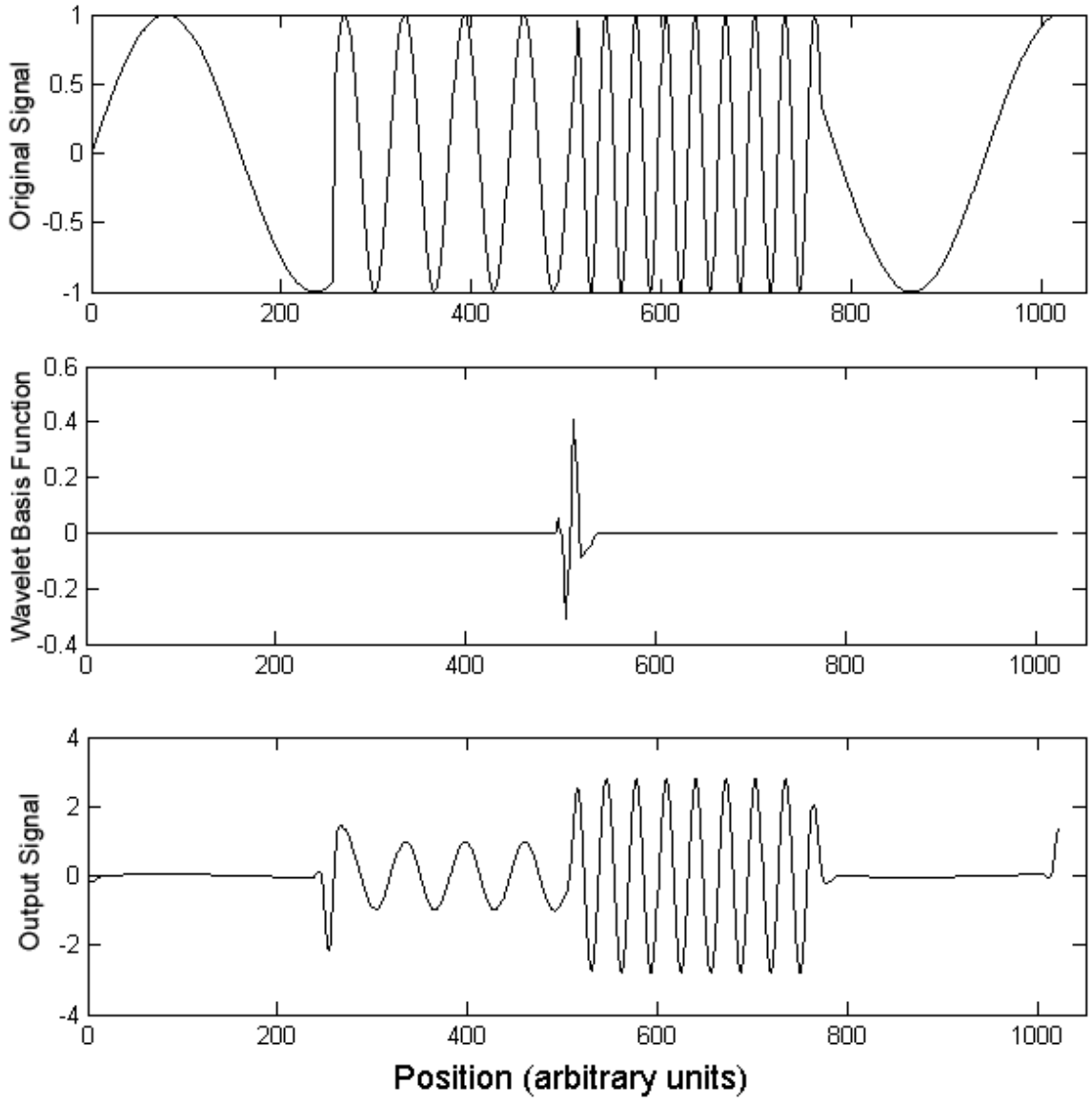


Figure 2.3 – Signal convolved with Daubechies 2 wavelet at a small scale

The third and final convolution, shown in Figure 2.3, uses a basis with a short duration of 32 units to probe the high frequency content of the signal. Although the

output coefficients respond to the second quarter of the signal, they respond most strongly to the high frequency third quarter of the input signal. The output includes almost no contribution from the low frequency first and last quarters of the signal, since the wavelet basis is insensitive to such low frequencies when it has such short duration.

2.2 The discrete wavelet transform and multiresolution analysis

The continuous wavelet transform maps a one-dimensional signal that depends only on position into a two-dimensional space of position and wavelet scale; hence, the transformed signal is highly redundant. The discrete wavelet transform samples only a subset of scales and discretizes the spatial domain, eliminating signal redundancy while maintaining full information content about the original signal. The most common technique for carrying out this discretization is known as dyadic sampling [25].

Dyadic sampling is a recursive procedure for breaking a signal down into its content at varying scales or frequencies. The first pass of the discrete wavelet transform samples the signal at a rate sufficient to resolve the highest frequency component present in the original signal; according to the Nyquist criterion, this sampling rate is twice the maximum frequency present in the original signal. The wavelet basis function is a finite sequence of numbers in the discrete case and is convolved with the original sampled signal to produce an output corresponding to the smallest scale and the highest frequency measured by the transform. Two output signals are maintained: the high frequency or detail signal, which is the output of the convolution between the basis function and the sampled signal; and the low frequency or approximation signal, which is the difference between the original signal and the detail component. Dyadic sampling

keeps only every second point in the two output signals; this eliminates redundancy without information loss, since the original signal can still be recovered from the two low-resolution output signals, so long as the wavelet basis is orthogonal [24]. The approximation signal is then used in the next pass of the wavelet transform to measure the signal content at a lower frequency or larger scale.

To change the wavelet scale, the discrete wavelet transform uses the approximation signal as the input to another pass of the same wavelet basis function. Because the approximation signal contains only half as many points as the input signal, the next pass of the same wavelet function naturally covers twice as large of an extent in space. This process of halving the resolution at each decomposition level is known as down-sampling [25]. The approximation signal at the new level is convolved with the basis function to produce another detail and another approximation signal at a new scale that each contain one-fourth the number of data points that the original signal did. This process can be recursively repeated on the approximation signal until down-sampling has reduced the number of points in the signal to the point that convolution with the wavelet basis function is no longer possible.

Another interpretation for the detail and approximation signals uses the concept of high pass and low pass filters [49]. The detail signal can be regarded as the output from passing the signal through a high pass filter whose response is described by the wavelet basis function. The approximation signal can be regarded as the output from passing the signal through a low pass filter whose response is described by the scaling function. The scaling function is a function complementary to the wavelet basis function that generates the approximation signal when it is convolved with the input

signal. In the discrete case, defining both a scaling function for the low pass filter and a wavelet basis function for the high pass filter greatly simplifies the calculation of the wavelet transform of a signal: at each scale, the input signal is convolved with each of the two functions to produce the outputs, just as is the case for any set of digital filters.

The term multiresolution refers to the structure of the output signals from the discrete wavelet transform. The output consists of detail and approximation signals at each of a set of resolution scales that each differ in resolution by a factor of two. Pictorially, the set of signals produced by the discrete wavelet transform can be represented by a binary tree: each level of the tree consists of a detail signal and an approximation signal; the approximation signal then has two children in the next level of the tree corresponding to the detail and approximation signals at the next, coarser level of resolution. The highest levels of the tree have the finest resolution and correspond to the smallest scales and the highest frequency components of the original signal, while the lowest levels of the tree have the coarsest resolution and correspond to the largest scales and lowest frequency components of the original signal. Figure 2.4 shows a tree structure for a wavelet transform with three levels of decomposition; the topmost node in the tree is the original signal, while each lower level in the tree corresponds to successively coarser resolutions of the signal's detail and approximation components.

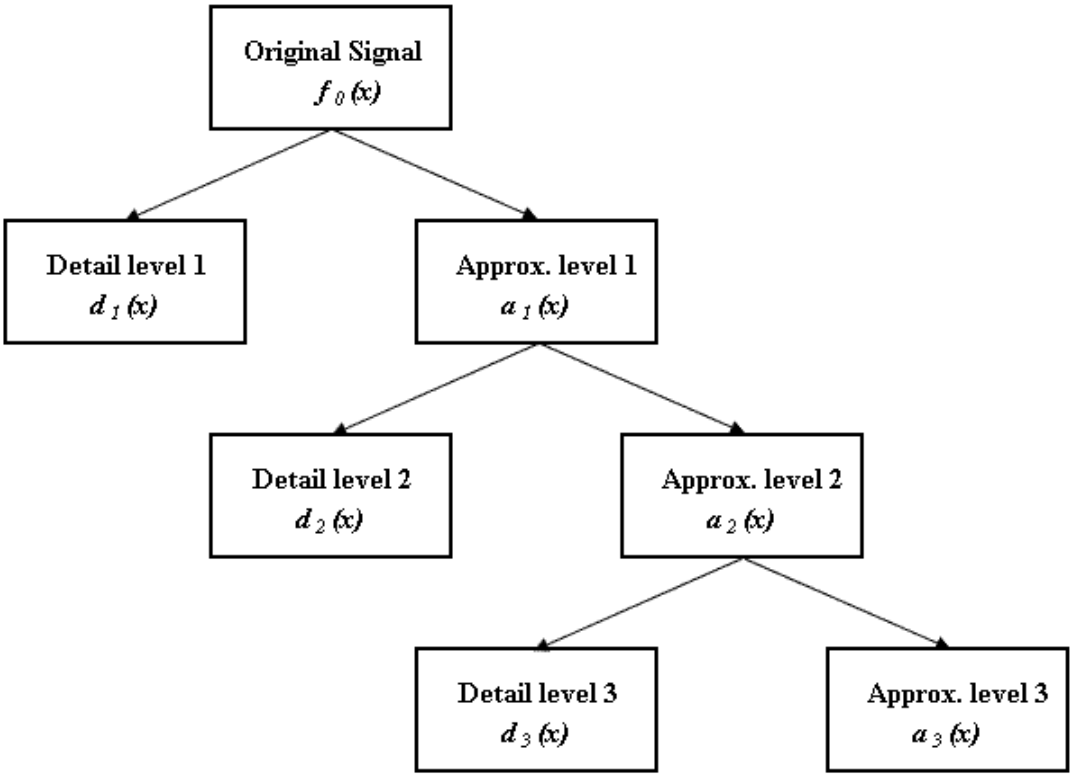


Figure 2.4 - Wavelet transform tree showing three levels of decomposition

Note that once the signal has been sampled, it is no longer a function of the continuous variable x, but is instead a sequence of sample points at $x = 2\pi n/\omega_s$, where n is an integer and ω_s is the sampling frequency; hence, the output signals should strictly be written as, for example, d_{1n} instead of $d_1(x)$, but such notation would be more cumbersome.

To reconstruct the original signal $f_0(x)$, only the three detail signals $d_1(x)$, $d_2(x)$, and $d_3(x)$, and the coarsest approximation signal $a_3(x)$ are needed: $a_2(x)$ can be found from $d_3(x)$ and $a_3(x)$, then $a_1(x)$ can be found from $d_2(x)$ and $a_2(x)$, and finally $f_0(x)$ can be found from $d_1(x)$ and $a_1(x)$. This decomposition also clearly shows that the information in the original signal is transformed without redundancy. The signals at

each successively lower level consist of half the number of sample points as signals in the level above, so the total number of points needed to represent the three detail signals and the one approximation signal are 1/2 + 1/4 + 1/8 + 1/8 of the number of points in the original signal, which sums to exactly the number of data points in the original sampled signal.

2.3 The two-dimensional discrete wavelet transform

The wavelet transform may be extended into two or more dimensions, making it a viable tool for analyzing images. The wavelet transform in two dimensions uses combinations of the scaling function and the wavelet function applied in one dimension at a time. The scaling function acts as a one-dimensional low pass filter while the wavelet function acts as a one-dimensional high pass filter. In two dimensions, four combinations of the two filter types are possible, leading to four transformed images from a single original:

1. horizontal detail – high pass filter vertically, low pass filter horizontally
2. vertical detail – low pass filter vertically, high pass filter horizontally
3. diagonal detail – high pass filters applied both horizontally and vertically
4. approximation – low pass filters applied both horizontally and vertically

The three detail images contain the fine information at the current level of the wavelet decomposition, while the approximation image contains the larger, coarse information that is passed down to the remaining levels of the decomposition.

Dyadic sampling, discussed above for the one-dimensional wavelet transform, can also be applied to the two-dimensional transform. When a filter is applied along a dimension, it is exactly analogous to the one-dimensional discrete wavelet transform, and the number of output data points is half the number of data points in the original signal. Thus, once filters are applied along both dimensions of a discrete two-dimensional signal, such as a pixel map, the output image contains one-quarter the number of data points as the original image. It is notable that since four images are produced at a given decomposition level, the total number of data points produced by the transform is equal to the number of data points in the original image: this agrees with the discrete wavelet transform's property of preserving the information content of a signal without redundancy.

2.4 Example decomposition using two-dimensional discrete wavelet transform

The application of the two-dimensional discrete wavelet transform using dyadic sampling can be illustrated by a sample decomposition of a small 4x4 pixel image using the Haar wavelet. The Haar wavelet is the simplest wavelet basis, consisting of a step function. The discrete form of the wavelet is $(1/\sqrt{2})(1,-1)$, which is proportional to the difference between two successive data points. The corresponding scaling function for the discrete transform is $(1/\sqrt{2})(1,1)$, which is proportional to the average of two successive data points.

Consider the following 4 x 4 array of pixel values:

$$\begin{bmatrix} 0 & 2 & 2 & 0 \\ 0 & 2 & 2 & 0 \\ 0 & 6 & 6 & 0 \\ 0 & 6 & 10 & 0 \end{bmatrix}.$$

We can apply two levels of wavelet decomposition to this image: the first will produce four 2x2 images, and the second will take the 2x2 approximation image and produce four 1x1 images.

Vertical detail:

Applying the high pass filter ($(1/\sqrt{2})(1,-1)$) in the horizontal direction to each row and down-sampling by a factor of two produces a two column, four row image:

$$\frac{1}{\sqrt{2}}\begin{bmatrix} -2 & 2 \\ -2 & 2 \\ -6 & 6 \\ -6 & 10 \end{bmatrix}.$$

The low pass filter ($(1/\sqrt{2})(1,1)$) is then applied vertically to each column and the output is down-sampled by a factor of two to produce the final 2x2 image:

$$\begin{bmatrix} -2 & 2 \\ -6 & 8 \end{bmatrix}.$$

The combined action of the two filters can be combined into a single 2x2 filter that can be applied to the image. Consider the general 2x2 region of pixels:

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}.$$

The high pass filter leaves the following one column, two row region:

$$\begin{bmatrix} (1/\sqrt{2})(a_{11} - a_{12}) \\ (1/\sqrt{2})(a_{21} - a_{22}) \end{bmatrix}.$$

The low pass filter then combines these two values into a single result for the 2x2 region:

$$\left[(1/2)(a_{11} - a_{12} + a_{21} - a_{22})\right].$$

This can be represented in shorthand by the matrix:

$$\frac{1}{2}\begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix}.$$

The sum of the four values produced by multiplying this matrix on an element-by-element basis with a region of the image is then the output pixel in the horizontal detail image. By a similar analysis, 2x2 matrices can be constructed to represent all four filtering operations for the two-dimensional discrete Haar wavelet transformation:

Horizontal Detail

$$\frac{1}{2}\begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}$$

Vertical Detail

$$\frac{1}{2}\begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix}$$

Diagonal Detail

$$\frac{1}{2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

Approximation Image

$$\frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

Similar sets of matrices can be constructed for other wavelet bases besides the Haar basis, simplifying the computation of the transformation. Other wavelet bases typically consist of longer basis function number sequences, making the computations more involved, but the application is exactly the same. For example, the Daubechies family of wavelets are described by 2*N* points for the db*N* basis, and *N* can range from one (equivalent to the Haar basis) to eight or more.

The four output images from the original image are then:

Horizontal Detail

$$\begin{bmatrix} 0 & 0 \\ 0 & -2 \end{bmatrix}$$

Vertical Detail

$$\begin{bmatrix} -2 & 2 \\ -6 & 8 \end{bmatrix}$$

Diagonal Detail

$$\begin{bmatrix} 0 & 0 \\ 0 & -2 \end{bmatrix}$$

Approximation Image

$$\begin{bmatrix} 1 & 1 \\ 3 & 4 \end{bmatrix}$$

The wavelet transform is then applied to the approximation image to produce a second level of detail images. In this case, the four resulting images are 1x1 scalar values and no further decomposition is possible: the approximation image is a scalar value representing the mean intensity of the entire original image and cannot be broken down into further levels of detail.

The four output images from the second level of the decomposition are:

Horizontal Detail

$$\begin{bmatrix} -2.50 \end{bmatrix}$$

Vertical Detail

$$\begin{bmatrix} -0.50 \end{bmatrix}$$

Diagonal Detail

$$\begin{bmatrix} 0.50 \end{bmatrix}$$

Approximation Image

$$\begin{bmatrix} 2.25 \end{bmatrix}$$

Note that the second level of detail captures coarser features in the image: for example, the horizontal detail in the first level misses the change from low to high values between the top and bottom halves of the image, but the second level horizontal detail captures this feature clearly.

CHAPTER 3 – PATTERN RECOGNITION TECHNIQUES

Automated pattern recognition is the process of assigning a new pattern to a class according to an algorithm or classifier [48]. The challenge of pattern recognition is to classify patterns that have not been observed before, based only on the knowledge of the class membership of a known set of patterns. For example, one may wish to classify a piece of produce as either a fruit or a vegetable; based on past experience, one notes that all fruits that they have seen contain seeds and classifies the new piece of produce as a fruit if it contains seeds or a vegetable if it does not. More complex classification schemes are needed when there is more variation within a class, such as the set of all mammographic images showing abnormalities.

Template matching is the simplest form of pattern recognition in imaging. The image to be classified is convolved with an image representative of a class; if the correlation between the images is sufficiently high, then the image is assigned to that class. This method is best in low noise situations with little variation within a class, such as classifying silhouette images of machined parts as either acceptable or defective. This method works poorly when there is large variation in size, shape and orientation of image structures within a class, such as in the variety of ways that different abnormalities appear in mammographic images, making it unsuitable for this work.

The most widely implemented approaches to pattern recognition are statistical and rely on measuring a set of attributes, or features, of a pattern and determining class membership based on their values. The set of all N features measured for a single pattern is called a feature vector, and maps the pattern into an N-dimensional feature space where the classifier can then differentiate between the feature vectors corresponding to particular classes. Features may be continuously valued, such as mean image intensity, or discretely valued, such as the number of foreground objects present in an image.

Classifiers are designated as hard or soft depending on their assumptions regarding the uniqueness of feature vectors. A hard classifier assumes that all patterns that produce nearly identical feature vectors belong to the same class. In this case, when a particular feature vector is measured, it is assigned to a particular class with absolute certainty. Soft classifiers relax this restriction, but instead assume that a certain fraction of feature vectors in a small neighbourhood belong to a particular class and the rest belong to one or more other classes. In this case, a feature vector is assigned a probability p of belonging to the class of interest. A hard classifier can be seen as a limiting case of a soft classifier, where all feature vectors producing a probability greater than some chosen threshold are assigned to the class with certainty. Soft classification provides more useful diagnostic information in medical pattern recognition, since it provides a measure of the confidence in a particular classification and can be used to rate the importance of re-examining a particular patient or image. The main disadvantage of soft classifiers is the difficulty in quantifying the probability of class membership.

3.1 Training and testing methodologies for classifiers

To develop a classifier in practice, some technique is needed to use an input data set to train and test the classifier. Two learning methods are possible, based on whether the class memberships of the patterns in the input data set are known [48]: supervised learning, where the class memberships are known and used; and unsupervised learning, where the classifier is trained without using knowledge of the true class memberships of the input patterns, and in some cases even without knowledge of the number of classes. In supervised learning, some fraction of the input images are used to train the classifier, and the remaining images are used for testing to see how effectively the trained classifier assigns the test patterns to their actual classes. In unsupervised learning, the class memberships are chosen for all of the patterns in such a way as to best separate the data set into some number of classes; various measures exist to determine when the optimal set of class designations have been given. One common measure is to minimize the mean squared distance from the feature vectors in a class to the mean feature vector for that class; this measure makes each class as tightly clustered as possible, rejecting samples whose feature vectors are too distinct from the norm for a given class.

In supervised learning, some fraction of the data set must be set aside for testing, while the remaining samples are used to train the classifier. There are three typical ways to segment a data set for this purpose: the leave-one-out method, cross validation, and the half-and-half method [48].

The half-and-half method uses half the data set for training the classifier, then tests it on the other half of the samples. This method works well for large data sets where half the data set is sufficient to accurately represent the sample population during

training. Because of its inefficient use of a data set for training, though, it is less useful in more complex applications where high in-class variability requires as many samples as possible to accurately model the population; mammographic images display this high in-class variability, and so this method was not used.

The cross validation technique partitions the data set into M equally-sized subsets; one subset is used for testing, and the other M-1 subsets are used to train the classifier. The classifier is trained and tested M times, using each of the subsets as a test set once and as part of the training set the other M-1 times. This method uses a larger fraction of the data set for training while still maintaining a significant number of test samples, making it more suitable for slightly smaller data sets where a limited number of samples are available from each class.

The leave-one-out technique can be seen as a limiting case of the cross validation technique, where the equal subsets each consist of only one sample. In this case, all but one of the samples are used to train the classifier, and the classifier is tested on the lone remaining sample. The overall performance of the classifier is measured by averaging the classification results from when each sample in the data set was used as the test sample. This method maximizes the size of the training set, making it a viable choice for testing the mammographic x-ray image classifier designed in this work. Since there is so much variability in the appearance of normal and of abnormal mammographic x-ray images, the training set size should be maximized to best represent the distribution of samples within these two classes.

3.2 Common types of classifiers

There are several widely used classifiers in image recognition research today; among them are *c*-means classifiers, *k*-nearest neighbour classifiers, neural networks, and Bayesian classifiers [48]. All four classifiers are naturally hard classifiers, although confidence measures may be extracted from each by extending their basic algorithms. The classifiers all work in two phases: the training phase takes a set of feature vectors with a known class designation to give the classifier reference values to measure against new vectors, and the operating phase takes new feature vectors as input and assigns class labels based on the information gained during the training phase.

3.2.1 *C*-means classifier

A *c*-means classifier defines a prototype feature vector for each class as the average of all training vectors for that class. The feature vector of a new sample is assigned to the class with the nearest prototype vector. This algorithm trains quickly and performs the classification quickly as well, making it useful as a first pass classifier for testing the effectiveness of new features. This type of classifier can be used for an unsupervised learning experiment by iteratively choosing potential class memberships and calculating the corresponding prototype feature vectors until some desired segmentation rate is achieved among the different classes. Because class distinctions are determined by a simple distance measurement, however, this approach fails when the boundaries between classes in feature space are non-linear.

3.2.2 *K*-nearest neighbour classifier

A *k*-nearest neighbour classifier assigns a new feature vector to the class to which the largest fraction of the *k* nearest training vectors belong. Nearness is measured by the geometric distance between two feature vectors in a space defined by the component features used to construct the vector. This algorithm does not require a distinct training phase, other than inputting the training vectors and their classes, but has a relatively slow operating phase, since the distance to every training vector must be measured for each new sample, though optimizations are possible to reduce the number of measurements needed for each new vector. The classifier can handle classes with complex boundaries in feature space: since it is only concerned with the *k* training vectors in the local neighbourhood, it can handle complex cases, such as classes with two disjoint regions in feature space, quite well. The robustness, ease of implementation, and relatively high efficiency, when optimized, of this algorithm make it highly popular, especially in research where the emphasis is on generating new types of feature vectors and measuring their performance directly against older types of feature vectors using a common classifier.

3.2.3 Neural networks

A neural network uses a complex set of decision nodes to create non-linear relationships between the input feature vector and the output class designation. The *N* features of a feature vector are input to *N* nodes of a directed graph; each node takes in a set of input values and produces a single output value based on a weighted sum of its inputs. The output may be binary, where the output is 1 if the inputs are above a

particular threshold and 0 otherwise, or it may be continuous, such as $1/(1+e^{-x})$ where x is the weighted sum of the inputs. Several layers of nodes can be arranged and connected to determine a unique output for a given input vector; for example, in a classification scheme with four classes, the output may consist of four nodes defining a class vector with a 1 for the component representing the class chosen for the input and zeroes for the other components of the output. A neural network is trained by repeatedly updating the values of the weights at each node until the vectors in the training set are classified at a sufficiently high rate. In particular, the back-propagation technique measures the error at the output and feeds this back through the network, adjusting the weight at each node by an amount related to the magnitude of the misclassification at the output [48]. The iterative training process can be very time consuming, but the classification of new feature vectors afterwards is quite rapid.

3.2.4 Naïve Bayesian classifier

A Bayesian classifier uses probabilistic measures to assign a feature vector to the class most likely to produce it. Specifically, the classifier uses Bayes' rule to measure the conditional probability *P(f|c)*, the probability that class *c* could produce feature vector *f*, for all classes *c* and assigns *f* to the class with the highest probability of producing it:

$$P(c \mid f) = \frac{P(f \mid c)P(c)}{\sum_i P(f \mid c_i)P(c_i)}. \qquad (3.1)$$

Equation 3.1 is the basic equation used to calculate probabilities for a Bayesian classifier. Here, *P(c|f)* is called the posterior probability and describes the likelihood of

feature vector *f* belonging to class *c*. *P(f|c)* is called the likelihood, since it describes the likelihood that a class *c* would produce feature vector *f*. *P(c)* is called the prior probability for class *c*, and describes the *a priori* probability of class *c* occurring, compared to all other classes; since only about 1 in 400 mammographic x-ray images are abnormal, *P(c)* would be 0.0025 for the abnormal class and 0.9975 for the normal class. Finally, the denominator is called the evidence and normalizes the posterior probability so that the probability of feature vector *f* belonging to any class sums to one.

A naïve Bayesian classifier, the most common implementation, makes the strong assumption that the components of the feature vector are independent, so that $P(f|c)$ is equal to the product of the probabilities $P(f_i|c)$ that class *c* could produce each component f_i of the feature vector. This independence assumption is often demonstrably wrong, since many components of a feature vector may share some relationship, but the naïve Bayesian classifier can still produce very high classification rates with this assumption [52]. It is believed that the independence assumption does not destroy the validity of the classifier because, although it changes the probabilities $P(f_i|c)$, the Bayesian classifier makes decisions based on which probability is largest, and the independence assumption does not typically change the relative ordering of the probability magnitudes [21]. Because the magnitudes of the probabilities are not valid, however, the naïve Bayesian classifier cannot use them to determine a confidence measure for its classifications. The classifier is trained by measuring the conditional probabilities $P(f_i|c)$ from the training vectors; typically binning is used to discretize continuously-valued features to form a useable probability distribution. The Bayesian classifier takes longer to train than the other classifiers, except for neural networks, but

classifies new patterns quickly and often very accurately; the challenges with this classifier are binning the feature values effectively and incorporating soft classification.

This research used a novel extension of a naïve Bayesian classifier to classify mammographic x-ray images into one of two possible classes – normal or suspicious – and provide a confidence level for the classification.

3.3 Survey of current approaches in computer aided detection

In recent years, a large number of Computer Aided Detection (CAD) methods have been proposed to assist radiologists in the detection of breast cancer from x-ray mammography images. These approaches use a variety of image processing techniques and pattern recognition tools to extract information from raw mammograms; the differing approaches often reflect the different physical and biological phenomena that indicate abnormalities.

A number of comprehensive review articles cover the variety and effectiveness of current CAD methods; the following sections give only a sample of current approaches and techniques. S. Ciatto *et al.* [12] examined the CAD readers currently used in clinical settings as second readers to measure their effect on diagnostic accuracy. Susan Astley provided two discussions of the potential of CAD systems and the improvements that must occur for them to become effective [3,4]. Kunio Doi discussed the history of CAD and the limitations that make it an aid but not a replacement for human radiologists [16]. Maryellen Giger examined the state of CAD and its use in other modalities, such as ultrasound and MRI, where similar challenges face physicians attempting to process a large number of images to find a small number of abnormal

cases [20]. Finally, H. D. Cheng *et al*. provided a particularly comprehensive overview of techniques common to most CAD systems, especially image processing techniques [9,10].

Three representative CAD methods are discussed in some detail below to provide a context for the advantages of the method developed in this work. Other CAD methods exist; the methods discussed here are those with some common features to the current work, such as the use of wavelets or of statistical pattern recognition techniques.

3.3.1 Spatial grey level dependence (SGLD) matrices

The team led by Heang-Ping Chan [7,8] developed a classification tool that differentiated between benign and malignant calcifications in mammographic images using a textural analysis and a fuzzy-neural network. This approach was only designed to differentiate between benign and malignant calcifications and not to analyze an arbitrary image, so the algorithm acted only on a region of interest (ROI) defined by a radiologist: the ROI contained a cluster of calcifications that the classifier analyzed for malignancy.

The analysis of the ROI began by removing the relatively low spatial frequency background to isolate the small, high spatial frequency calcifications. The resulting region was cropped to 512 x 512 pixels, and 40 SGLD matrices $p_{\theta,d}(i,j)$ were developed for each region. Each matrix $p(i,j)$ for a given θ, d pair measures the number of times that the grey levels i and j are found a distance d apart and at an angle θ from each other; because the region is discrete due to digitization, the angle θ was limited to four values – 0, 45, 90 and 135 degrees – and the distance d was limited to have the values from 4 to

40 pixels in four pixel increments. The grey levels were binned to have only four possible values; thus, each of the 40 matrices measured 4 x 4 elements in size. From these matrices, 13 textural features were measured, such as energy, entropy, and inertia, which are statistical characterizations of the nature of grey level changes within the region [22].

The smallest subset of features which provided the maximal discrimination between the benign and malignant classes were chosen iteratively by adding or removing one feature at a time from the complete feature set. The Wilks lambda function was used to measure the separation between the two classes and is equal to the ratio of the within-class sum of squared errors to the sum of squared errors for the total data set. A feature was added if its effect on the Wilks lambda was above a chosen threshold F_{in}, while a feature was removed if its effect on the Wilks lambda was below a chosen threshold F_{out}; raising either threshold lowered the total number of features chosen for the final classifier, an artificial neural network. The classifier was trained with the leave-one-out methodology, and typically selected six to seven features to classify the 86 images. The classifier was able to detect 100% of malignant cases while correctly identifying 11 of 28 (39%) of the benign calcification cases.

3.3.2 Multiresolution detection of spiculated lesions using binary tree classifier

Sheng Liu, Charles F. Babbs and Edward J. Delp proposed a detection system for identifying spiculated lesions on mammograms [33]. Spiculated lesions are masses that lack a smooth boundary and have an irregular shape; these masses are far more likely to be cancerous than smooth lesions, so their detection is of great diagnostic

importance. Wavelet decomposition was used on the image to generate coarser, low resolution versions of the image, which allowed the same algorithms to be applied to different size scales. For example, a region of interest 20 pixels in diameter would correspond to regions 1, 2, 4 and 8 mm in diameter as the resolution was successively halved, making the algorithm sensitive to lesions of various sizes. The wavelet used in the decomposition was a linear phase non-separable 2D wavelet: this means that the phases were transformed linearly and that the wavelet could not be decomposed into separate horizontal and vertical transformations. This wavelet was used to avoid bias towards the horizontal and vertical orientations in the image, since the masses could have arbitrary orientation.

The distinguishing feature of spiculated lesions used in the analysis is the random orientation of their edges and spicules compared to the relatively parallel orientation of the ductile tissue within the healthy region of a breast. To measure this lack of directionality, four parameters were measured for each pixel at each of four levels of image resolution, using a neighbourhood 30 pixels in diameter to measure statistical values. The four parameters were mean intensity, standard deviation of pixel intensity, standard deviation of the edge orientation histogram and the standard deviation of the intensity gradient. The edge orientation histogram measures the number of pixels in the neighbourhood with each possible edge orientation; for healthy tissue, edges should be relatively parallel and the histogram should have a narrow peak and a small standard deviation, while for a spiculated lesion the edges should be more randomly oriented, resulting in a broad or non-existent peak and a larger standard deviation.

The images were then classified, starting with the coarsest: once an image had a positive result, finer resolutions were not analyzed to reduce computational complexity; images with negative results at all resolutions were combined with a weighted average to make a final search for any missed lesions. Each pixel within an image was run through a binary tree classifier, which gave a measure of suspiciousness, and the results for the image were run through a median filter, a low pass filter and then thresholded to mark any regions above a certain level of suspiciousness that were most likely to correspond to a spiculated mass. A binary tree classifier makes a decision at each node based on the value of a particular feature so that the path continues to the child selected by that decision, and the leaf nodes correspond to a particular class designation and a confidence level; in this case there were two class designations: suspicious and not suspicious.

This scheme was tested on 19 images of spiculated lesions and 19 normal images taken from the MIAS database also used in the current work [51]. The algorithm achieved 84.2% sensitivity with 1 false positive per image or 100% sensitivity with 2.2 false positives per image, depending on the choice of the threshold when selecting suspicious pixels from the output of the classifier.

One drawback of this approach is the decision to stop searching finer resolutions once a positive result is found. Firstly, because of the large number of false positives per image, the majority of suspicious regions marked are false positives, and finding a positive result at a coarse resolution does not mean that a true positive does not exist at a finer resolution. Secondly, the finer resolution images are most sensitive to the small lesions that mammography can find better than physical examination; stopping at coarse

resolutions limits the size of lesions that will be found and may reduce the effectiveness of mammography when used with this approach.

3.3.3 Multiresolution segmentation of calcifications using fuzzy c-means analysis

The primary focus of the work by S. Sentelle, C. Sentelle and M. A. Sutton was to use wavelet analysis to rapidly process images to detect calcifications, as processing time is a problem for a number of current algorithms [47].

The algorithm first down-sampled an image, taking one pixel from each 64 x 64 region in order to more rapidly process the high resolution mammography images. The down-sampled image pixels were segmented into seven classes according to their intensity: bright structures, background, adipose tissue, glandular tissue, and three classes for the air-skin interface region. The segmentation was performed using fuzzy c-means analysis, which iteratively updates the location of class prototypes in a feature space and assigns samples to the class with the nearest prototype. After each iteration, the centroid of all points assigned to a particular class is used as that class's prototype for the next iteration; the algorithm stops iterating when the adjustments made to the prototypes after an iteration are all below a chosen threshold, at which point the class memberships are finalized. Because this method is statistical, its accuracy is not significantly reduced by using a smaller set of samples, so long as they are representative of the entire population; this means that the prototypes determined from the down-sampled image are also applicable to higher resolution images.

The class prototypes found from the first down-sampled image were then directly applied to classify the pixels of an image down-sampled by a factor of 4 in each

direction from the original image. This mid-resolution image was coarse enough to make class assignment rapid but fine enough to accurately segment the image into the seven classes. Windows of the image containing enough pixels classified as bright structures that could potentially be calcifications were then located and examined at full resolution Wavelet decomposition was performed on the windowed region using the biorthogonal 4.4 wavelet base, and the resulting wavelet maps were tuned to enhance the appearance of small, bright regions that could correspond to calcifications. A potential calcification was marked if the image contrast and intensity at a point were together above a chosen threshold.

The algorithm was tested using 20 images containing malignant calcifications and 5 images of normal breast tissue. The algorithm achieved 75% sensitivity with 3.0 false positives per image and 94% sensitivity with 17.0 false positives per image. The relatively low specificity was a compromise with the fast processing speed of the algorithm, which was relatively high due to the selection of windows of interest from the low resolution image before performing the wavelet analysis and calcification detection.

CHAPTER 4 - METHODOLOGY

4.1 Introduction

The primary objective of this research was to design a tool that could accurately determine whether a given mammography image contained abnormalities that could signify breast cancer. This chapter discusses the components of the tool's design. Section 4.2 discusses the full system and breaks it down into its component stages, each of which are discussed in subsequent sections of this chapter.

The digitized images were first pre-processed to remove artifacts and reduce noise, making the images as uniform as possible to highlight meaningful differences between images; this step is discussed in Section 4.3. Section 4.4 describes the decomposition of the images using wavelet analysis a set of wavelet maps and discusses the generation of scalar features from these maps. Section 4.5 introduces the modified naïve Bayesian classifier used in this work. Section 4.6 outlines the problem of feature selection, or choosing the smallest subset of the generated features that optimizes classification efficiency. Finally, in Section 4.7, the novel concerted-effort set of classifiers and the process of constructing a network of classifiers working in tandem are discussed.

4.2 Complete image analysis system

The entire image analysis system, from the reading of the original image to the final classification as either normal or suspicious, is represented by the block diagram of Figure 4.1. The system consists of two distinct stages: the image processing system, which reads in the original image and produces a set of wavelet map images for the classifier to use; and the classification system, which measures features from the wavelet images and classifies the image as either normal or suspicious based on the results from the ensemble of classifiers.

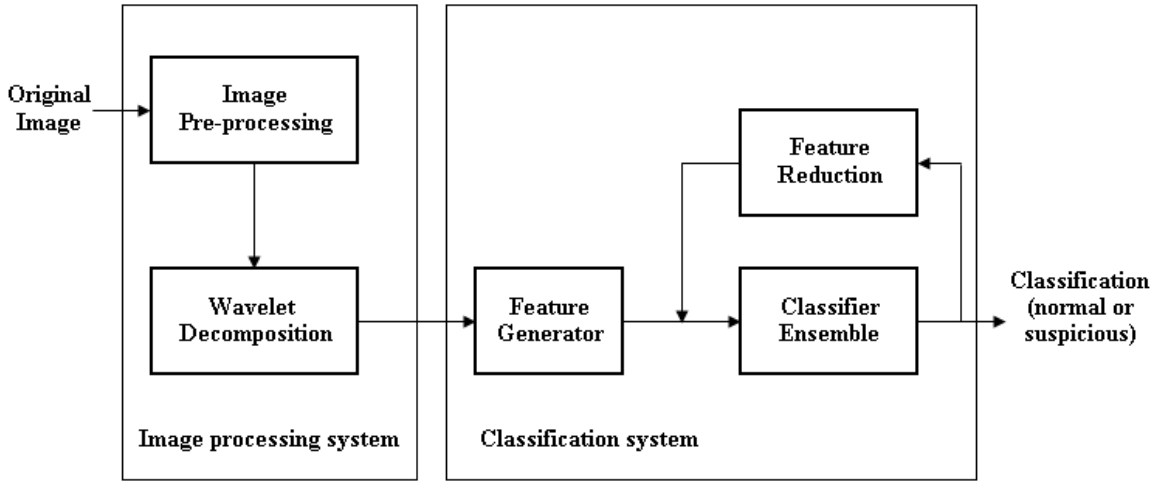


Figure 4.1 – Block diagram of complete image classification system

The image processing system consists of two discrete stages: the image pre-processor and the wavelet decomposition system. The image pre-processor takes the original digitized mammography image as an input and outputs a normalized image. The normalized images are flipped, if necessary, to all point in the same direction, have had their intensities scaled to a common maximum value, have had background artifacts removed and have had the background thresholded to zero to remove noise. The pre-

processing system is discussed in detail in Section 4.3. The wavelet decomposition system takes the normalized images and carries out a wavelet analysis on them using one of a number of possible wavelet bases. Multiple levels of decomposition are used, and four images are produced at each level of the decomposition, so that the output of the wavelet decomposition system is a set of images forming the wavelet maps of the normalized version of the original input image. The wavelet decomposition scheme used is discussed in Section 4.4.

The classification system consisted of three stages: the feature generator, the classifier ensemble, and the feature reduction system. The feature generator read in the set of wavelet maps produced by the image processing system and reduced them to a set of scalar features. Section 4.4 discusses the generation of features from the wavelet maps.

The classifier ensemble consisted of a set of naïve Bayesian classifiers working in tandem to produce a single classification output of normal or suspicious using the features from the feature generator as input. Section 4.5 discusses the design and tuning of a single classifier in the ensemble, while Section 4.7 discusses the process of linking several classifiers together to form the ensemble.

Feature reduction was used to minimize the number of features used to make the classification. The problem of feature selection is discussed in Section 4.6. Feature selection is also closely related to the problem of training and testing the classification system; it is discussed in that context in Chapter 5.

4.3 Image pre-processing

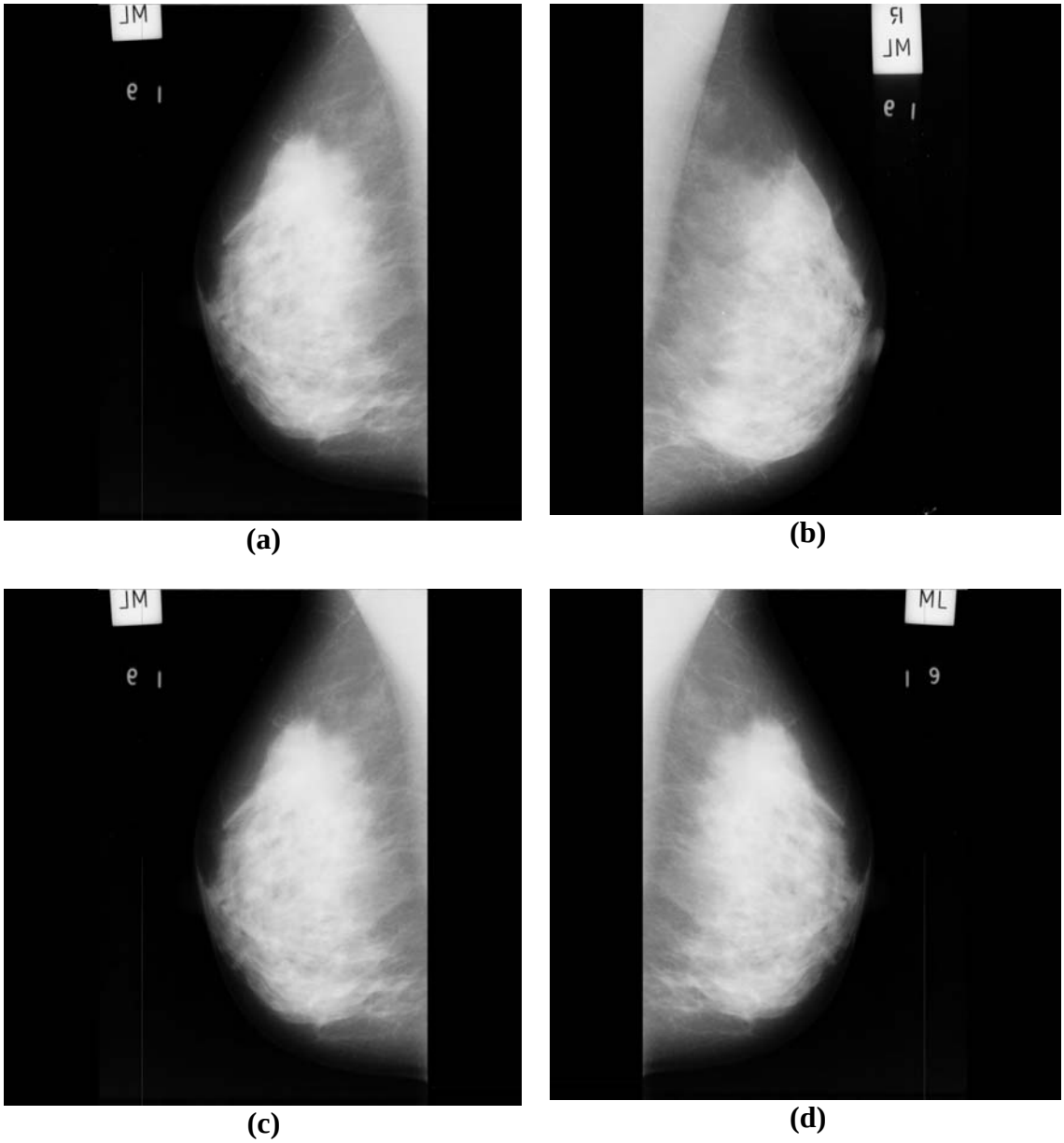
The images to be analyzed in this work were taken from the Mammographic Images Analysis Society's digital mammogram database [51] that consisted of 303 images: 205 images of normal breasts and 98 images showing one of four pathologies: benign masses, cancerous masses, benign calcifications or cancerous calcifications. There were an approximately equal number of images of right and left breast images, all of which were medial-lateral images. A larger second data set taken from the Digital Database for Screening Mammography [23] was also used to test the system; this data set consisted of 1714 medial-lateral images, including 1065 normal images and 649 images showing some form of pathology. Images from both databases contained artefacts and noise unrelated to the presence or absence of abnormalities in the breast that needed to be addressed.

In order to reduce the influence of information content not related to pathology, several pre-processing steps were implemented to regularize the appearance of the images and remove any unnecessary artefacts. The steps taken were: orientation matching, background thresholding, artefact removal and intensity matching.

4.3.1 Orientation matching

Because the images of right and left breasts point to the left and right sides of the image, respectively, the images of right breasts were flipped horizontally to have the same orientation. This step ensured that all images pointed in the same direction, preventing changes in the wavelet transform coefficients due only to the directionality change between right and left images. A major feature in all images that this affects is

the sharp vertical line between the tissue and the dark background where the original film ends: on left breast images the intensity rises left to right across this edge, while on right breast images it falls, changing the sign of the calculated wavelet coefficient; matching the orientations of all images prevents this type of artefact from appearing in later analysis.



(a) **(b)** **(c)** **(d)**

Figure 4.2 – Right (a) and left (b) breast images, no abnormalities. Right breast image before (c) and after (d) orientation matching.

Figure 4.2 shows the process of orientation matching. Figures 4.2 a) and b) respectively show the left and right breast images of a patient with no abnormalities in her breast tissue. Figures 4.2 c) and d) show the right breast before and after being reflected. Note that after flipping the right breast images, the two images from the patient are qualitatively much more similar, reducing errors in subsequent analysis.

4.3.2 Background thresholding

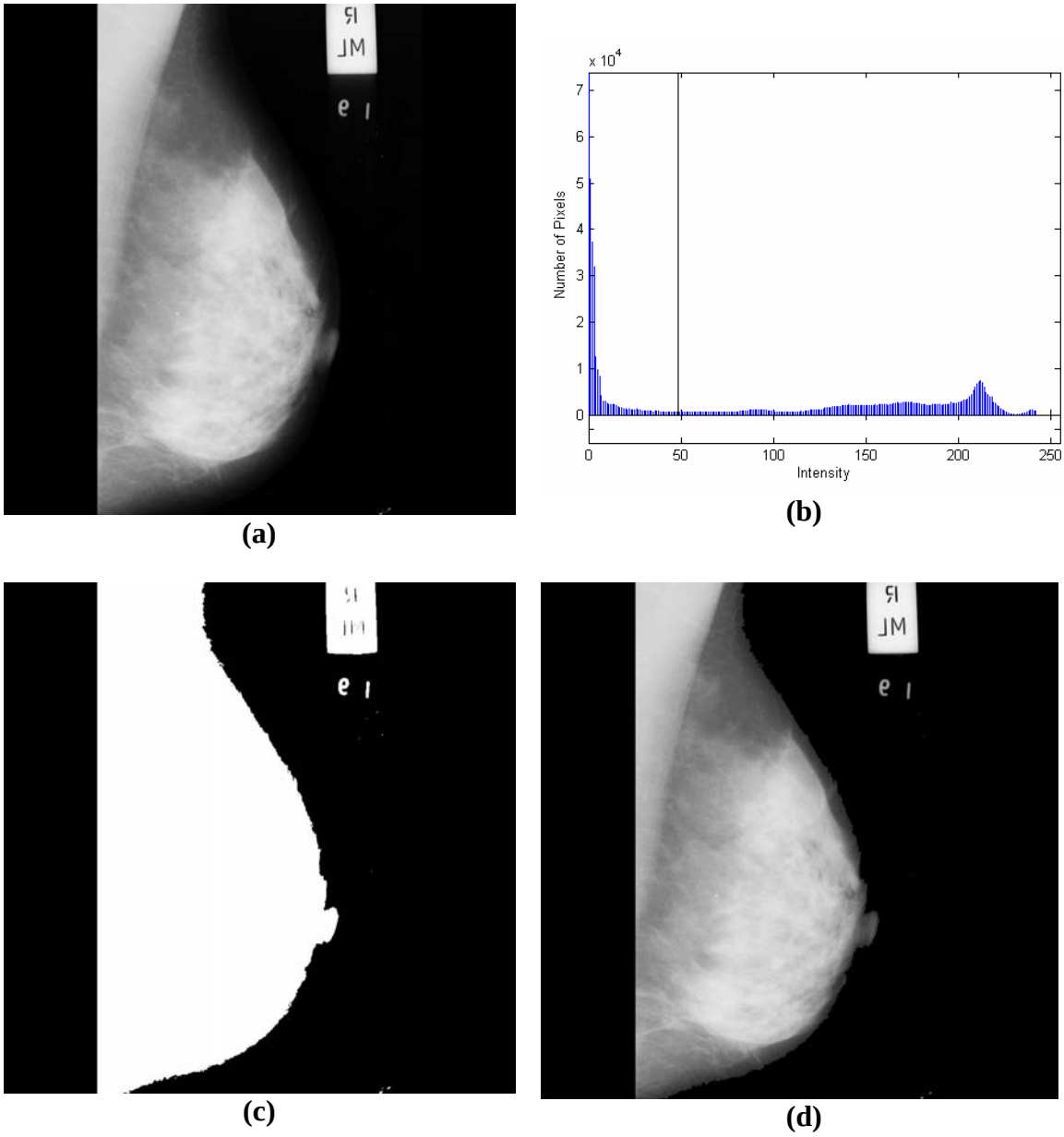
Since the pathology information is contained entirely within the tissue region of the image, no signal should be present in the dark background. Semi-thresholding [48] is a technique which sets all pixels below a set intensity level to zero; if the threshold is chosen well, this procedure can zero out the majority of background pixels, which do not contain useful signal, and leave foreground objects unaltered. To accomplish this, a conservative threshold value was chosen to be half of the threshold predicted by Otsu's Method [48]. This method assumes that the distribution of grey levels within the image is bimodal, consisting of a largely bright object of interest against a largely dark background. The brightness of the breast tissue region relative to the dark background in the images makes them appropriate for Otsu's method.

In Otsu's method, a threshold level is chosen such that all background pixels have an intensity below the threshold value and all foreground object pixels have intensity above the threshold value. In practice, the intensities of foreground and background pixels will both have broad, overlapping distributions, and it will not be possible to choose a threshold value which does not misclassify any pixels. Otsu's method assumes that the foreground and background intensities are normally distributed

and chooses the threshold level which minimizes the number of misclassified pixels between the background and foreground regions.

In this work, the actual threshold level used is half of that predicted by Otsu's method; this is done to reduce the number of pixels from the tissue which are misclassified as background pixels and thus removed from the image. Since the intensity in the images is directly related to the attenuation of x-rays passing through the tissue, and since this attenuation depends largely on the thickness and density of the tissue, tissue pixels which fall below the conservative threshold are predominantly from the edges of the tissue region where the breast tissue is thin and uncompressed. Any pathology which exists this close to the surface of a patient's skin should be readily detectable by conventional examination without the aid of mammography.

The thresholding process was implemented in conjunction with the next step, artefact removal; because of this, it was convenient to perform a binary thresholding, where all pixels below the threshold were set to an intensity of zero and all pixels above the threshold were set to an intensity of one. The resulting image was used as a mask for the original image: the output image of the process was the pixel-by-pixel product of the binary image with the original, so that all background pixels were set to an intensity of zero and all foreground pixels were unaffected.



(a) **(b)**

(c) **(d)**

Figure 4.3 – Mammogram image before(a) and after(d) background thresholding. The intensity histogram is shown in (b) with the threshold indicated by a vertical line; (c) shows the thresholded binary image used to mask the original image.

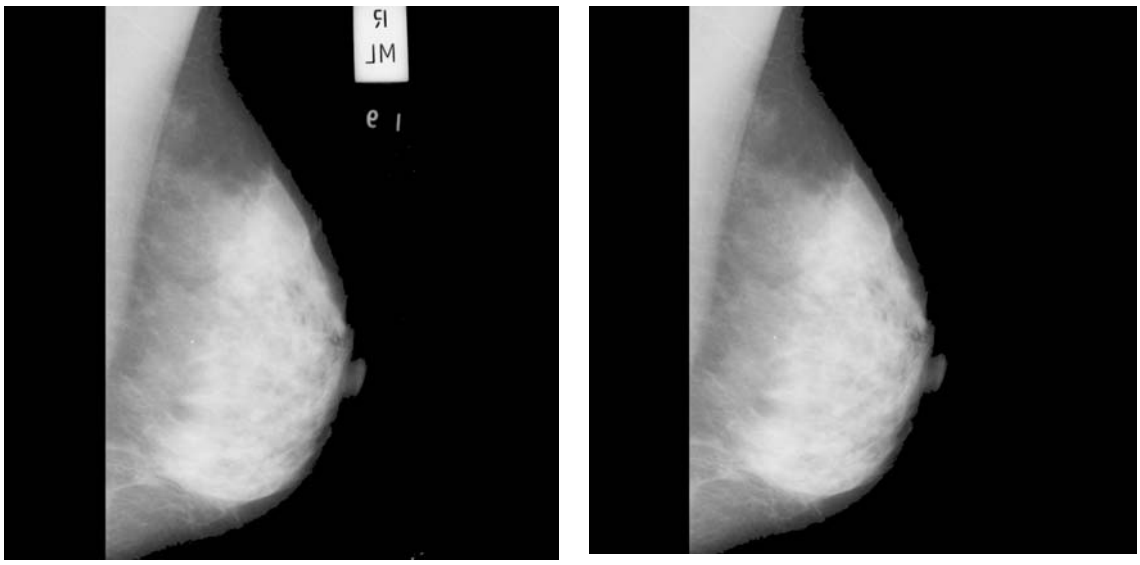
Figure 4.3 shows the process of thresholding the image. Figure 4.3 a) is the original image after orientation matching. Figure 4.3 b) shows the intensity histogram of the image: note that the distribution has a narrow peak at low intensity corresponding to the large number of very dark background pixels and a broad peak at higher

intensities corresponding to the brighter foreground tissue region. The vertical line in the histogram corresponds to the conservative threshold value; all pixels with intensities below the threshold appear black in the binary image in Figure 4.3 c) while all pixels with higher intensities appear white. Figure 4.3 d) shows the masking of the original image in a) with the binary image in c): note that the background has had any low intensity noise removed; the effect is subtle in the printed figure but is apparent near the "ML" view tag[2] in the upper-right corner that now appears much more sharply.

4.3.3 Artefact removal

Once the binary image has been created through thresholding, it is possible to remove artefacts in the image. The image is labelled as a set of contiguous regions, typically 100 – 125, consisting of 4-connected groups of white pixels in the binary image. Two pixels are part of a 4-connected region if it is possible to travel between them by moving only vertically or horizontally (not diagonally) from pixel to pixel, travelling only on pixels that are part of the 4-connected region. To preserve the tissue, all labelled regions are removed from the image except for the one containing the centre pixel in the image: since the breast tissue is centred in all images, this approach isolated the breast tissue in all 303 images used. The most noticeable artefact removed is the "ML" view tag, although a large number of small regions in the background which were brighter than the threshold level in the previous step are caught and removed by this procedure. Figure 4.4 shows an image before and after the artefact removal step has been applied.

[2] This tag appears on the original mammography x-ray film, specifying that this is a medial-lateral, or side view. The tags are metal plates which strongly absorb x-rays and thus appear extremely bright on the developed films.



(a) **(b)**

Figure 4.4 – Mammography image before (a) and after (b) artefact removal procedure

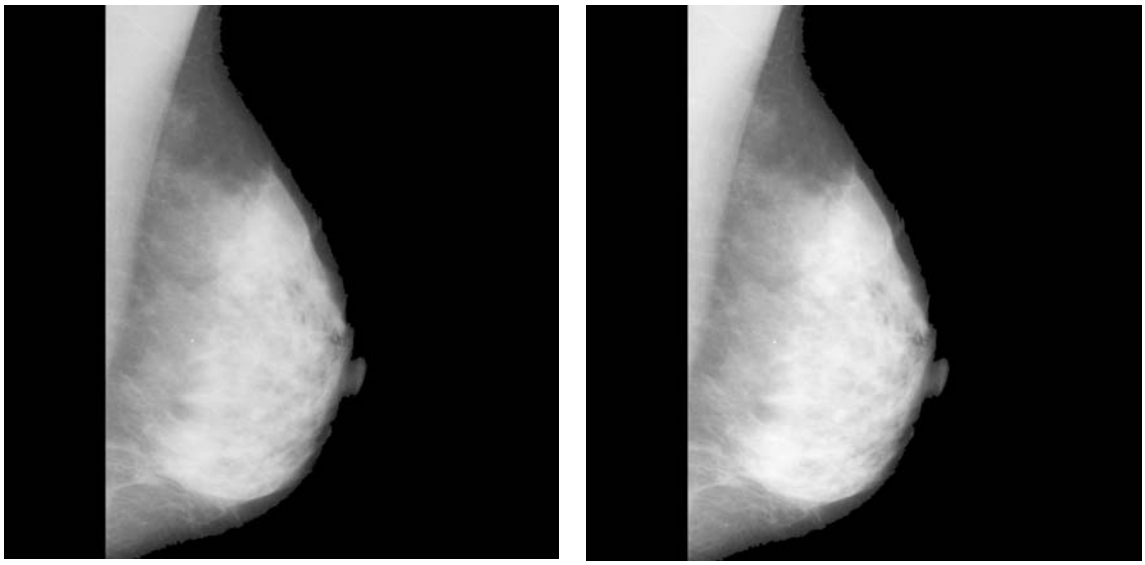
4.3.4 Intensity normalization

The final pre-processing step applied to the images before they are ready for wavelet decomposition is the intensity normalization step. This step scales all images so that the intensity of the brightest pixel in the image has a relative intensity of 1.0 and linearly scales all other image pixels accordingly. The transformation is described by:

$$img\_out = \frac{img\_in}{\max(img\_in)}, \tag{4.1}$$

where *img_in* is the input image following the artefact removal step and *img_out* is the intensity matched image whose pixel intensities range from zero to one. This step ensures uniformity across different images, which may be taken at different times under slightly different machine settings or by different personnel. Figure 4.5 shows the subtle

difference created by the intensity matching procedure. The broader spread in intensities (the maximum relative intensity prior to normalization was 0.92) creates stronger differentiation between variations in tissue types and densities in Figure 4.5 b).



(a) **(b)**

Figure 4.5 – Mammography image before (a) and after (b) intensity matching procedure

4.4 Wavelet decomposition of processed images

Once the images were pre-processed to minimize the differences between images that are not related to differences in the physical composition of the breast tissue, wavelet analysis was performed on the images. The two-dimensional discrete wavelet transform was discussed in detail in Section 2.3 and was applied directly to these images. The images were all sampled to 1024x1024 pixels, which would allow 10 levels of decomposition, since dyadic sampling reduces the dimensions by a factor of two in each direction after each pass. The images produced at each level of the decomposition were then sensitive to structures in the tissue of different sizes, making

this approach effective for detecting the large variety of abnormalities in the tissue that could indicate breast cancer.

In practice, only eight levels of decomposition were used, for several reasons. Since the final two levels would consist of four pixel and one pixel images and would be sensitive only to structures comparable in size to the entire breast, these levels were omitted from the wavelet analysis to speed calculation. This limitation of the number of levels in the decomposition reduced the computation time for the complete system by a factor of 2.44. More importantly, the lowest level of the decomposition had only one data point, and so images from this level would have all had a standard deviation of zero and undefined skewness and kurtosis, the statistical measures used to compare among different images. Maps in the first eight levels of the decomposition each have at least 16 pixels, making statistical calculations more meaningful than for the four or one pixel images at the lowest two levels of the decomposition.

4.4.1 Choice of wavelet basis

A large number of wavelet bases have been developed in the literature, and new wavelet bases may be constructed easily, making the selection of an optimal basis for a particular task a complex project on its own. For this research, a representative sample of widely used wavelets were tested. Though this approach was unlikely to find the optimal wavelet basis for this type of signal processing, it is believed that the use of any sufficiently carefully chosen wavelet base can produce strong results, and the difference between optimal and sub-optimal wavelet bases is typically small [25]. Eleven wavelet bases were tested in this work: the Haar wavelet; the Daubechies 2, 4 and 8 wavelets;

and the biorthogonal 1.5, 2.2, 2.8, 3.7, 4.4, 5.5 and 6.8 wavelets. The Haar wavelet was chosen for its simplicity and for its effectiveness at detecting sharp contrasts, such as the presence of microcalcifications against a relatively dark background. The Daubechies wavelets were chosen for their sensitivity to various types of intensity gradients. The biorthogonal wavelets were chosen for their ability to provide exact reconstruction. Figure 4.6 shows the Haar wavelet and the 3 Daubechies wavelets used in this work, and Figure 4.7 shows the 7 Biorthogonal wavelets used in this work. The wavelets and their associated scaling functions are shown in their discrete form, since this was the form used to decompose the mammographic images. Note that the wavelet functions correspond to the high pass filters and the scaling functions correspond to the low pass filters applied along either the horizontal or the vertical directions in an image.

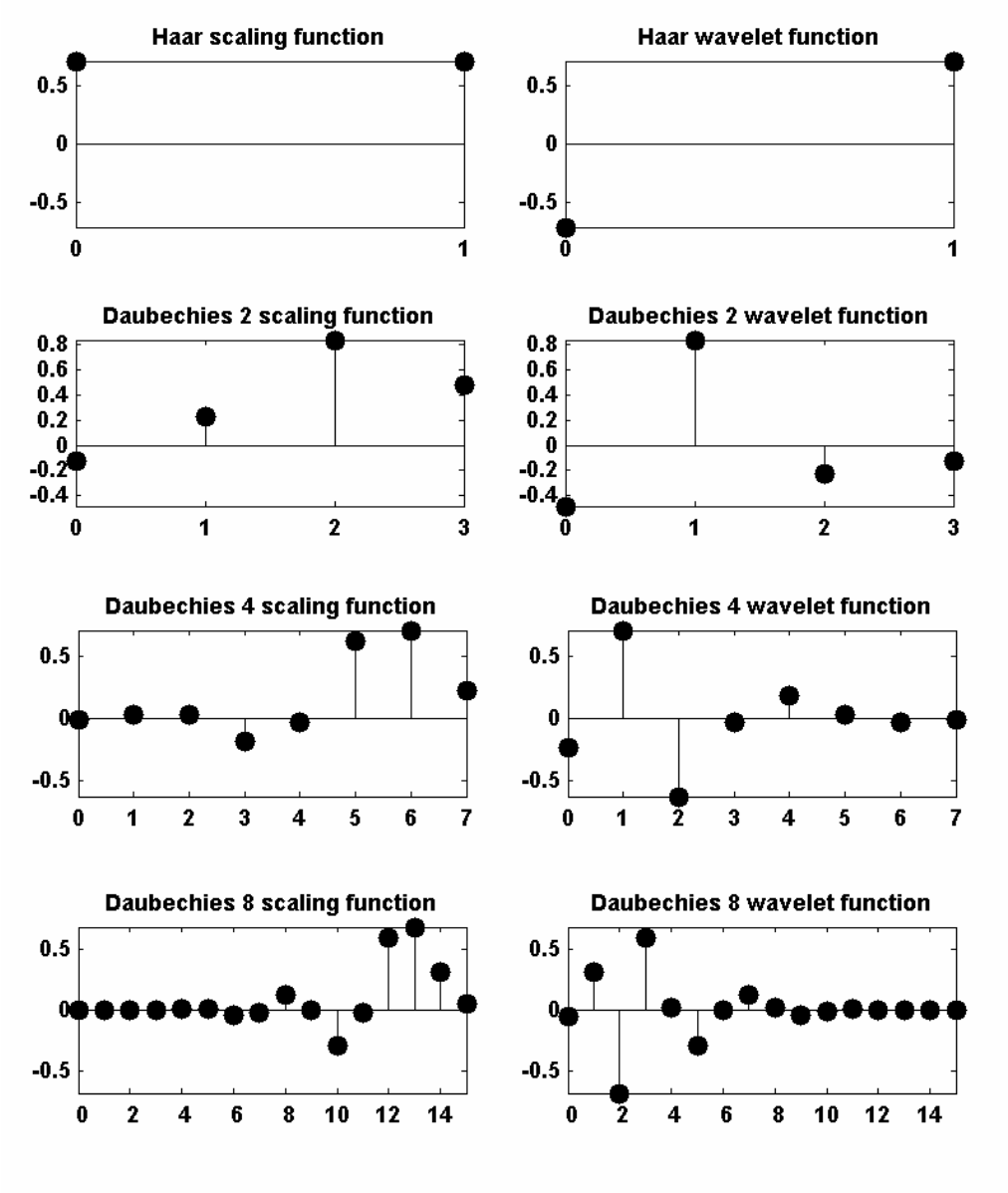


Figure 4.6 - Wavelet functions (high pass filters) and scaling functions (low pass filters) for Haar and Daubechies wavelet bases used in this work

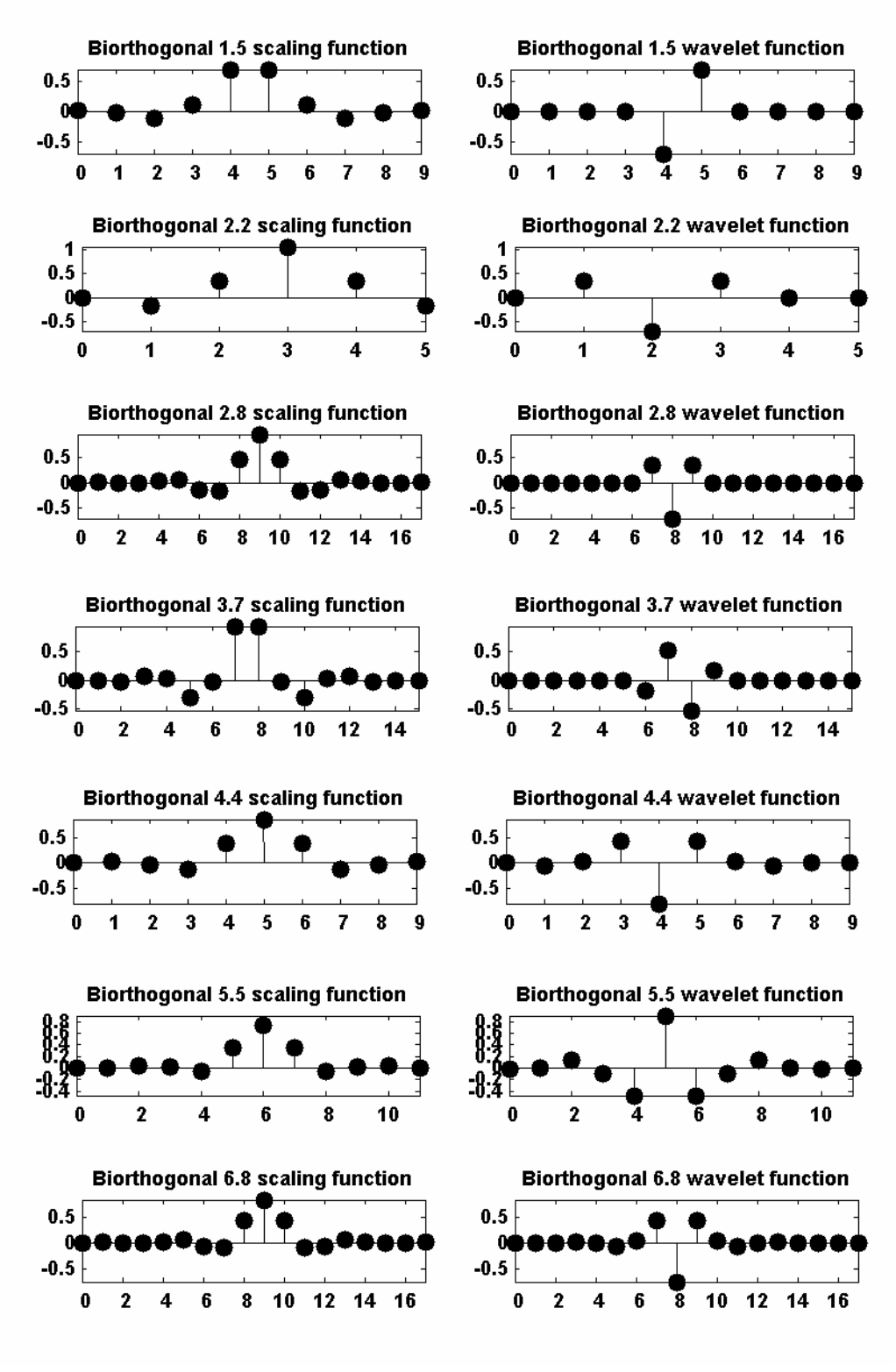


Figure 4.7 - Wavelet functions (high pass filters) and scaling functions (low pass filters) for Biorthogonal wavelet bases used in this work

Figure 4.8 shows the four views obtained at the third decomposition level when the Haar wavelet basis is used. Note that the wavelet maps have a lower resolution than the original image, and that each view is sensitive to different features in the image: the horizontal detail detects vertical changes in intensity, the vertical detail detects horizontal changes in intensity, the diagonal detail responds when the intensity is varying in both directions, and the approximation image is a low resolution version of the original image used as an input to the next coarser level of the decomposition.

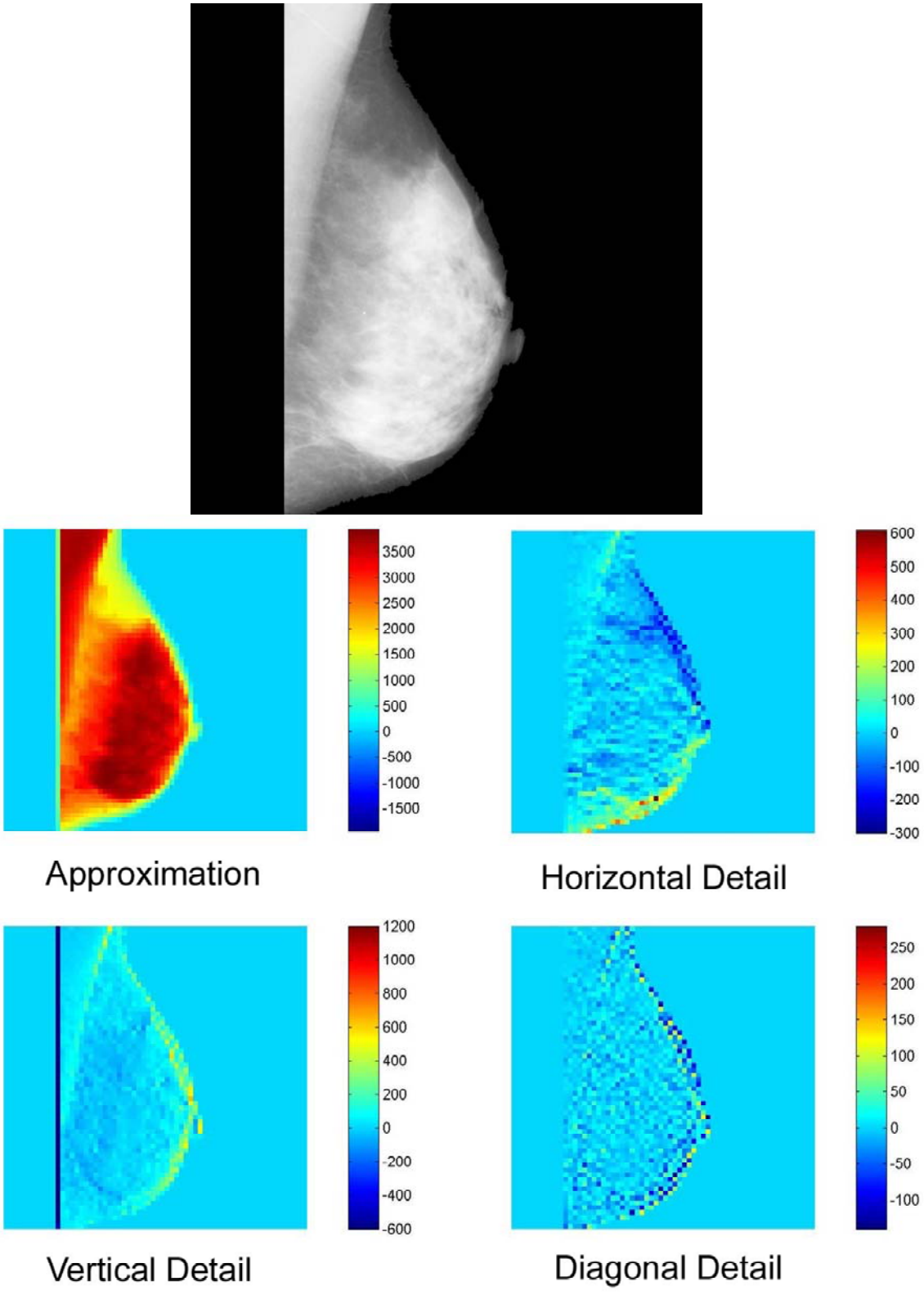


Figure 4.8 - Original mammography image (top) and 4 output views at the third level of decomposition using the Haar wavelet basis. Resolution is 128x128 pixels at this scale.

The use of the wavelet transform offers several advantages over other transforms, such as the Fourier transform, that could have been used to analyze the mammography images. The primary advantage is the property of multiresolution: the wavelet maps at different levels emphasize features of different sizes. Though masses may range in diameter from a few millimetres to a few centimetres, the same algorithm may be sensitive to them by analyzing wavelet maps at different scales. Microcalcifications tend to be quite small and are fairly similar in size, so multiresolution is less useful for detecting these abnormalities.

The second advantage of the wavelet transform is its retention of spatial location information. The produced maps show the spatial distribution of information at particular size scales; in contrast, the Fourier transform would lose the spatial information and simply produce a map of the relative contributions of different frequencies over the entire image. This spatial dependence is useful for finding localized structures, such as microcalcifications and small masses, which remain localized after the wavelet transform has been applied and can then be distinguished from a more homogeneous background.

4.5 Generation of scalar features

Once the wavelet maps had been generated from the pre-processed images, scalar features were extracted to be used in the classification process. All features extracted were whole-image statistical features, and the resulting classification scheme classified whole images as being normal or suspicious, rather than locating suspicious

regions within an image. Four features were extracted from each wavelet map: the mean intensity, the standard deviation of the pixel intensities, the skewness of the pixel intensities and the kurtosis of the pixel intensities, each of which are discussed below. Several corrections were applied to the wavelet maps before calculating these features to emphasize the effects of abnormalities on their values; these corrections are discussed in Section 4.5.1. Following that, Sections 4.5.2 to 4.5.5 discuss how each statistic is calculated and their physical interpretation, which governs their utility in distinguishing between normal and suspicious images.

4.5.1 Corrections for breast size and directionality of wavelets

To make the scalar features extracted from the wavelet maps as physically meaningful as possible, several corrections were applied to the raw statistical measures taken from the images. Specifically, the absolute values of the pixels in the wavelet maps were used rather than their signed values to focus only on the strength of the correlation between the wavelet basis function and the underlying image at the current scale. Secondly, a correction was made for the size of the breast tissue region within the image to keep it from skewing the feature values, which should ideally only vary between images because of differences in pathology and structure in the tissue.

A wavelet map contains both positive and negative values, corresponding to positive and negative correlations between the wavelet and the image. For example, the Haar wavelet transform produces positive values for a rising edge and negative values for a falling edge as it moves left to right across an image. However, the directionality of the structures in the images, such as microcalcifications, was not as important as their

presence or absence, and so only the magnitudes of the wavelet maps were used. As well, the mixture of positive and negative values can partly compensate for each other in the calculation of mean intensity and other features; this reduces the difference between the values measured from different images, making it more difficult to classify images based on this feature Figure 4.9 shows the difference between a raw wavelet map and the absolute value of a wavelet map for the level 3 horizontal detail view of the Haar wavelet basis.

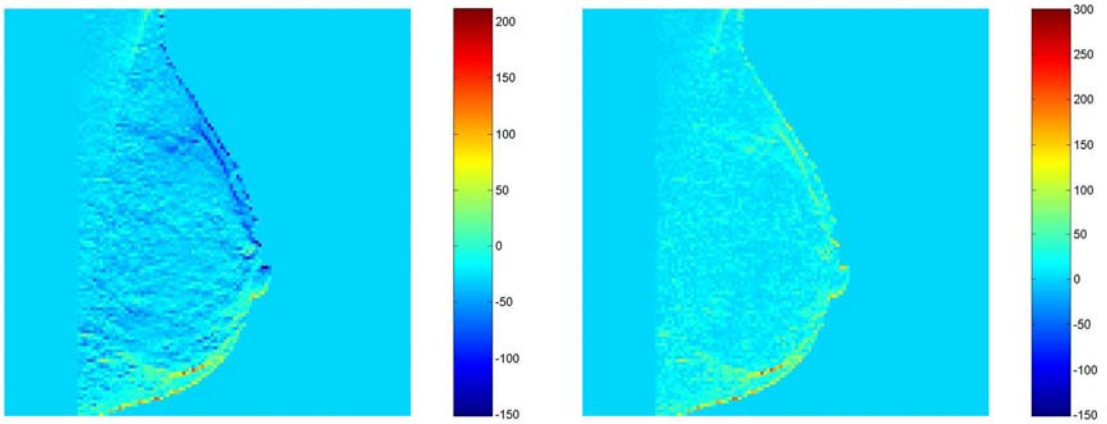


Figure 4.9 - Wavelet map (left) and absolute value of wavelet map (right) for level 3 horizontal detail view of Haar wavelet decomposition

To reduce the effect of breast size alone on the resulting scalar features, all statistical values were measured only for pixels within the region corresponding to the actual breast tissue. This was done by using the approximation image at each level as a mask: a given pixel in an image was only used in a calculation if the corresponding pixel in the approximation image at the same level was non-zero, meaning that it corresponded to tissue. Without this correction, the statistical features would have been skewed to higher values for larger breasts, obscuring the more meaningful differences in values caused by the presence or absence of abnormalities. For example, the

measured mean intensity of an entire image would be twice as large if the tissue covered twice as much of the image even if the mean intensities within the tissue regions themselves were the same.

4.5.2 Mean intensity

The first scalar feature used in this research was the mean intensity of the wavelet map. The mean intensity μ was calculated only for pixels within the tissue area and was calculated as follows:

$$\mu = \frac{1}{N}\sum_{i,j} I(i, j), \tag{4.2}$$

where N is the number of pixels in the tissue region of the image, $I(i,j)$ is the pixel intensity of the pixel in row i, column j of the image and the summation runs over all pixels in the region defined by the image mask.

The mean intensity feature value gives a measure of the fraction of the total information in the original image present at the current scale. The horizontal, vertical and diagonal detail images' mean intensities show the high frequency information at the current scale, while the approximation image shows the information left in all larger scales. Lower levels of the decomposition corresponding to larger scales will then parse the approximation image and show how its information is in turn distributed among all larger scales.

The presence of microcalcifications should skew the total energy of an image towards the higher resolution, lower spatial scale maps, since the deposits are small in size and bright in appearance. In tissue containing calcifications, then, the high resolution maps should have a slightly higher intensity and the low resolution maps

should have a slightly lower intensity than the corresponding maps produced from a normal sample. In reality, the effect of the small microcalcifications on the total image intensity is small and easily obscured by other, normal variations between different samples; this drawback does not make this statistic invalid, but it motivates the use of multiple measures from several images in tandem to fully differentiate between normal and suspicious tissue.

The other sign of pathology is the presence of a mass in the image. A mass may have almost any size, from a few millimetres to several centimetres in width. As well, a mass may have a sharp boundary, or it may have a speculated appearance with tendrils extending into surrounding tissue, especially for the case of malignant, cancerous masses. This large variety in the appearance of masses in an image means that no single scale or wavelet basis will naturally extract all masses from the background tissue. All masses do share the property of being localized in one region of the tissue, though, and they typically appear as slightly brighter regions due to their slightly higher density as compared to healthy tissue. Hence, a particular wavelet basis may measure a slightly larger than normal mean intensity at a particular scale and view when a mass is present, and this may make it possible to identify images showing the presence of masses automatically, especially when the feature values from several different scales are used in conjunction to detect masses of differing sizes.

4.5.3 Standard deviation of pixel intensities

The second scalar feature generated from the wavelet maps was the standard deviation of the pixel intensities in the wavelet maps. The standard deviation σ is calculated over all N pixels within the tissue region as follows:

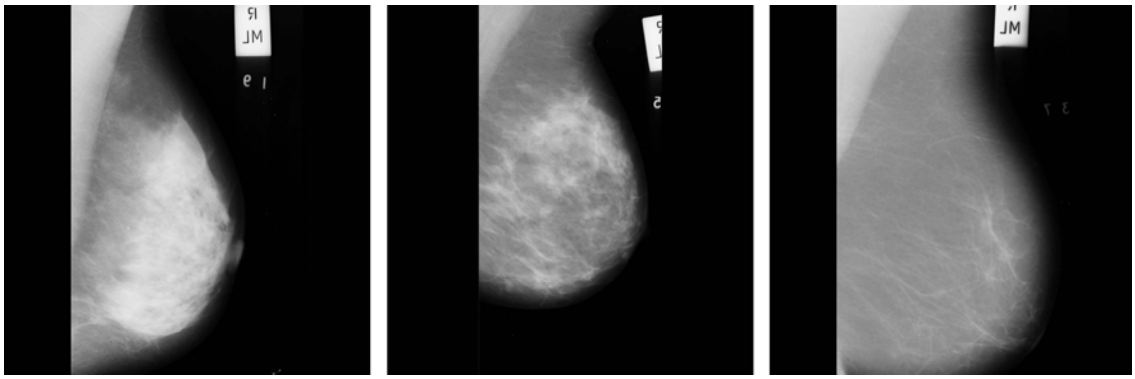
$$\sigma = \sqrt{\frac{1}{N}\sum_{i,j}\left[I(i,j) - \mu\right]^2} \ . \tag{4.3}$$

The standard deviation measures the variability in the brightness of the image over the tissue region. Microcalcifications much brighter than the mean image intensity will raise the standard deviation measure of the high spatial resolution levels of the wavelet map images as compared with the corresponding images from a sample with no visible microcalcifications. Masses may affect the standard deviation of the image in a similar fashion, but at lower spatial resolution scales corresponding to the approximate size of the mass.

The major drawback of using the standard deviation as a measure for detecting abnormal pathologies is the variation in tissue appearance among healthy patients. One contributor to this variation is the amount of stromal and glandular tissue present in different patients.

Breasts can be generally classified by the fraction of glandular tissue to stromal tissue into several categories. Glandular tissue is composed of the ductile tissue and milk-producing lobules of the mammary glands and appears in a mammography image as a set of relatively bright lines that radiate back from the nipple. Stromal tissue does not absorb x-rays as efficiently at the energies used for breast imaging and hence appears relatively dark; it is composed of connective tissue and fatty adipose tissue. Since the amount of glandular tissue is relatively constant between individuals, the

difference between patients' breast sizes is controlled mainly by the amount of stromal tissue present. The MIAS divides its images into three categories according to the relative amounts of stromal and glandular tissue present: dense breasts contain primarily glandular tissue, fatty breasts contain a relatively large amount of stromal tissue, and glandular breasts are somewhere in between. The boundaries of the classifications are arbitrary, as the fractions of tissue types present falls along a continuum, but this classification highlights the large inherent variability in breast images that makes accurate interpretation so challenging. Figure 4.10 shows a sample image from each of the three categories: note how the bright glandular tissue is progressively more diffuse from the dense to the glandular to the fatty images.



(a) **(b)** **(c)**

Figure 4.10 - Comparison of dense (a), glandular (b) and fatty (c) breast images, all showing normal tissue

The variation in tissue composition between the three images in Figure 4.9 shows how strongly the standard deviation is affected by normal biological variation. The relatively uniform appearance of the fatty sample in (c) means that the standard deviation of the image intensity is low, while the split between bright and dark tissue regions in (a) and (b) means that they will have a much larger standard deviation. The

standard deviation can still act to differentiate between normal and abnormal images, but only with careful choices of scale and wavelet bases that minimize the effects of normal biological variation on the feature values. Section 4.7 discusses the problem of feature selection, where the optimal scales and views are determined for a given wavelet basis.

4.5.4 Skewness of pixel intensities

The third statistic measured from each wavelet map image is the skewness of the pixel intensities. The skewness of a distribution of values is defined as the third central moment of the distribution, normalized by the cube of the standard deviation. Formally, the skewness S is calculated according to:

$$S = \frac{1}{N} \sum_{i,j} \left[\frac{I(i, j) - \mu}{\sigma} \right]^3 . \tag{4.4}$$

The skewness of a distribution measures the degree of asymmetry. Consider, for example, a distribution which is approximately Gaussian. If the left tail is slightly larger than the right tail, the distribution has a negative skewness; if the right tail is slightly larger, the distribution has a positive skewness. This statistic is sensitive to the addition of a small number of unusually small or large values to a distribution, which may alter the skewness even if the mean value or standard deviation is not significantly altered.

Because of its sensitivity to a small number of additional large-valued points, skewness is a good candidate for detecting the presence of microcalcifications in an image. The calcifications, which are only a few pixels in size at even the highest resolutions but are unusually bright, can skew the distribution of pixel intensities in the wavelet map of an image in a measurable way.

The skewness measurement is also sensitive to the presence of masses in an image. Dense masses appear as slightly brighter than normal regions within the tissue, raising the number of pixels with intensities larger than the mean value of the distribution. Since skewness measures the imbalance between the parts of the distribution above and below the mean, the presence of a dense mass will raise the skewness relative to a healthy image. Wavelet bases and levels that correlate particularly well with the shape of a mass will show a larger effect in their skewness measure, making this approach most useful when used in conjunction with an appropriate basis.

4.5.5 Kurtosis of pixel intensities

The fourth and final statistic measured from the wavelet maps is the kurtosis of the pixel intensities. The kurtosis of a distribution of values is defined as the fourth central moment of the distribution, normalized by the fourth power of the standard deviation of the distribution. Formally, the kurtosis K is calculated according to:

$$K = \frac{1}{N} \sum_{i,j} \left[\frac{I(i, j) - \mu}{\sigma} \right]^4 . \tag{4.5}$$

Qualitatively, kurtosis measures the narrowness of the central peak of a distribution compared with the size of the distribution's tails. A distribution with a narrow peak and tails that drop off slowly has a large kurtosis compared with a distribution with a relatively wide peak but suppressed tails. Kurtosis is positive definite for a real-valued distribution of values. The kurtosis and standard deviation of a distribution may be similar, though kurtosis is more sensitive to points distant from the mean than the standard deviation is.

Because kurtosis depends on the fourth power of the distance between an outlying point and the mean, it is highly sensitive to the addition or loss of points far from the mean of the distribution. Microcalcifications may raise the kurtosis by increasing the number of unusually bright pixels in a wavelet map, especially at the higher resolution scales where the calcifications can be differentiated.

The kurtosis measure appeared to be sensitive to the presence of masses as well. One possible mechanism for this is that masses appear slightly brighter than normal stromal tissue and introduce additional structure into the wavelet maps at several scales. This may reduce the number of unusually low intensity points in a wavelet map by adding intensity into normally dark regions; this shift would lower the kurtosis and may explain its effectiveness at detecting the presence of masses in an image. Because of normal biological variation, however, the kurtosis of a single scale and view is not sufficient to classify all images, and a combination of features will be necessary to achieve an acceptable detection rate.

4.5.6 Sample distributions of each feature type

From each wavelet map in the decomposition, four statistical features were measured: mean intensity, standard deviation of intensity, skewness of intensity and kurtosis of intensity. Figure 4.11 shows the distributions of these four features for one particular wavelet map, the level 7 horizontal detail map using the Haar wavelet basis. The feature values are grouped into 8 bins in the histograms: each bin represents the fraction of the input image set that produced feature values in that range. Note that the normal and suspicious distributions are very similar for the mean intensity and standard

deviation features, making it difficult to differentiate between them in a classifier, while the skewness and kurtosis features show better differentiation: the peaks of the distributions are offset slightly and the probabilities in many of the bins are significantly different between the normal and suspicious distributions. Because there is such great overlap between the normal and suspicious distributions for a given feature, multiple features are needed in the classifier to effectively distinguish between the two classes.

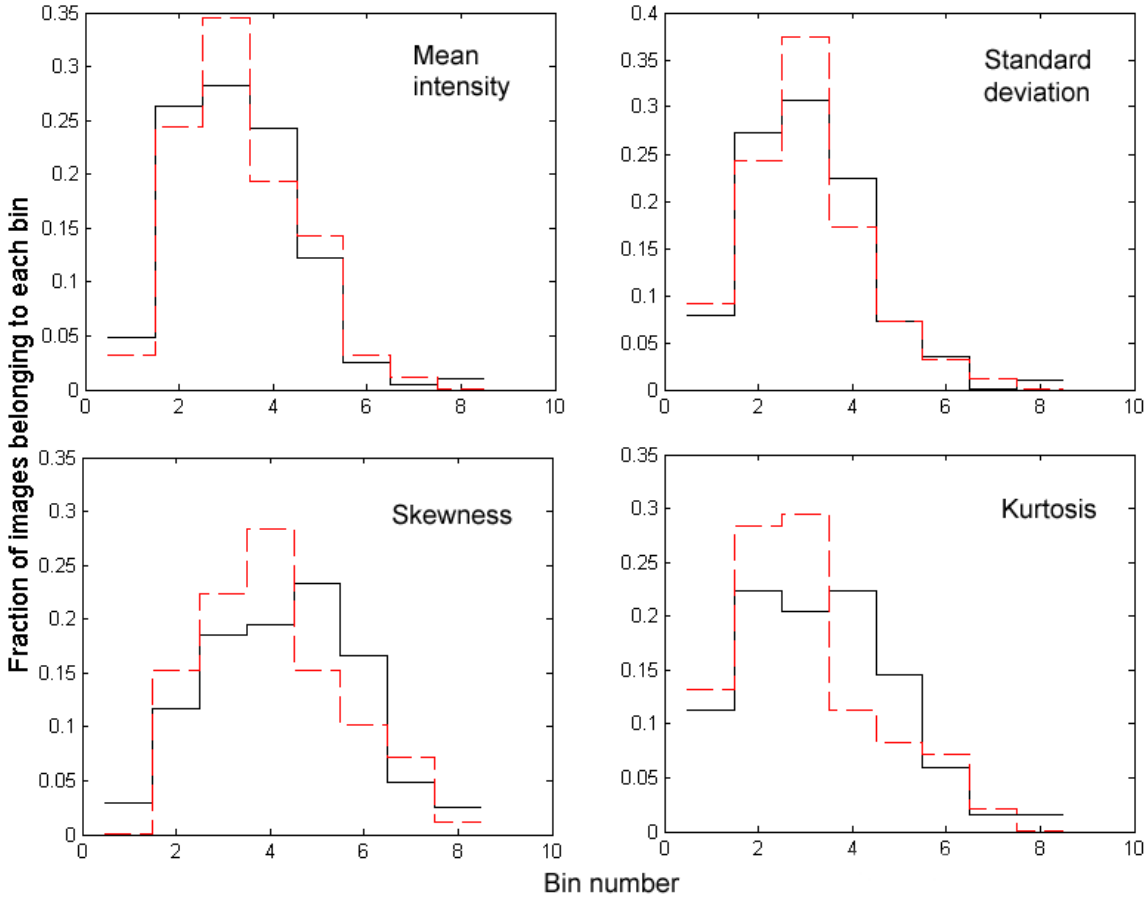


Figure 4.11 – Probability distributions for the four features measured from level 7 horizontal detail map using Haar wavelet basis for normal (solid black) and suspicious (dotted red) images

For reference, the bin sizes and ranges of possible values for the four types of features shown in Figure 4.11 are listed in Table 4.1. The kurtosis values had the largest range due to the fourth power of the pixel intensity appearing in the numerator of

equation 4.5. The skewness values were all positive, meaning that the pixel intensities in every images' level 7 horizontal detail wavelet map was skewed towards higher intensities than a Gaussian distribution of intensities would have been.

Table 4.1 – Feature value ranges and bin sizes for level 7 horizontal detail map using Haar wavelet basis

Feature type	Minimum Value	Maximum Value	Bin size
Mean intensity	1.88	7.66	0.72
Standard deviation	2.92	10.51	0.96
Skewness	0.32	6.58	0.78
Kurtosis	1.59	20.23	2.33

4.6 Single naïve Bayesian classifier

Naïve Bayesian classifiers were used in concert to form the full classification system; each individual classifier was constructed and trained as an individual classifier before being combined into the complete system.

The construction of a single Bayesian classifier was relatively straightforward. The input feature values from the wavelet analysis were discretized into a small number of bins to create probability distributions for each feature's value. The binned probabilities were then employed to determine the probabilities that a new sample was either normal or suspicious and classification was made based on which class carried the larger probability of producing the sample. Several techniques used to streamline the algorithm are discussed at the point where they occur in the algorithm, since high efficiency was beneficial for the feature selection step to be able to rapidly explore a sufficient number of feature subsets.

4.6.1 Discretization of scalar feature values to form probability distributions

Each scalar feature measured from the wavelet maps of the normalized mammographic images varied over a continuous range of possible values. Since the leave-one-out training methodology was applied, there were approximately 95 suspicious and 200 normal sample values available for each feature to construct a probability distribution when the MIAS image set was used [51].

Many approaches to binning continuous data exist [54], though the differences between their accuracies are relatively small. Because of this, the simplest approach was used in this work: the data was binned into a pre-selected number of bins of equal width. To be able to compare the probability that a new sample's feature value came from the normal or the suspicious class, the normal and suspicious probability distributions for a feature needed to have the same bin widths and locations. The number of bins was estimated using Sturges' rule [50], which estimates the optimal number of bins for representing a distribution given a set of sample points. Sturges' rule estimates that the number of bins should be $1 + \log_2 N$, where N is the number of points available for binning; this gives 7 bins for the suspicious distribution and 8 bins for the normal distribution. To make the two distributions comparable, a common number of bins was chosen for both. The value of 7 bins given by Sturges' rule was used as a starting point, though choices from 3 to 28 bins were tested to determine their effect on the final classification rate; ultimately, 8 bins were used to discretize both distributions. Section 5.2.1 shows the effect of varying the number of bins on the classification rate for a fixed set of features.

Once the number of bins was selected, the locations of the binning boundaries was determined based on the available samples. The lowest bin's lower bound was set equal to the smallest sample feature value, and the highest bin's upper bound was set equal to the largest sample feature value; the remaining bin boundaries were equally spaced between these two extremes. Note that the upper and lower bounds were set by the largest and smallest points in the combined normal and suspicious training samples, since both distributions need to use a common set of bin locations.

For each sample whose feature value fell between the bounds of a particular bin, a count was added to that bin. Figure 4.12 shows the binning results for the normal and suspicious samples for the kurtosis feature of the level 3 horizontal detail wavelet map using the Haar wavelet.

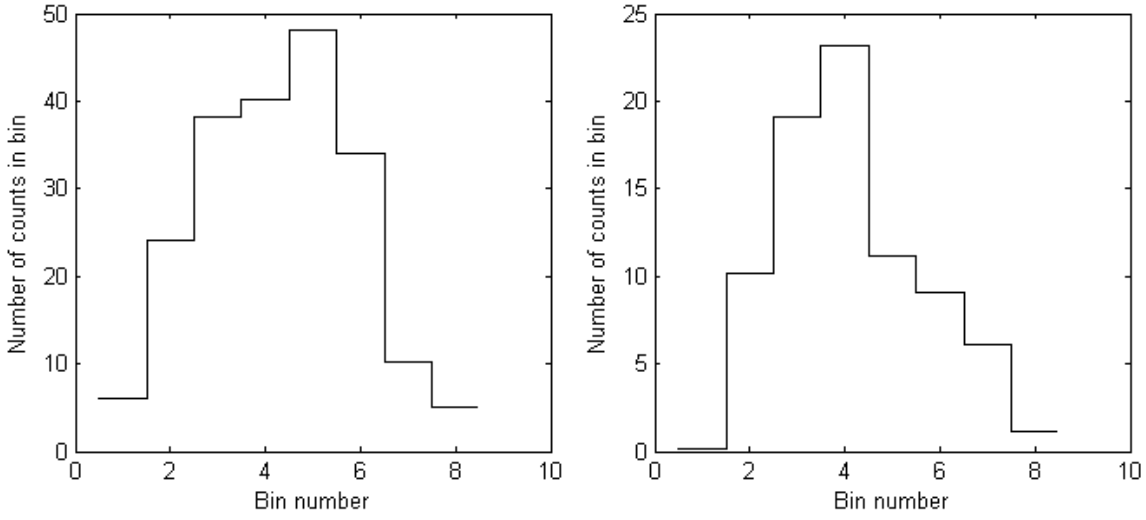


Figure 4.12 - Normal (left) and suspicious (right) bin counts for skewness feature of the level 3 horizontal detail map using the Haar wavelet

4.6.2 Correction for empty bins

Since the relative probabilities of the two distributions are used to classify each sample, the presence of an empty bin, corresponding to a probability of zero, biases a classifier: if an image has a feature value falling in the range of an empty normal bin, for example, the probability of the image being from the normal class is automatically zero, regardless of the relative probabilities of the other features used in the classifier. To mitigate this bias, a small correction factor was applied to all bins, including the empty ones, to account for the uncertainty inherent in estimating a probability distribution from a finite sample of data points.

The uncertainty in a given bin's count was taken to be the inverse of the square root of the number of data points used to estimate that distribution; the correction factor was 0.113 counts for the suspicious distribution and 0.070 counts for the normal distribution. This factor was chosen in analogy to the uncertainty in counting experiments for random processes like radioactive decay, where the relative uncertainty scales with the inverse square root of the number of observed counts. The factor was added to all bins, whether or not they were empty, to avoid introducing additional bias; if the corrective factor is taken as the uncertainty in the number of counts, then the corrected bin values correspond to the maximum possible number of counts in each bin, rather than the most probable number of counts in each bin. This correction reduced the effect of empty bins, but still kept their counts small, especially when a large number of samples were available; this agrees with the notion that an empty bin does have the significance of carrying a low probability, and that an empty bin's significance increases as the number of samples used to build the distribution increases.

Once the bins were filled and the correction factor was applied, the bin counts were normalized to convert the count rates into a probability density. Each bin was normalized according to the following equation:

$$P_c(j) = \frac{1}{Wk} \frac{N_c(j)}{\sum_i N_c(i)} \tag{4.6}$$

where $P_c(j)$ is the probability of class c producing a feature value within bin j, W is the width of a single bin, k is the number of bins used, $N_c(j)$ is the number of training samples from class c falling into bin j, after the correction factor is applied, and the summation runs over all bins in the distribution. In this way, the integral of the probability density over the range of possible values for a particular feature becomes equal to one. Figure 4.13 shows the probability distribution obtained by normalizing the bin counts shown in Figure 4.12 and correcting for empty bins.

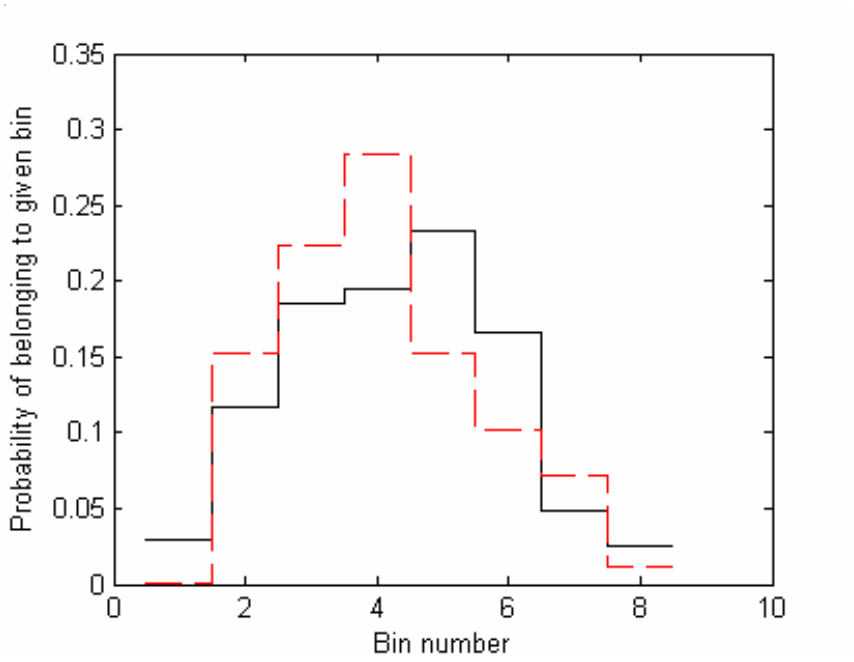


Figure 4.13 – Normal (solid black) and suspicious (dashed red) binned probabilities for skewness feature of the level 7 horizontal detail map using the Haar wavelet

4.6.3 Classification of whole image based on feature probabilities

The discrete probability densities for each feature, once generated, were used to classify an input image as either normal or suspicious. Equation 3.1 was used to calculate the probabilities of class membership; the prior probabilities *P(c)* were set equal for both classes, though other values were tested in Section 5.2.3 The probability that an image was normal or suspicious was the product of the probabilities that each feature used from the image was normal or suspicious; three to five features were used at a time from an image for classification and were selected according to the process in Section 4.7. The ratio of the suspicious to the normal probability was taken and used to classify the image: if the ratio was greater than one, the image was classified as suspicious; if the ratio was less than one, the image was classified as normal.

To improve efficiency, the implementation of the probability comparison was performed slightly differently. The ratio of the suspicious to the normal probabilities for each individual feature was calculated first before the product over all features used was performed, rather than multiplying the different feature probabilities together first and then calculating the suspicious to normal probability ratio. Further, the logarithm of the probabilities was used in practice, so that a product of probabilities became a sum of the logarithm of the probabilities, and the logarithm of the ratio of the probabilities was used for classification instead: an image was suspicious if the logarithm of the suspicious to normal probability ratio was greater than zero or normal if it was less than zero. The use of logarithms and pre-calculated ratios eliminated the repetitive multiplication and division operations from the classification step, speeding the

algorithm by an order of magnitude and allowing a larger family of possible feature subsets to be explored, as discussed in Section 4.7.

4.7 Feature selection and reduction

Since a large number of potential classification features were generated from each image, and since a classifier should use only approximately one feature for every ten training samples available, a selection process was needed to choose those features that are most effective at differentiating between normal and suspicious images. Specifically, their were four parameters measured from each wavelet map; as there were four wavelet maps per level and eight levels of decomposition, 132 potential features were created for each of the 11 wavelet bases tried in this work. Since multiple classifiers would be used in tandem to perform the final classification, each individual classifier was limited to use no more than three features, though one or two features could be used if it produced better results. Selecting such a small subset of the candidate features added flexibility to the design of each individual classifier: for example, one classifier could use a feature subset sensitive to microcalcifications while another could use a feature subset sensitive to masses.

Feature selection was carried out through a semi-exhaustive process, since there were too many potential features to test the performance of every possible triplet of features. Each classifier was limited to use only one wavelet basis and two of the four types of parameters generated for the maps. The 64 features this left were then searched exhaustively: every possible triplet, doublet and singlet of features was tried on the

available data, and a performance metric was developed to select the most effective combination.

The performance metric used to select the most effective feature subset was a weighted sum of the number of true positive classifications, *NTP*, and the number of true negative classifications, *NTN*. A true positive was an image with abnormalities that was correctly classified as suspicious, and a true negative was an image with no abnormalities that was correctly classified as normal. The score *S* produced from this was calculated according to:

$$S = w(NTP) + (1 - w)(NTN), \tag{4.7}$$

where *w* is the weighting factor and varies between zero and one. A high weighting factor places more importance on correctly classifying suspicious images, a low weighting factor places more importance on correctly classifying normal images, and a weighting factor of 0.5 makes the score depend only on the number of images classified correctly, regardless of type. Another interpretation of the weighting factor is that it measures the importance of sensitivity relative to specificity; a large weighting factor favours a more sensitive classifier while a small weighting factor favours a more specific classifier.

The nature of the classification done in this work, that is, that images classified as normal are not subject to further analysis, means that any false negatives cannot be corrected for later and will correspond to a missed abnormality; therefore, the true positive fraction must be maximized. To that end, the weighting factor was chosen to be 0.995; this made the true positive fraction paramount, but in the event that two feature

subsets produced the same number of true positives, the tie would be broken by selecting the feature subset which had the higher true negative fraction.

Since the individual classifiers were combined after they were tested, individual classifiers could also be designed to maximize the number of masses detected, or the number of microcalcifications detected, or any other specific type of abnormality. By looking for a specific type of abnormality, the appearance of images within that group should be more uniform than across all types of abnormalities, and a classifier may be better able to distinguish those images from all others. To search for a particular type of abnormality, *NTP* in equation (4.7) was replaced with the number of correctly classified images with the specified abnormality, and *NTN* was replaced with the number of correctly classified images of all other types, including normals. Again, this selected the most sensitive feature combination for the specific type of abnormality, breaking ties by selecting the most specific classifier with that sensitivity. Note that by searching for only one type of abnormality, the performance metric had little penalty for misclassifying other types of abnormalities as normal; to ensure that these other abnormalities were not missed by the complete system, the outputs of the individual classifiers had to be combined carefully.

4.8 Formation of concerted-effort set of classifiers

By combining the output from several classifiers, two advantages occur: firstly, the overall accuracy of the classifier system may increase significantly; and secondly, it becomes possible to provide confidence levels for the classifications made by the complete system. Section 4.8.1 discusses the individual classifier and how confidence

levels may be extracted from its outputs. Section 4.8.2 develops the general method for calculating the confidence levels for a concerted-effort set of classifiers based on testing with small data sets. Section 4.8.3 discusses the sequential classifier design and the determination of confidence levels from it. Section 4.8.4 discusses the vote-taking classifier scheme and its confidence measure as a possible alternative to the sequential method. Section 4.8.5 discusses the method used in the final system: it used a more complex set of connections between the individual classifiers and customized classifiers to search for different types of abnormalities at different stages of the process.

4.8.1 Confidence levels from a single classifier in a concerted-effort set

The chance that an image classified as normal by a given classifier is actually normal depends on the sensitivity and specificity of the classifier, as well as on the relative number of normal and abnormal images presented to the classifier. Specifically, the probability that a sample in a bin is normal is equal to the number of normal images in that bin divided by the total number of normal and suspicious images in the bin. Figure 4.14 shows the probabilities of the four possible types of outputs from a single classifier; $P_i(N)$ is the fraction of input images to the i^{th} classifier which are actually normal, $P_i(S)$ is the fraction of input images to the i^{th} classifier which are actually suspicious, TN_i is the true negative fraction for the classifier, TP_i is the true positive fraction for the classifier, FN_i is the false negative fraction for the classifier, and FP_i is the false positive fraction for the classifier. $P_{i+1}(a,B)$ are the probabilities that a sample of class B is classified as belonging to class a, where a and B can be s (or S) for

suspicious or *n* (or *N*) for normal. The outputs from one classifier can then be used as inputs to another classifier, making the equations in Figure 4.14 recursive.

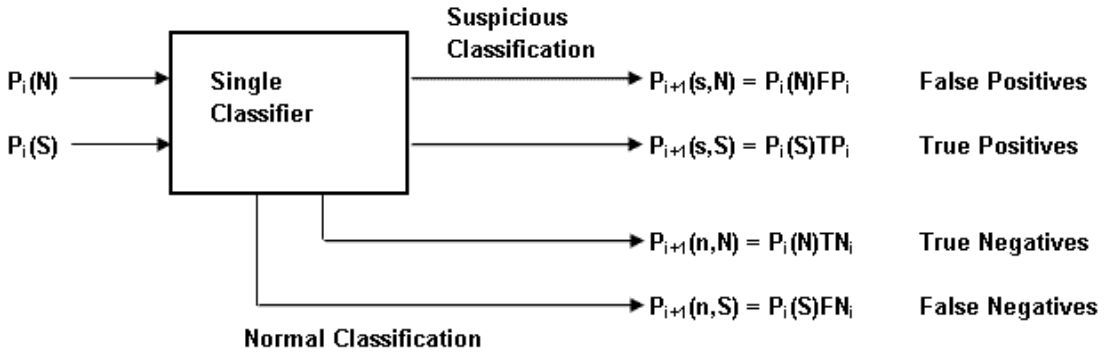


Figure 4.14 – Probabilities for 4 possible outputs from a single classifier

The four output probabilities from a single classifier together sum to one, so the probabilities among all samples classified as normal or as suspicious must be normalized to give a confidence level. The confidence that a sample in the normal bin is actually normal, $C_{i+1}(N)$ is the probability $P_{i+1}(n,N)$ divided by the total number of images classified as normal, $P_{i+1}(n,N) + P_{i+1}(n,S)$.

$$C_{i+1}(N) = \frac{P_{i+1}(n,N)}{P_{i+1}(n,N) + P_{i+1}(n,S)} \tag{4.8}$$

The confidence that a sample in the suspicious bin is actually suspicious, $C_{i+1}(S)$ is similarly:

$$C_{i+1}(S) = \frac{P_{i+1}(s,S)}{P_{i+1}(s,S) + P_{i+1}(s,N)}. \tag{4.9}$$

4.8.2 Classification confidence when classifiers share non-zero correlation

In practice, classifiers should share some correlation, since they are trained on the same data sets and are making the same class distinctions. The classification

confidence for a set of classifiers must then be determined experimentally, since it is not possible to develop a generalized expression for the correlation among more than two classifiers. The experimental confidence levels may be measured by testing the set of classifiers on a large number of images and counting the number of images of each class that are placed into each of the bins of the classifier network.

The predicted confidence levels for a realistic distribution of normal and suspicious images may be inferred from the results from a small data set. In the MIAS data set, for example, 98 of the 303 images are suspicious; this frequency of suspicious images is much higher than for the approximately 1 in 20 images that are suspicious in a typical clinic [14]. To correct for this discrepancy, the relative probabilities $P_i(N)$ and $P_i(S)$ in Figure 4.14 must be rescaled, or, equivalently in the experimentally measured case, the counts of images in each bin must be rescaled.

If the number of normal images counted in a normal bin experimentally is $\eta_{exp}(n,N)$, then the expected fraction of all images from a realistic distribution that are normal and are in the same bin, $F_{real}(n,N)$ can be calculated as:

$$F_{real}(n,N) = \eta_{\exp}(n,N)\frac{P_{real}(N)}{T_{\exp}(N)} \tag{4.10}$$

where $P_{real}(N)$ is the probability of an image from the realistic distribution being normal and $T_{exp}(N)$ is the total number of normal images used in the experimental data set.

Similarly, the realistic fraction of suspicious images in a normal bin, $F_{real}(n,S)$, can be found from the experimentally counted number of suspicious images in the bin, $\eta_{exp}(n,S)$, according to:

$$F_{real}(n,S) = \eta_{\exp}(n,S)\frac{P_{real}(S)}{T_{\exp}(S)} \tag{4.11}$$

where $P_{real}(S)$ is the probability of an image from the realistic distribution being suspicious and $T_{exp}(S)$ is the total number of suspicious images used in the experimental data set.

The predicted confidence level for an image from a realistic distribution to be correctly placed into a certain normal bin, $C_{real}(N)$, can then be calculated for each bin using the results measured from a small data set:

$$C_{real}(N) = \frac{F_{real}(n,N)}{F_{real}(n,N) + F_{real}(n,S)} \tag{4.12}$$

$$C_{real}(N) = \frac{1}{1 + \frac{\eta_{\exp}(n,S)}{\eta_{\exp}(n,N)} \frac{P_{real}(S)}{P_{real}(N)} \frac{T_{\exp}(N)}{T_{\exp}(S)}} = \frac{1}{1 + \frac{1}{\alpha} \frac{\eta_{\exp}(n,S)}{\eta_{\exp}(n,N)}}, \tag{4.13}$$

where α is a constant defined by:

$$\alpha = \frac{P_{real}(N)}{P_{real}(S)} \frac{T_{\exp}(S)}{T_{\exp}(N)}. \tag{4.14}$$

α characterises the frequency of normal and suspicious images in the experimental data set and in a realistic data set. For the MIAS data set with 98 suspicious and 205 normal images and for a clinic where 1 in 20 images are suspicious, α = 9.54. For the full DDSM data set [23] with 649 suspicious and 1065 normal images and for a clinic where 1 in 20 images are suspicious, α = 12.16.

By the same argument, the confidence level for an image from a realistic distribution to be correctly placed into a certain suspicious bin, $C_{real}(S)$, is calculated according to:

$$C_{real}(S) = \frac{F_{real}(s,S)}{F_{real}(s,S) + F_{real}(s,N)} \tag{4.15}$$

$$C_{real}(S) = \frac{1}{1 + \frac{\eta_{\exp}(s,N)}{\eta_{\exp}(s,S)} \frac{P_{real}(N)}{P_{real}(S)} \frac{T_{\exp}(S)}{T_{\exp}(N)}} = \frac{1}{1 + \alpha \frac{\eta_{\exp}(s,N)}{\eta_{\exp}(s,S)}} \tag{4.16}$$

Using equations 4.13 and 4.16, the confidence levels for any scheme of classifiers can be calculated directly by counting the number of normal and suspicious images assigned to each bin of the classifier network and using the value of α appropriate for the data set in question.

In practice, confidence levels from the case where an equal number of input images are normal and suspicious may be more useful than the realistic confidence levels. The difference between the two types of confidence levels is the relatively low number of suspicious images that occur in practice that dominates the realistic confidence levels and makes all bins have a large confidence for containing normal images. By applying equation 4.13 to both cases and comparing, the realistic confidence levels, $C_{real}(N)$ can be found from the case with an even number of normal and suspicious images, $C_{even}(N)$, by the following transformation:

$$C_{real}(N) = \frac{C_{even}(N)}{\frac{P_{real}(S)}{P_{real}(N)} + \left(1 - \frac{P_{real}(S)}{P_{real}(N)}\right) C_{even}(N)}. \tag{4.17}$$

The inverse transformation is given by:

$$C_{even}(N) = \frac{C_{real}(N)}{\frac{P_{real}(N)}{P_{real}(S)} - \left(\frac{P_{real}(N)}{P_{real}(S)} - 1\right) C_{real}(N)}. \tag{4.18}$$

The transformations for confidence levels in suspicious bins for a realistic distribution, $C_{real}(S)$, and an even distribution, $C_{even}(S)$, are given by:

$$C_{real}(S) = \frac{C_{even}(S)}{\frac{P_{real}(N)}{P_{real}(S)} + \left(1 - \frac{P_{real}(N)}{P_{real}(S)}\right) C_{even}(S)} \tag{4.19}$$

and

$$C_{even}(S) = \frac{C_{real}(S)}{\frac{P_{real}(S)}{P_{real}(N)} - \left(\frac{P_{real}(S)}{P_{real}(N)} - 1\right) C_{real}(S)}. \tag{4.20}$$

The mapping is monotonic, so the bins with the highest confidence levels are the same using either method. The choice of which confidence level measure to use depends on the situation: $C_{real}(N)$ gives a more exact measure of the realistic likelihood that an image in a bin is normal, while $C_{even}(N)$ may be more useful for comparing the relative confidence levels of different bins when deciding which images merit further analysis and which images are least likely to be suspicious.

4.8.3 Sequential series of individual classifiers

The first scheme devised for combining multiple classifiers was a sequential series, as shown in Figure 4.15. An image is presented to the first classifier: if it is classified as normal, it is removed from the system and placed in the first normal bin. If it is classified as suspicious, though, it is passed to the second classifier. If the second classifier finds the image to be normal, the image is removed from the system and placed in the second bin. The process is repeated until the images being binned as normal are no longer classified as normal with a sufficiently high confidence rate, and the final classifier's suspicious bin contains images which merit further study. The images removed as normal by each classifier carry a confidence level based on how many classifiers they were passed through before being found normal.

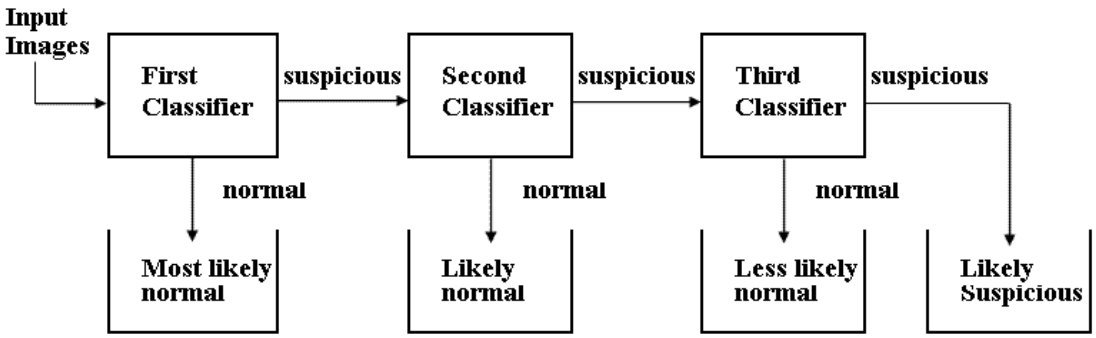


Figure 4.15 - Sequential series of classifiers and binned outputs

The confidence levels for each normal bin are calculated from the statistical behaviour of each individual classifier. A problem arises if two sequential classifiers have some correlation between their classifications. Images presented to the second classifier will be more likely to be reclassified in the same way that they were by the first classifier, altering the classification confidence levels for the second classifier and any others that follow. In the extreme case, applying the same classifier twice, the second classifier would find no images to be normal, since the suspicious images from the previous classifier would again be found suspicious. The next two subsections discuss two cases. The first case is the ideal, when there is no correlation between two sequential classifiers: the images presented to the second classifier will be classified at exactly the statistical rates predicted from running the second classifier on the whole data set. The second case generalizes to classifiers sharing some non-zero correlation in their classifications, and discusses how to determine the classification confidence for systems using such classifiers.

4.8.3.1 Confidence levels for sequence of uncorrelated classifiers

By applying the relations from Figure 4.14 recursively, the confidence levels for images classified as normal by the first, second, third classifiers and so on may be determined. Table 4.2 collects the confidence levels, $C_{real}(N)$ and $C_{real}(S)$, for the first 8 stages of a sequential set of fictitious, uncorrelated classifiers, each with a sensitivity of 90% and a specificity of 40% with 1 in 400 input images being abnormal. The table also shows the confidence levels, $C_{even}(N)$ and $C_{even}(S)$, for the case where the input images are equally likely to be normal or suspicious, to give a better sense of how the confidence levels decrease over successive stages of the sequential classifier design.

Table 4.2 – Confidence levels for uncorrelated sequential classifiers

Classifier	Normal bin outputs		Suspicious bin outputs	
	$C_{real}(N)$ (%)	$C_{even}(N)$ (%)	$C_{real}(S)$ (%)	$C_{even}(S)$ (%)
1	98.8	80.0	0.375	60.0
2	98.2	72.7	0.561	69.2
3	97.3	64.0	0.839	77.1
4	95.9	54.2	1.253	83.5
5	94.0	44.1	1.868	88.4
6	91.3	34.5	2.78	91.9
7	87.5	26.0	4.11	94.5
8	82.4	19.0	6.04	96.2

Note that the confidence levels of the normal bins decrease as more classifiers are used to classify an image as normal. This is reasonable, since an image placed in a later normal bin is actually classified as suspicious by all previous classifiers and as normal only by the last. Similarly, the suspiciousness of an image increases as more and more classifiers all classify the image as suspicious.

4.8.3.2 Confidence levels for sequence of classifiers with non-zero correlation

The sequential classifier design with M classifiers has M+1 bins: M normal bins and 1 suspicious bin. Thus, this concerted-effort set of classifiers is best suited to finding and removing normal images by removing all images assigned to the M normal bins, if their confidence levels are acceptably high. Images reaching the final classifier and being assigned to its suspicious bin are then subject to further analysis. Section 5.5.1 and Section 5.6.2 show the confidence levels predicted by training on testing on the MIAS and the DDSM image databases, respectively.

4.8.4 Vote-taking scheme for combining individual classifiers

An alternative to a sequence of classifiers is a vote-taking scheme. In the sequential case, an image is only classified as suspicious if every classifier classifies it as suspicious, limiting the sensitivity of the complete system. The vote-taking scheme considers the classification given by each of several individual classifiers and develops a confidence level based on the number of classifiers that agree in their classification. The output bins for the vote-taking scheme are the case where all M classifiers classify an image as normal, where M-1 classifiers classify the image as normal, where M-2 classifiers classify the image as normal, and so on.

In the ideal case, the expected number of images in a particular bin is calculated as the product of the probabilities that each classifier classified the image as it did. For example, if all classifiers except classifiers j and k classified an image as normal, the probability of this classification is $P_{\sim jk}(n,N)$ if the image is actually normal and $P_{\sim jk}(n,S)$ if the image is actually suspicious, where $P_{\sim jk}(n,N)$ and $P_{\sim jk}(n,S)$ are given by:

$$P_{\sim jk}(n,N) = \prod_{m=j,k} FP_m \prod_{i\neq j,k} TN_i \tag{4.20}$$

and

$$P_{\sim jk}(n,S) = \prod_{m=j,k} TP_m \prod_{i\neq j,k} FN_i \tag{4.21}$$

where TN_i and TP_i are the true negative and true positive fractions of the i^{th} classifier, and FN_i and FP_i are the false negative and false positive fractions of the i^{th} classifier. The confidence level for this image being normal, $C_{\sim k}(N)$, is then:

$$C_{\sim jk}(N) = \frac{P_{\sim jk}(n,N)P_{real}(N)}{P_{\sim jk}(n,N)P_{real}(N) + P_{\sim jk}(n,S)P_{real}(S)} \tag{4.22}$$

$$C_{\sim jk}(N) = \frac{1}{1 + \frac{P_{real}(S)}{P_{real}(N)} \prod_{m=j,k} \frac{TP_m}{FP_m} \prod_{i\neq j,k} \frac{FN_i}{TN_i}}, \tag{4.23}$$

where $P_{real}(N)$ and $P_{real}(S)$ are the probabilities of a given input image being normal or suspicious, respectively.

This equation can be used for any result from a set of classifiers: all classifiers that classified an image as suspicious are used in the TP_m/FP_m product, and all classifiers that classified an image as normal are used in the FN_i/TN_i product. Table 4.3 shows the confidence levels, $C_{real}(N)$, for 3 ideal classifiers: the classifiers have sensitivities of 85, 90 and 95 % and specificities of 45, 40 and 35 % respectively. The images are assumed to be realistically distributed, with 1 in 20 images being suspicious. A second set of confidence levels, $C_{even}(N)$, are also given for the case where half the input images are normal and half are suspicious; though this is not realistic, the confidence levels given in this case are distributed over a broader range and give a better sense of which bins are most likely to contain a large number of suspicious images.

Table 4.3 Performance of uncorrelated, three classifier vote-taking scheme

Classifier's classification (*N* or *S*)			$C_{real}(N)$	$C_{even}(N)$
1	**2**	**3**	**(%)**	**(%)**
N	*N*	*N*	99.97	99.4
S	*N*	*N*	99.85	97.1
N	*S*	*N*	99.81	96.3
N	*N*	*S*	99.39	89.1
S	*S*	*N*	99.12	84.9
S	*N*	*S*	97.25	63.9
N	*S*	*S*	96.48	57.8
S	*S*	*S*	85.52	22.8

Note that the confidence levels for each bin depend both on the number of classifiers that agree in their classification and on the relative sensitivities and specificities of those classifiers. These data also show the importance of sensitivity over specificity for this application: the third classifier is the most sensitive but the least specific, yet the overall confidence level depends most strongly on this classifier's output. For example, when the third classifier finds an image to be normal, the confidence level is always at least 84.9 %, even when the other two classifiers find the image suspicious

4.8.5 Network of classifiers customized to detect particular abnormalities

The most flexible, and potentially the most powerful approach is to link the outputs of different classifiers together in a more complex way than the linear sequential method of Section 4.8.3. Several classifiers can be used to remove images suspicious for the presence of calcifications; the images classified as normal by these classifiers should then contain mostly masses and normal images. Several classifiers can then be used to remove images suspicious for masses; images classified as normal by both

sections can then be removed as normal images. Figure 4.16 shows one potential scheme for organizing 5 individual classifiers into a network. The confidence levels for every output bin can be calculated using equations 4.13 and 4.16 as necessary. Section 5.5.3 and Section 5.6.4 discuss this approach in greater detail, since the exact structure of the network depends on the classifiers chosen and their relative correlations.

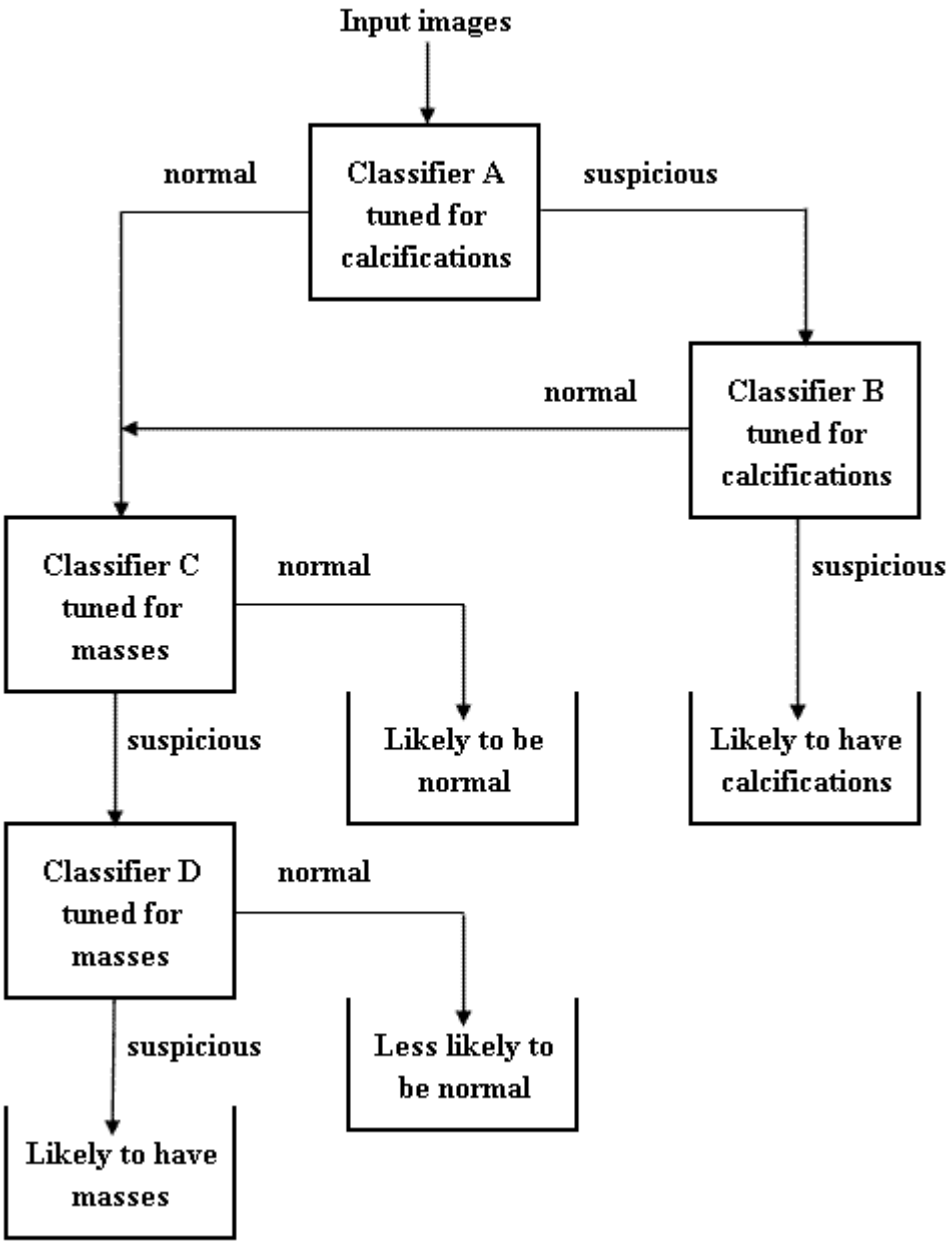


Figure 4.16 - Potential network design for concerted-effort set of classifiers

CHAPTER 5 – TESTING AND RESULTS

The complete image analysis and classification system discussed in Chapter 4 was built and tested in several stages. The image pre-processing was tested to ensure that the output images were free from artefacts and were regularized in appearance. The single Bayesian classifier was tested in great detail, due to the large number of tuneable parameters in the classifier and in the procedure for building the feature probability distributions. Once the individual classifier was operating as desired, testing was performed on the methods for combining the output from several classifiers. Finally, the complete system was tested on the MIAS database. To confirm that the classifier was not over-specified to the data set used in the testing process, the final classifier system was tested again on a set of images from the DDSM database.

5.1 Image Pre-processing testing

The images produced by the image pre-processing step were examined visually as they were produced to ensure the correct operation of this part of the system. The output image from the pre-processing step was displayed onscreen for every image processed. The displayed images were inspected to ensure no artefacts remained, to ensure that the image intensities were properly normalized and to ensure that the tissue

region was not altered by the artefact removal process. Two images had to be removed from the MIAS data set because of their poor quality and are shown in Figure 5.1: the images contained artefacts which overlapped the tissue regions and could not be removed without altering the information content therein; the images also showed significant artefacts due to poor scanning of the film plate into the digital image. The image on the left was a fatty normal image, while the image on the right was a fatty cancerous mass image. As the system being developed in this research is designed to be automated, it is expected that images provided to it will be more uniform in appearance, as the remaining 303 images in the MIAS database are, and so the omission of the two poor quality images was not considered a fault of the algorithm.

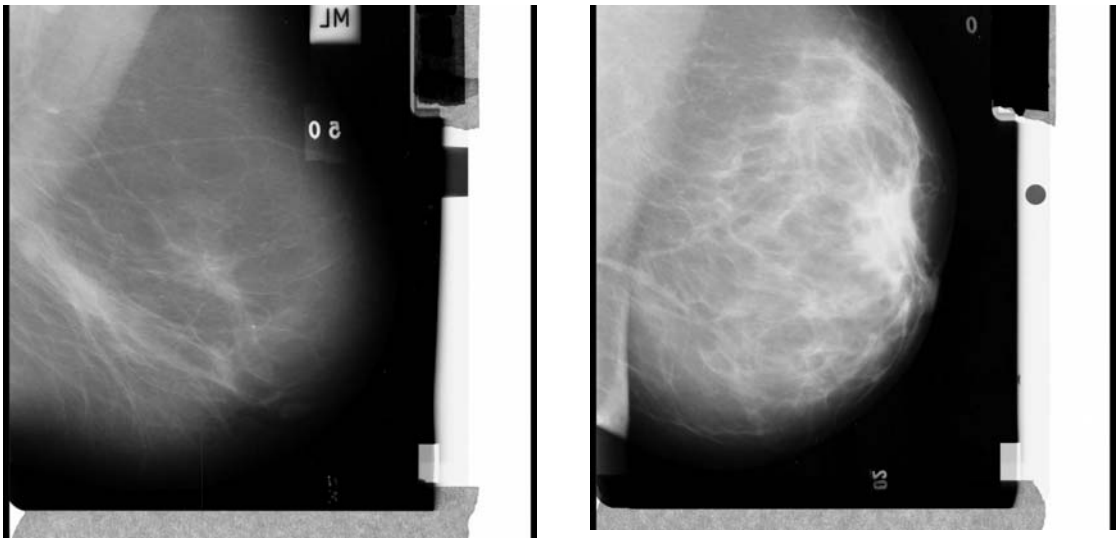


Figure 5.1 – Two poor images removed from the MIAS database before analysis

The remaining images were kept for analysis and are tabulated in Table 5.1. The images are listed by their tissue type – dense, glandular or fatty – and by their abnormality, if any – benign mass, cancerous mass, benign calcification, cancerous calcification or normal. Further, 19 of the images of masses showed architectural

distortions but were not given a designated tissue type within the database; they are included in their own column but are not differentiated by tissue type. Since the classification system does not use the provided information about the tissue type, this lack of information does not affect the performance of the system and the architectural distortion images could be used in the testing process.

Table 5.1 - MIAS database images by type

Abnormality (if any)	Dense	Glandular	Fatty	Architectural Distortion	Totals
Normal	75	62	68	---	205
Cancerous mass	3	7	8	10	28
Benign mass	11	13	14	9	47
Cancerous calcification	4	4	4	---	12
Benign calcification	4	5	2	---	11
Totals	97	91	96	19	303

5.2 Testing parameters for single Bayesian classifier

The naïve Bayesian classifier used to analyze the wavelet maps of the images had several tuneable parameters. Each parameter is discussed in turn below, along with test results used to validate the final parameter value used in the complete system.

The probability distributions used to select the most likely class to have generated a particular image had to be developed from the training data. Two approaches were used to build the distributions: the first approach was to bin the probability distribution and populate it by counting the number of training samples that mapped into each bin; the second approach was to approximate the distribution as a normal distribution and estimate its mean and standard deviation from the training data.

The results obtained while investigating the optimal number of bins to use for the data are given in Section 5.2.1. The selection between the binning approach and the normal distribution approximation approach is reported in Section 5.2.2.

The performance metric, used to select the optimal subset of features for the classifier, was tested for its effect on the final sensitivity and specificity of the complete system, as presented in Section 5.2.3. Finally, the *a priori* probabilities $P(c_j)$ of each class' relative frequency was adjusted, and evidence validating the choice to make this probability equal for both classes is shown in Section 5.2.4.

To simplify the listing of those features selected for a given classifier, a shorthand method is used. A feature is given the code *A-bN*, where *A* is the feature type (such as skewness), *b* is the wavelet view (such as horizontal detail or approximation), and *N* is the level of the decomposition (from 1 to 8). Table 5.2 lists the single letter codes for the feature types and wavelet views.

Table 5.2 – Shorthand for representing feature types

Feature type code	Corresponding Feature Type	Wavelet view code	Corresponding Wavelet view
M	Mean Intensity	h	Horizontal Detail
σ	Standard Deviation	v	Vertical Detail
S	Skewness	d	Diagonal Detail
K	Kurtosis	a	Approximation

Thus, for example, the code *S-h4* means that the feature is the skewness of the fourth level horizontal detail map of a given wavelet decomposition.

5.2.1 Number of bins for probability distributions

The final approach used to generate the probability distribution for each scalar feature was a binning process. The accuracy of approximating the probability distribution by a set of discrete bins depended on the size of the bins and the number of training samples: if there were too few bins, the approximation may not effectively represent the underlying distribution, while if there were too many bins, quantization effects would reduce the approximation's accuracy because of the small number of discrete counts in each bin.

Table 5.3 shows the sensitivity, specificity and overall correct classification rate for classifiers built using various numbers of bins. The overall classification rate is defined as the fraction of all input images that are correctly classified, regardless of whether they are suspicious or normal. The biorthogonal 3.7 wavelet basis was used, and the weight factor in the scoring metric (equation 4.7) was set to 0.5, making it selective for the highest possible overall correct classification rate, regardless of the classes of the misclassified images. This weight factor was used to measure the overall accuracy of a particular number of bins, since a weight factor that favoured sensitivity alone could emphasize statistically poor features where almost every sample is classified as suspicious. The most effective triplet of features was selected in each case from the pool of skewness and kurtosis features only.

Table 5.3 – Classification rate for different numbers of bins, biorthogonal 3.7 basis

Number of Bins	Best Feature Triplet			Sensitivity	Specificity	Overall Classification Rate
3	*K-v6*	*K-d6*	*K-d7*	15.2 %	98.1 %	75.0 %
5	*S-v5*	*S-v7*	*K-v1*	58.2 %	80.0 %	73.9 %
7	*S-v2*	*K-v1*	*K-d7*	67.1 %	78.1 %	75.0 %
8	*S-v1*	*S-d1*	*K-d2*	50.6 %	86.8 %	76.8 %
10	*S-h1*	*S-d1*	*K-v5*	39.2 %	92.7 %	77.8 %
12	*S-h3*	*S-h7*	*K-v1*	57.0 %	80.0 %	73.6 %
15	*S-d1*	*S-d7*	*K-v1*	69.6 %	76.1 %	74.3 %
20	*S-h3*	*S-a8*	*K-d7*	39.2 %	86.3 %	73.2 %

Several trends are visible from these data. The overall classification rate, the parameter being maximized by this choice of weight factor, is highest for 8 or 10 bins, though it does not vary greatly among all of the choices. This shows that there is some flexibility in the choice of the number of bins without greatly altering the effectiveness of the final classifier design. The choice of 8 bins was made for its agreement with Sturges' rule's prediction of 8 bins for the population size of the MIAS database of images. The choice of 8 bins also minimized the number of empty bins: although a correction was made for this as outlined in Section 4.6.2, populated bins are more representative of the training samples and should offer more predictive power for classification.

Another trend in this data is the large number of high resolution wavelet map views used in the selected classifiers. Except for those using the largest and smallest numbers of bins, each classifier used a feature taken from one of the level 1 views of the wavelet maps. Almost all of the classifiers also use a low-resolution view, with level 7

views being most common. This use of features from different resolution levels supports the usefulness of using wavelet analysis to parse the original images into multiple scale sizes for better classification.

Figure 5.2 shows the shape of the normal and suspicious probability distributions for the choices of bin numbers in Table 5.3 for the *S-h7* feature. Although the normal and suspicious distributions are very similar, the peaks of their distributions are slightly offset, allowing a Bayesian classifier to select between them. The difference between the two distributions becomes more apparent for larger numbers of bins, as the finer structure of the two distributions becomes more visible. The use of larger numbers of bins, however, creates more bins with no counts; these bins are ineffective for classifying new images. Larger numbers of bins also accentuates statistical variation: since there are more bins to populate using the same number of training samples, each bin will have a lower count of images and, as such, will have a larger uncertainty associated with it. The distribution using 20 bins shows this clearly, as the suspicious distribution has several bins with far higher or lower values than their immediate neighbours: this variation is more likely due to poor statistics in each bin than to an underlying distribution with such a complex shape.

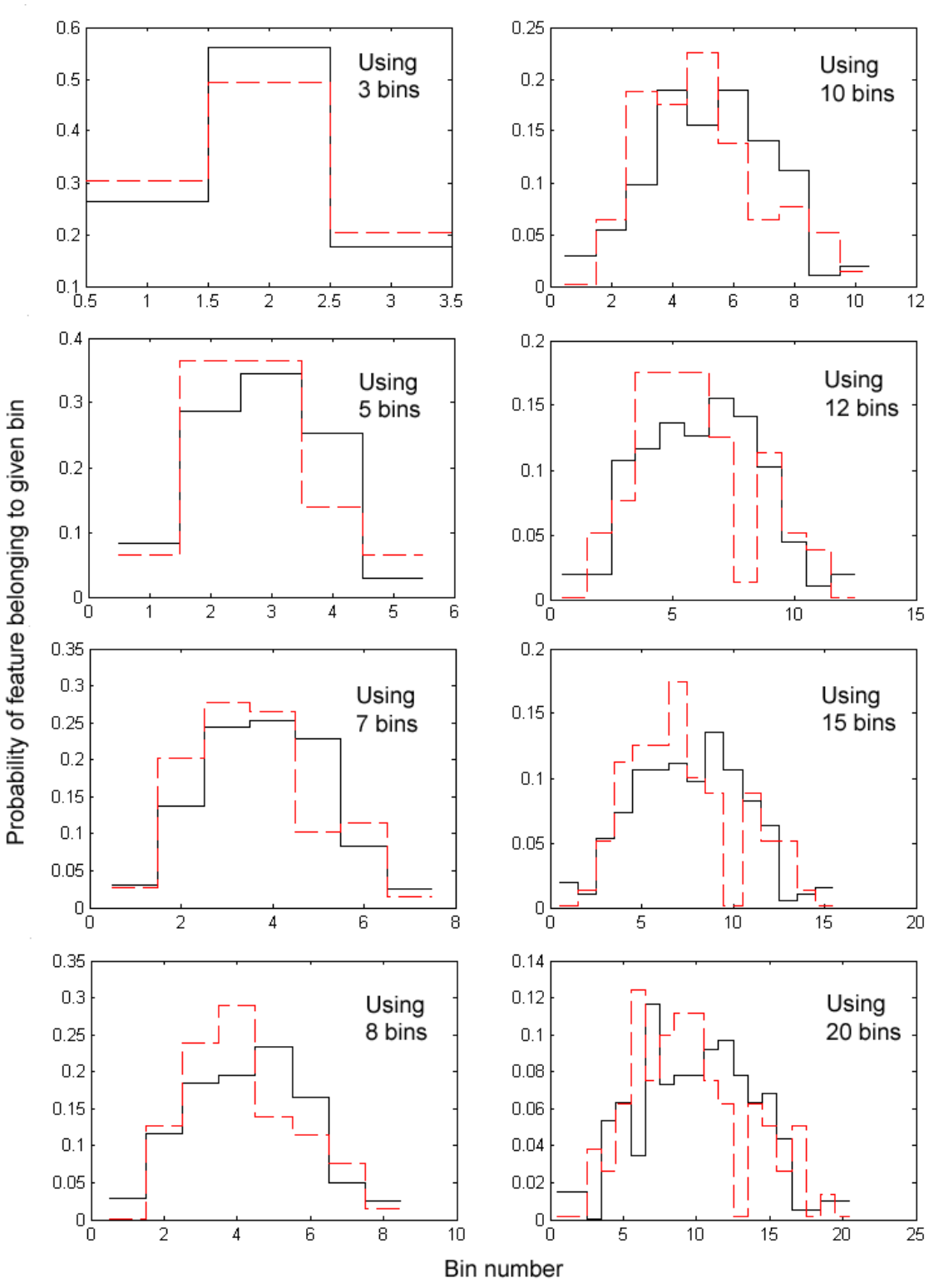


Figure 5.2 – Normal (solid black) and suspicious (dotted red) distributions for *S-h7* feature, Haar wavelet basis for different numbers of bins

5.2.2 Weight factor in performance metric

The feature reduction step used a scoring metric to select the subset of features that produced the highest score according to equation 4.7. The weight factor *w* made it possible to select feature subsets that were more sensitive or more specific, depending on the goals of the classifier. Table 5.4 shows the sensitivity and specificity of the classifiers selected for a range of possible choices of *w*. Three features were used in the feature subset and were taken from the skewness and kurtosis features measured from a biorthogonal 3.7 decomposition. Figure 5.3 shows the trends of Table 5.4 graphically.

The most noticeable trend as the weight factor varies is the compromise between sensitivity and specificity in the classifier using the selected feature subset. The sensitivity increases as *w* increases, while the specificity decreases as *w* increases. The overall classification rate, which depends on the total number of suspicious and normal images classified correctly, reaches a maximum for a weight of approximately 0.3 – 0.6. A crossover occurs for $w = 0.78$, where the sensitivity becomes greater than the specificity; at the crossover point, the sensitivity, specificity and overall classification rate all have the same value of approximately 70 %. Since sensitivity is paramount for breast cancer screening, a weight factor of 0.995 was used in the remainder of the analysis, unless specifically stated otherwise. This weight factor selects the most sensitive feature subset first, and breaks ties between different subsets with the same sensitivity by selecting the one also showing the highest specificity.

Table 5.4 - Sensitivity and specificity of feature subsets selected by different values of weight factor in scoring metric, biorthogonal 3.7 basis

Weight Factor	Best Feature Triplet			Sensitivity (%)	Specificity (%)	Overall Classification Rate (%)
0.00	*K-v2*	*K-v7*	*K-d5*	26.6	87.8	70.8
0.05	*K-v2*	*K-v7*	*K-d5*	26.6	87.8	70.8
0.10	*S-h1*	*K-d7*	*K-a5*	40.5	87.3	74.3
0.20	*S-v1*	*S-d1*	*K-d2*	50.6	86.8	76.7
0.30	*S-v1*	*S-d1*	*K-d2*	50.6	86.8	76.7
0.50	*S-v1*	*S-d1*	*K-d2*	50.6	86.8	76.7
0.70	*S-v1*	*S-a5*	*K-d7*	63.3	76.6	72.9
0.80	*S-a3*	*K-v1*	*K-d7*	86.1	46.8	57.8
0.90	*S-a5*	*K-v8*	*K-a7*	92.4	25.6	44.4
0.95	*S-a5*	*K-v8*	*K-a5*	93.7	20.5	40.9
1.00	*S-a5*	*K-v8*	*K-a5*	93.7	20.5	40.9

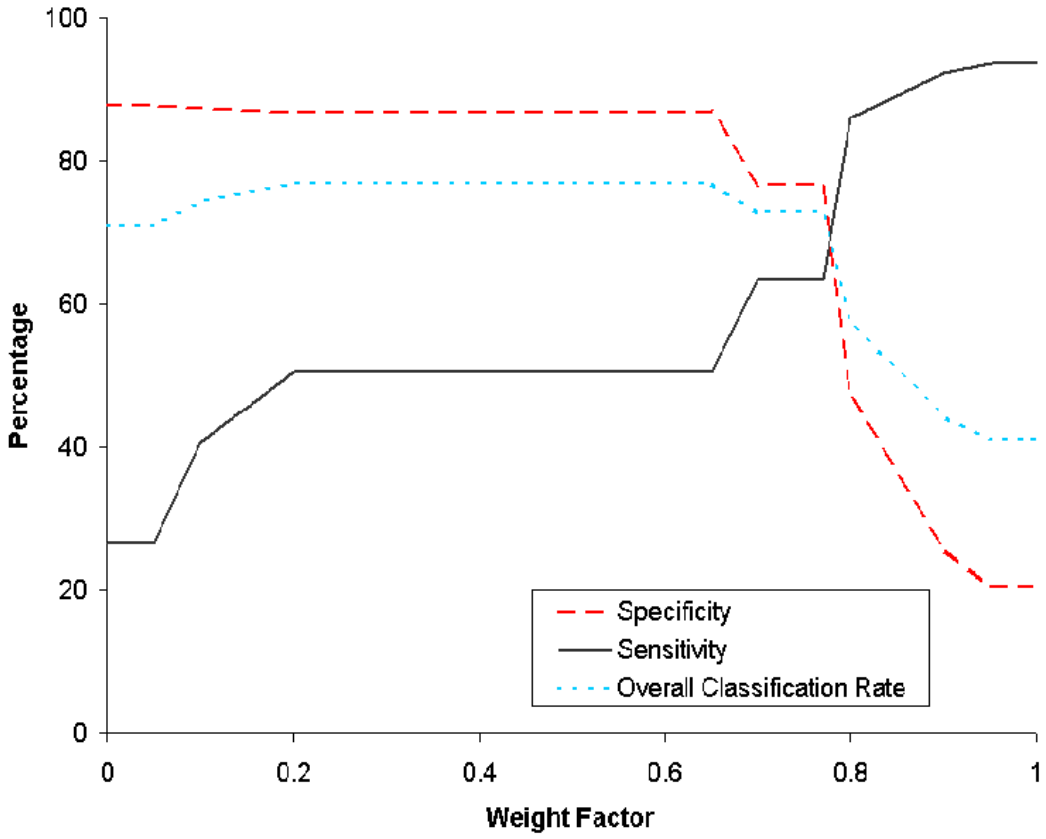


Figure 5.3 - Performance of best feature subset selected by scoring metric vs. choice of weight factor

5.2.3 *A priori* probability of relative frequency of each class

As stated in Section 4.6.3, the prior probability *P(c)* of each class appearing in Bayes' rule (equation 3.1) was chosen to be the same for all classes. This assumption was made because of the relatively greater consequences of misclassifying a suspicious sample compared with misclassifying a normal sample. The actual relative rate of incidence of the two classes, based on large-scale statistics, are that approximately 1 in 10 patients, or 1 in 20 images, show abnormalities [14]; . By this ratio, the prior probability for the normal class should be 0.95, and the prior probability for the suspicious class should be 0.05.

The problem with one class being far more likely than another is that, when the classifier is working correctly, it should then classify the vast majority of new samples into the more probable class. In the case of this research, the likelihood factor $P(f_i|c)$ in Bayes rule, calculated from the images' scalar features, would have to be 20 times larger for the suspicious than the normal class for a particular image in order for the whole probability to be larger for the suspicious class. Making the prior probabilities equal for both classes makes it more likely that an image will be classified as suspicious; this is desirable for this system, since false negatives are of more severe consequence for patients than false positives.

Table 5.5 shows the maximum sensitivity possible, given a particular choice of the prior probability of the suspicious class, using the skewness and kurtosis features from the biorthogonal 3.7 wavelet decomposition. For the final system, the probabilities were equal, so the prior probability was set to 0.50. For the prior probability to be representative of the actual incidence rate of cancer, it would have to equal 0.05.

Several intermediate values are also shown to demonstrate the continuous increase in sensitivity possible as the prior probability of the suspicious class is increased.

The drawback of increasing the prior probability of the suspicious class is that the specificity of the classifiers decrease, forcing a compromise to be chosen: the choice of setting the prior probabilities to be equal simplifies and speeds calculation while balancing sensitivity with specificity, making it a reasonable choice for this work.

Table 5.5 - Sensitivity of classifiers vs. prior probability of suspicious class, biorthogonal 3.7 basis

Prior Probability of Suspicious Class	Best Feature Triplet			Sensitivity (%)	Specificity (%)	Overall Classification Rate (%)
0.05	*S-h1*	*S-a1*	*S-a5*	3.1	99.0	68.0
0.10	*S-h1*	*S-v2*	*K-d7*	7.1	98.0	68.6
0.20	*S-h1*	*K-v7*	*K-d7*	19.4	92.7	69.0
0.30	*S-d6*	*S-d7*	*K-d7*	36.7	84.9	69.3
0.40	*S-h3*	*S-a8*	*K-a7*	70.4	40.5	50.2
0.50	*S-d6*	*K-d8*	*K-a7*	93.9	23.9	46.5
0.60	*S-d7*	*K-h6*	*K-v8*	100.0	9.3	38.6
0.70	*S-v1*	*S-d7*	*K-d7*	100.0	20.0	45.9
0.80	*S-h3*	*S-d7*	*K-v1*	100.0	12.7	40.9
0.90	*K-v8*	*K-a4*	*K-a8*	100.0	7.8	37.6

The sensitivity increased and the specificity decreased as the prior probability of the suspicious class was increased. The possible choices for the remainder of the work were the true prior probability, 0.05, and the balanced prior probability, 0.5; the other choices are shown to illustrate the trends in the sensitivity and specificity as the prior probability is varied. Because of the importance of sensitivity in screening procedures, the prior probability of 0.5 was selected and used in the remainder of this work.

5.3 Relative performance of different feature sets

Once the general structure of the single classifier had been finalized, the effect of different sets of features as inputs to the classifier was tested. The parameters from Section 5.2 were finalized as follows: the probability distributions were constructed with 8 bins each; the weight factor for the performance metric was set to 0.9950; and the prior probabilities were set to 0.50 for both the normal and the suspicious class. The feature sets were constructed by using a single wavelet basis' set of features, and using any two of the four possible types of statistical features in a given classifier. Section 5.3.1 compares the performance of different pairings of types of features, while Section 5.3.2 compares the performance of feature sets taken from different wavelet bases.

5.3.1 Comparing different statistical parameters

Four statistical parameters were measured from each wavelet map as described in Section 4.5: the mean intensity, standard deviation, skewness and kurtosis of pixel intensities in the tissue region. Choosing features from only two of these four types of features limited the number of potential features to 64, making the search process an order of magnitude faster for selecting the optimal feature subset for a classifier.

Table 5.6 shows the sensitivities and specificities of classifiers constructed from different potential statistical feature types. The tables use shorthand for the feature types: M for mean intensity, σ for standard deviation of intensity, S for skewness of pixel intensity, and K for kurtosis of pixel intensity. The most sensitive subset of three features was chosen from the one or two types of features listed in the first column. To ensure that the optimal types of statistical features were selected for the full system,

regardless of wavelet basis, all wavelets were tested for all possible feature type combinations. Appendix A collects the results of this analysis for the MIAS database, and Appendix B collects the results of the same analysis for the DDSM database. Only the mean performance results for the 11 wavelet bases' decompositions are shown in Table 5.6, though they are representative of the results for each individual basis.

Table 5.6 - Mean performances of different statistical feature types across all 11 wavelet bases tested

Features	Mean Sensitivity (%)	Mean Specificity (%)	Mean Classification Rate (%)
M	86.8	24.1	44.4
σ	87.0	27.7	46.9
S	91.9	20.6	43.6
K	93.5	16.8	41.6
$M + \sigma$	89.7	23.7	45.0
$M + S$	91.9	25.1	46.8
$M + K$	94.0	19.6	43.6
$\sigma + S$	93.2	23.1	45.8
$\sigma + K$	94.3	18.0	42.6
$S + K$	93.9	19.5	43.6

Several trends are apparent in the results of Table 5.6. Using a pool of two different types of features proved more sensitive than using either feature type alone. For example, using mean intensity and standard deviation together gave a sensitivity of 89.7%, though, by themselves, the sensitivities for mean intensity and standard deviation were only 86.8% and 87.0%, respectively. This improvement may be due to the different aspects of the original image that each type of feature is sensitive to, so that a classifier using two different types of statistical features can detect a larger fraction of the suspicious images in the data set.

The second trend is that the higher order features, skewness and kurtosis, were more sensitive than the lower order features, especially mean intensity. Section 4.5 discussed the mechanisms that may have made skewness and kurtosis especially sensitive to abnormalities, and these results support those hypotheses.

As a result of testing the different combinations of feature types to probe for selecting the optimal feature subset for a classifier, the combination of skewness and kurtosis features was used exclusively for constructing the final individual classifiers.

5.3.2 Comparing different wavelet bases

Table 5.7 shows the performance of the 11 wavelet bases tested using skewness and kurtosis features. Note that in Table 5.7, for the first time in this work, a two feature subset was selected over any three feature subset. As all one, two and three feature subsets were tested, this is an allowable occurrence and ensures that the most effective feature subset is selected, even when it is smaller than the nominal subset size.

There was only a small difference in sensitivity between the different bases; thus, the final classifiers were constructed using any one of the 11 possible wavelet bases. As the different wavelet basis functions are sensitive to different patterns in the signal, as shown by their structure in Figure 4.6 and Figure 4.7, they should be sensitive to different aspects of the original images and produce relatively different classification patterns for a data set. By allowing each individual classifier to use a different wavelet basis, there was a greater possibility of reducing the correlation between different classifiers. This would increase the effectiveness of a concerted-effort set of classifiers, that is, a set of classifiers operating co-operatively to classify a single image.

Table 5.7 - Relative performance of different wavelet bases

Wavelet Basis	Best Feature Triplet			Sensitivity (%)	Specificity (%)	Overall Classification Rate (%)
Haar	*S-h1*	*K-a2*	*K-a8*	90.8	32.7	51.5
Db 2	*K-a3*	---	---	93.9	14.1	39.9
Db 4	*K-h5*	---	---	94.9	9.3	37.0
Db 8	*S-h3*	*S-a4*	*K-d4*	91.8	27.3	48.2
Bior 1.5	*K-h3*	*K-a1*	*K-a8*	94.9	13.7	39.9
Bior 2.2	*K-a2*	---	---	94.9	14.1	40.3
Bior 2.8	*S-d5*	*K-a4*	---	94.9	27.3	49.2
Bior 3.7	*S-d6*	*K-d8*	*K-a7*	93.9	23.9	46.5
Bior 4.4	*K-h6*	*K-a2*	*K-a5*	93.9	16.1	41.3
Bior 5.5	*K-a1*	---	---	93.9	14.1	39.9
Bior 6.8	*S-h3*	*K-h7*	*K-a3*	94.9	22.0	45.5

5.4 Performance of classifiers tuned for particular abnormalities

Once the feature types used in the feature reduction step had been selected and the parameters of a single Bayesian classifier had been finalized, classifiers were designed to detect particular types of abnormalities. Specifically, classifiers were designed to detect only calcifications or to detect only masses. By combining these more specialized classifiers, the final concerted-effort set of classifiers was capable of classifying at a significantly higher rate than generalized classifiers that treat all types of abnormalities as equal.

5.4.1 Classifiers tuned to detect calcifications

To construct a classifier that was sensitive to calcifications, the scoring metric (equation 4.7) was modified: the true positive fraction was replaced by the fraction of correctly classified images showing calcifications, and the true negative fraction was replaced by the fraction of correctly classified images of all other types. In this case, the

correct classification for a mass is unsuspicious, since it is not suspicious for calcifications: other classifiers can then sort the unsuspicious images from this classifier to separate the masses from the normal images after this classifier has removed images containing calcifications that could interfere with classification. The consequences of misclassifying a mass with this classifier are not as severe as for a general classifier, however: if the image is misclassified as suspicious, it will still be subject to further study and will not necessarily lead to a false negative decision.

Table 5.8 shows the performance of classifiers constructed to be sensitive for calcifications using each of the 11 wavelet bases tested. The MIAS database contained 23 images with calcifications, so the table lists the number of missed calcifications for each classifier, to a maximum of 23. The *specificity for masses* column lists the percentage of masses that were classified as normal; as stated above, this classification for masses was desirable but not vital. The *specificity for normals* column lists the percentage of normal images classified as normal. A good classifier of this type should classify almost all calcifications correctly while maximizing its true negative fraction.

Table 5.8 - Performance of classifiers tuned to detect calcifications only

Wavelet Basis	**Best Feature Triplet**			**Misclassified Calcifications (out of 23)**	**Specificity for masses (%)**	**Specificity for normals (%)**
Haar	*K-h5*	*K-v2*	*K-d8*	0	65.2	58.5
Db 2	*S-v5*	*K-v3*	*K-v4*	0	93.3	65.9
Db 4	*K-h6*	---	---	0	82.7	67.8
Db 8	*S-v6*	---	---	0	52.0	53.2
Bior 1.5	*S-v2*	*S-d7*	*K-d7*	1	93.3	79.5
Bior 2.2	*S-h8*	*K-h6*	---	1	70.7	83.9
Bior 2.8	*S-v2*	*S-v5*	*K-v3*	1	53.3	73.2
Bior 3.7	*S-v7*	*S-d7*	*K-v8*	0	62.7	65.9
Bior 4.4	*S-v7*	*K-v6*	*K-v7*	1	73.3	67.3
Bior 5.5	*K-v7*	---	---	1	49.3	50.2
Bior 6.8	*S-h6*	*S-a3*	*K-v3*	0	74.7	72.7

The results in Table 5.8 show the merits of searching for calcifications specifically: no classifier missed more than a single calcification image, and 4 classifiers also maintained a specificity for normal images above 72%.

5.4.2 Classifiers tuned to detect masses

In analogy to the classification scheme for detecting calcifications only, classifiers can be designed to detect images showing masses only. Table 5.9 shows the performance of classifiers tuned to detect masses using the 11 wavelet bases tested. The table lists the number of masses misclassified as normal, to a maximum of the 75 total masses in the MIAS database, and lists the fraction of calcification and normal images that were correctly classified as unsuspicious for masses.

Table 5.9 - Performance of classifiers tuned to detect masses only

Wavelet Basis	Best Feature Triplet			Misclassified Masses (out of 75)	Specificity for calcifications (%)	Specificity for normals (%)
Haar	*S-d5*	*K-a6*	*K-a7*	0	91.3	23.9
Db 2	*K-h5*	*K-h8*	*K-a3*	0	95.7	43.4
Db 4	*K-d5*	*K-d6*	*K-a3*	0	87.0	40.0
Db 8	*S-h1*	*K-h7*	*K-h8*	0	100.0	42.4
Bior 1.5	*S-d2*	*K-d8*	*K-a7*	0	78.3	42.0
Bior 2.2	*S-h7*	*K-d6*	*K-a1*	0	100.0	29.3
Bior 2.8	*K-v5*	*K-a2*	*K-a4*	0	100.0	40.0
Bior 3.7	*K-d5*	*K-a2*	*K-a6*	0	91.3	36.1
Bior 4.4	*S-v1*	*S-d6*	*K-a6*	0	91.3	46.8
Bior 5.5	*K-d1*	*K-d3*	*K-a6*	0	82.6	40.0
Bior 6.8	*S-d5*	*S-d7*	*K-a5*	0	87.0	41.0

The results in Table 5.9 are highly promising. All 11 classifiers detected every mass, though the specificities for normal images were somewhat low; further, over 78%

of the calcification images were correctly removed from the pool of images suspicious for masses. The specificity for normal images is lower than for the calcification case in Table 5.8; this is likely due to the more subtle appearance of masses in images as compared to the sharp contrast of the small, bright calcification images.

The extremely high sensitivity for masses should not be considered evidence that these classifiers are perfectly sensitive to masses; rather, this points to a limitation in the original data set. The data set may contain too few images to be representative of all possible images showing masses, motivating the need to retest the system on a more extensive set of images. The DDSM data set of images will be tested in Section 5.6 to provide an alternate measure of the classifiers' sensitivities to masses. Also, this strong segmentation of the mass and calcification images into separate pools makes it effective to use classifiers tuned to particular abnormalities on each output pool of images afterwards.

5.5 Testing full system on MIAS database

Once individual classifiers had been trained, they could be combined to form one of the concerted-effort classifier designs described in Section 4.8: a sequential series of classifiers, a vote-taking combination of classifiers, and a network of classifiers tuned to detect specific types of abnormalities.

Because several classifiers are trained and tested together, care must be taken to ensure that no image is ever used as a training and as a testing image simultaneously. Each classifier is trained with the leave-one-out methodology, and the lone testing image is then passed into each classifier and its classification is recorded. This training

method ensures that multiple classifiers do not violate the requirement that a test image cannot be used as part of the training set.

5.5.1 Sequential series of classifiers

Several classifiers tuned for maximal sensitivity to any type of abnormality were combined to form a sequential series of classifiers, according to the method of Section 4.8.3.2. To increase the independence between individual classifiers, no wavelet basis was used more than once for the individual classifiers: this left 11 possible classifiers to choose from for the sequential design, one for each wavelet basis tested. Table 5.10 lists the five classifiers used and the sequence that they were used in. Each individual classifier was the most sensitive classifier possible for that wavelet basis, chosen from the classifiers in Table 5.7. The confidence levels for each classifier's outputs are also given in the normalized forms $C_{even}(N)$ and $C_{even}(S)$.

Table 5.10 – Performance of sequential series of classifiers

Basis of individual classifier		Number of images classified as normal		Number of images classified as suspicious		Confidence levels	
		Actually normal	Actually suspicious	Actually normal	Actually suspicious	$C_{even}(N)$ (%)	$C_{even}(S)$ (%)
1	**Bior 2.8**	56	5	149	93	84.3	56.6
2	**Bior 6.8**	23	0	126	93	100.0	60.7
3	**Bior 3.7**	9	1	117	92	81.1	62.2
4	**Haar**	50	5	67	87	82.7	73.1
5	**Db 8**	5	3	62	84	44.3	73.9

The sequential classifier reached its best performance with four stages; the fifth stage added little specificity while lowering the sensitivity significantly, and additional stages were not able to detect any more normal images. This limitation suggests that the

62 normal images remaining after the fifth classifier were found to be suspicious by every classifier available to construct the ensemble. Despite this limitation to the specificity, the sequential design worked well, removing significant numbers of normal images from the queue after each classifier while removing a minimal number of suspicious images. A cut-off could be implemented to classify all images reaching the n^{th} classifier as suspicious and all images removed before then as normal, if a hard classification was desired.

5.5.2 Vote-taking combination of classifiers

The second approach for combining classifiers into a concerted-effort set was to use a vote-taking scheme as outlined in Section 4.8.4. Three and five vote combinations of classifiers were tested: the three vote classifier had 8 output bins based on the outputs of each of the three classifiers while the five vote classifier had 6 output bins based on the number of classifiers that found an image to be suspicious.

Table 5.11 shows the results and confidence levels for the three vote combination of classifiers. The confidence level that an image in a bin is normal is given; the confidence that an image in that bin is suspicious is then one minus this level. The three classifiers used to build the concerted-effort set were those that together found the highest sensitivity when two or more of the three classifiers found an image to be suspicious. This meant that as few suspicious images as possible were missed by more than one of the three classifiers. In the case where two sets of classifiers had the same sensitivity, the classifier set featuring the highest specificity was selected. The classifiers used were chosen from the 11 classifiers listed in Table 5.7.

Table 5.11 – Confidence levels for three vote combination of classifiers

Classification by each classifier (N = normal, S = suspicious)			Normal images in bin	Suspicious images in bin	Confidence $C_{even}(N)$ (%)
Db 8	Bior 2.8	Bior 3.7			
N	N	N	32	4	83.0
S	N	N	0	1	0.0
N	S	N	17	0	100.0
N	N	S	2	0	100.0
S	S	N	0	1	0.0
S	N	S	22	0	100.0
N	S	S	5	4	43.2
S	S	S	127	88	46.8

The results for the three vote combination of classifiers are not as promising as the results for the sequential series of classifiers. Four of the suspicious images were missed by all three individual classifiers, while 127 of the normal images were found to be suspicious by all three classifiers. Further, the bins were erratically populated, though this is most likely due to the relatively small sample size; testing with the larger DDSM data set in Section 5.6.3 will reduce this effect.

The five vote classifier required a different binning procedure, since 32 different combinations of votes were possible, leading to 32 potential bins. Since there were only 98 suspicious samples in the MIAS data set, dividing them among 32 bins would not give an accurate representation of each bin's confidence levels as each bin would have a very small number of images. To compensate for this, only six bins were used, based on the number of classifiers in the set that classified an image as suspicious, regardless of the classification given by any particular classifier. Table 5.12 collects the results and the confidence levels for this design. The five classifiers used were chosen from those in Table 5.7 so that as few suspicious images as possible were missed by more than two classifiers. The five classifiers used were: Haar, Db2, Db8, Bior2.8 and Bior3.7.

Table 5.12 – Confidence levels for five vote combination of classifiers

Number of classifiers that found image to be suspicious	Normal images in bin	Suspicious images in bin	Confidence $C_{even}(N)$ (%)
0	5	4	43.2
1	18	0	100.0
2	23	1	93.3
3	5	2	60.4
4	81	7	87.6
5	73	84	34.6

The results for the five vote combination of classifiers were still not promising. At least four out of five classifiers found 154 of the 205 normal images to be suspicious, making it difficult to use this approach to remove these images from the population. As well, four suspicious images were missed by all five classifiers, eliminating the increase in sensitivity that a concerted-effort set of classifiers should provide.

A challenge with implementing the vote-taking set of classifiers is finding an effective selection tool for deciding which individual classifiers to use. Different combinations will distribute the normal and suspicious images among the bins differently; the choice of classifiers then depends on the goal of the particular application. For example, a system that is to remove as many normal images as possible would select a triplet with the majority of the suspicious images in one of very few bins to increase the confidence level that the other bins are normal. Because of this flexibility, the results shown in Table 5.11 and Table 5.12 should be viewed as two potential solutions and not as the optimal solutions to selecting individual classifiers to form vote-taking sets of classifiers.

5.5.3 Network of classifiers working in tandem

As described in Section 4.8.5, a network of individual classifiers tuned to detect particular types of abnormalities can be constructed. Because of the high sensitivities of these tuned classifiers, listed in Table 5.8 and Table 5.9, networks constructed from them can offer extremely high classification rates.

The performance of such networks can be shown pictorially. Figure 5.4 shows the performance of one network design: calcification images are filtered out first, then the other images are sorted in several passes to isolate images with masses. The individual classifiers are shown as boxes containing the type of wavelet basis and the type of abnormality the classifier is sensitive to. The two outputs from each classifier, normal or suspicious, are shown along with the number of each type of image that were classified in that way: for example, 103*n*, 23*m*, 10*c* means that 103 normal images, 23 mass images and 10 calcification images were given a particular classification by an individual classifier. The output bins are listed in the same way. The complete network offers 100% sensitivity and 46.4% specificity, with the high sensitivity caused by the precision of the tuned classifiers described above.

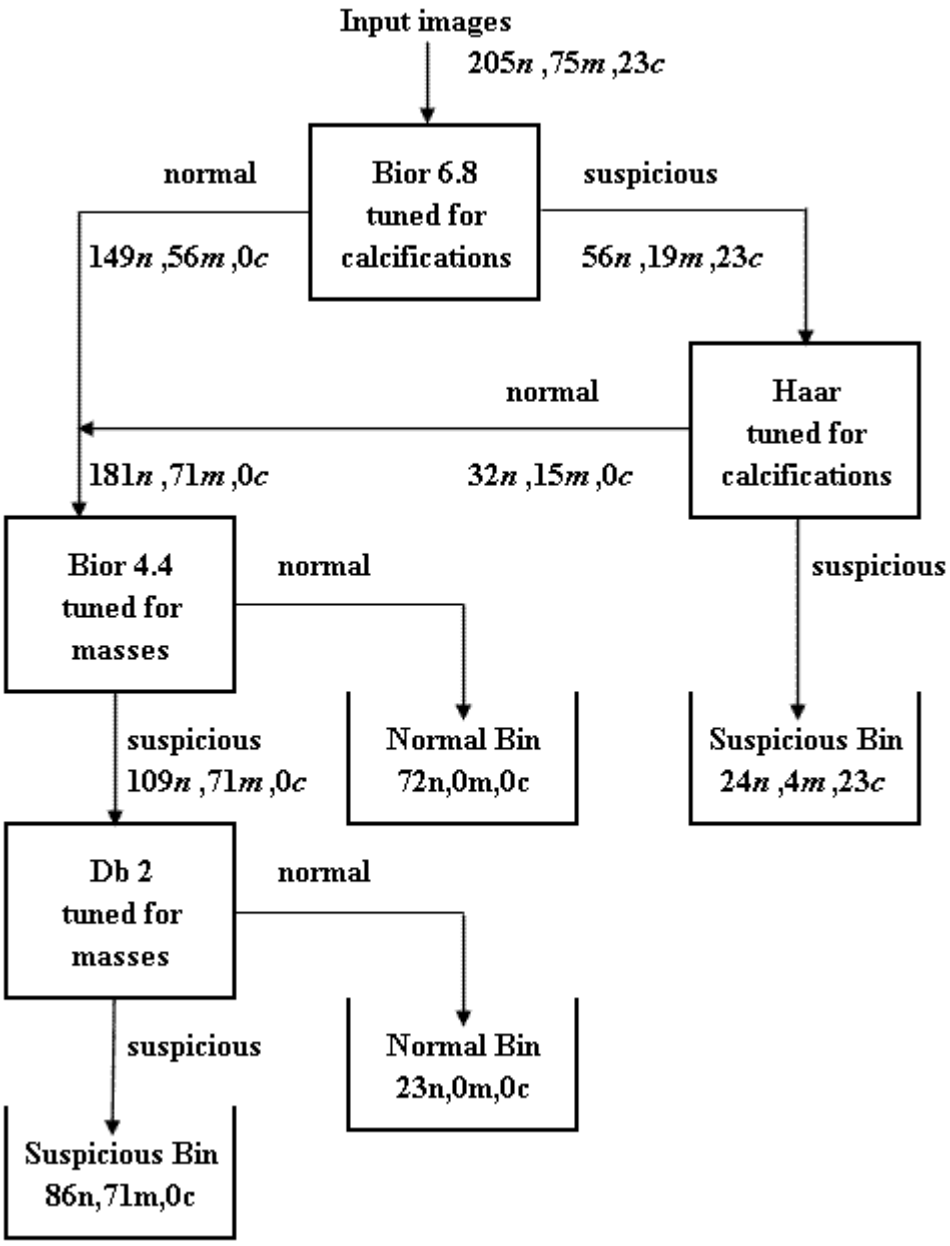


Figure 5.4 – Network for detecting abnormalities, detects calcifications first

Figure 5.5 shows an alternate design that first filters out the mass images, then sorts the remaining images to find the calcifications. Because the first classifier perfectly segments the calcification and mass images into two groups, two classifiers could be used on each output pool of images to achieve a high specificity. The additional classifier that is sensitive to masses is used to overcome the reduced

specificity that this type of classifier has compared with classifiers that are sensitive to calcifications. The whole network had a sensitivity of 100% and a specificity of 65.4%. While more classifiers could be added to such a network, minimizing the total number of features used to classify an image makes the system more flexible and less likely to become over-specified towards the data set used for training and testing.

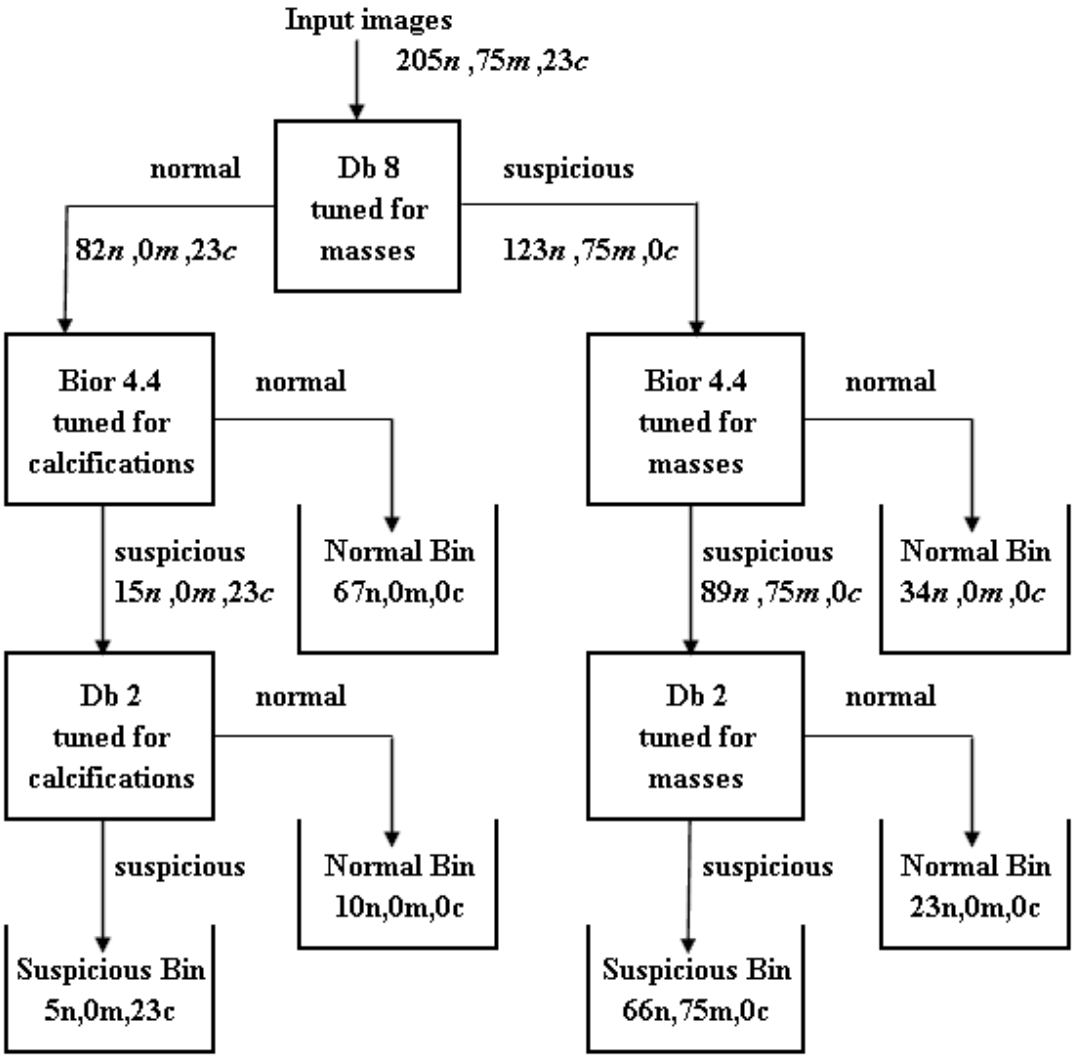


Figure 5.5 – Network for detecting abnormalities, detects masses first

5.6 Retesting full system on DDSM database

After testing and tuning the classification system using the MIAS data set, a fresh set of images were employed to retest the algorithm to ensure that it was not over-specified towards the MIAS data set. The DDSM data set [23] was used for this purpose: 1704 images were used including 1065 normal and 649 suspicious images. Table 5.13 lists the number of each type of image found in the data set. This larger data set also provided larger bin counts, improving the estimates for the confidence levels of the concerted-effort sets of classifiers developed in this work

Table 5.13 - Number and types of images in DDSM data set

Type of image	Number of images
Normal	1065
Total suspicious	649
Total benign	332
Total cancerous	317
All masses	410
Benign masses	213
Cancerous masses	197
All calcifications	239
Benign calcifications	119
Cancerous calcifications	120

The raw DDSM images were pre-processed in a fashion similar to the MIAS images as described in Section 4.3, with only a few changes. Specifically, the images were digitized using four different types of digitizers, leading to images of varying sizes and resolutions. All images were rescaled to have a resolution of 200 microns per pixel, matching the resolution of the MIAS database. As well, the images were cropped or padded as necessary to have a size of 1024 x 1024 pixels for uniformity and to ease the

computation of the 2D discrete wavelet transform of the images. Finally, one of the four digitizers converted the images into a percent transmission scale, rather than the optical density scale common in digital imaging. The optical density images were recovered from this digitizer's images by taking the natural logarithm of each pixel's value plus one in the percent transmission image. Adding one to the pixel intensity ensured that the output image pixels all had a value greater than or equal to zero, since an input intensity of zero would map to the natural logarithm of one, which is zero. The pixel values ranged from 0 to 65535 for the percent transmission image; thus, the pixel values in the optical density image ranged from 0 to 11.09. The intensities were then linearly rescaled to have pixel intensities in the range of 0 to 1 to match the intensity range of the images from the other three classifiers in the database.

Once the images were pre-processed, their wavelet transform was applied as in Section 4.4 and sets of scalar features were measured as in Section 4.5. The features were used to classify the images using the same algorithm as was used for the MIAS images. Section 5.6.1 gives the classification results for a single classifier, Section 5.6.2 gives the results for a sequential series of classifiers, Section 5.6.3 gives the results for a vote-taking scheme of classifiers, and Section 5.6.4 gives the results for a network of classifiers tuned to detect particular types of abnormalities.

5.6.1 Performance of individual classifiers

The DDSM images were classified by a single classifier at approximately the same sensitivity as the MIAS images. Table 5.14 gives the mean classification rates achieved using the different scalar feature type combinations, averaged over the 11

types of wavelet bases tested. Table 5.15 gives the classification rates and the best feature triplets for each wavelet base when the mean intensity and skewness features were used, as they provided among the highest sensitivities while maintaining relatively high specificities. Appendix B lists the performance of all 11 types of wavelets for each type of scalar feature used.

Table 5.14 - Mean performances of different statistical feature types across all 11 wavelet bases tested using DDSM data set

Type of Statistical Feature(s)	Mean Sensitivity (%)	Mean Specificity (%)	Mean Classification Rate (%)
M	89.2	26.6	50.3
σ	94.0	27.6	52.8
S	90.8	29.4	52.7
K	92.8	23.7	49.8
$M + \sigma$	97.4	33.9	57.9
$M + S$	97.2	38.1	60.5
$M + K$	96.1	35.6	58.5
$\sigma + S$	95.6	29.1	54.3
$\sigma + K$	96.3	28.5	54.1
$S + K$	94.2	32.2	55.7

Table 5.15 - Relative performance of different wavelet bases on DDSM data set using mean intensity and skewness features

Wavelet Basis	Best Feature Triplet			Sensitivity (%)	Specificity (%)	Overall Classification Rate (%)
Haar	*M-h1*	*M-d1*	*S-h3*	99.2	36.6	60.3
Db 2	*M-h3*	*M-d8*	*S-h5*	97.4	42.7	63.4
Db 4	*M-h8*	*M-d1*	*S-h5*	95.2	20.8	49.0
Db 8	*M-h6*	*S-v8*	*S-d3*	97.5	40.4	62.0
Bior 1.5	*M-d4*	*S-h6*	---	96.9	38.8	60.8
Bior 2.2	*M-h5*	*M-v2*	*S-d2*	98.8	44.8	65.2
Bior 2.8	*M-d4*	*S-d2*	*S-a5*	92.9	46.9	64.4
Bior 3.7	*M-d4*	*S-h4*	*S-d4*	98.9	28.1	54.9
Bior 4.4	*M-h1*	*M-d4*	*S-d2*	96.1	43.0	63.1
Bior 5.5	*M-h6*	*M-d5*	*S-d2*	98.5	38.1	61.0
Bior 6.8	*M-v3*	*M-d4*	*S-d2*	98.0	39.0	61.3

The performance of the system when tested and trained on the larger DDSM data set alone was significantly better than the performance using MIAS data set alone. This is likely due to the larger training set available to each classifier: a larger training set is likely to be more representative of all possible images from the normal and suspicious classes, improving the ability of the classifier to recognize and classify a new image correctly. The trends of the data are similar to the MIAS results: using two types of features together always outperformed using only one type of feature, though the standard deviation worked much better in this data set than in the MIAS data set.

Classifiers were also designed using the DDSM data set that were sensitive to masses or to calcifications in analogy with the classifiers using the MIAS data set in Section 5.4.1 and Section 5.4.2. Table 5.16 shows the results for a classifier tuned to detect calcifications only, and Table 5.17 shows the results for a classifier tuned to detect masses only. The classifiers used only the mean and skewness features in the

feature reduction step because of their high performance on the data set as described above.

Table 5.16 - Performance of classifiers tuned to detect calcifications only, using DDSM data set

Wavelet Basis	Best Feature Triplet			Misclassified Calcifications (out of 239)	Specificity for masses (%)	Specificity for normals (%)
Haar	*M-h1*	*M-d2*	*S-v4*	0	93.4	80.8
Db 2	*M-h1*	*M-d4*	*S-d8*	0	92.7	76.0
Db 4	*M-h1*	*M-v1*	*S-d8*	0	97.1	73.9
Db 8	*M-d3*	*S-v5*	*S-a3*	0	99.0	76.2
Bior 1.5	*M-h4*	*M-h8*	*S-a6*	0	99.5	65.3
Bior 2.2	*M-h2*	*M-d6*	*S-v4*	0	91.5	75.9
Bior 2.8	*M-h1*	*M-d5*	*S-a5*	0	99.3	67.4
Bior 3.7	*M-h4*	*M-v2*	*S-v7*	0	87.1	73.9
Bior 4.4	*M-h1*	---	---	0	97.3	65.6
Bior 5.5	*M-d1*	*M-d3*	*S-v3*	0	87.1	71.5
Bior 6.8	*M-h6*	*M-d3*	*S-h6*	0	96.3	75.7

Table 5.17 - Performance of classifiers tuned to detect masses only, using DDSM data set

Wavelet Basis	Best Feature Triplet			Misclassified Masses (out of 410)	Specificity for calcifications (%)	Specificity for normals (%)
Haar	*M-d3*	*S-h3*	*S-d4*	1	79.1	70.5
Db 2	*M-d3*	*S-h6*	*S-d4*	1	92.9	72.3
Db 4	*M-d6*	*S-h4*	*S-d5*	1	86.2	66.9
Db 8	*M-v7*	*M-d5*	*S-d3*	1	74.1	64.8
Bior 1.5	*M-d5*	*S-d1*	*S-d3*	1	85.8	61.2
Bior 2.2	*M-d4*	*S-d2*	*S-d6*	1	73.6	67.1
Bior 2.8	*M-d5*	*S-d3*	*S-d5*	1	63.6	62.5
Bior 3.7	*S-h6*	*S-v7*	*S-d5*	1	90.0	56.8
Bior 4.4	*S-d4*	*S-d5*	*S-a8*	1	75.7	72.5
Bior 5.5	*M-d4*	*S-h4*	*S-d4*	1	89.1	71.9
Bior 6.8	*M-v8*	*S-v1*	*S-d5*	1	95.4	65.2

The classifiers tuned to detect calcifications showed a higher specificity than those tuned to detect masses, as was the case when using the MIAS data set in Section 5.4. The classifiers tuned to detect masses all missed one mass; further, the same mass was missed by all 11 classifiers. The image that was universally misclassified is shown in Figure 5.6. The mass' location is shown with a white arrow; the mass is classified as subtle with obscured borders according to the DDSM, and appears to be circular and slightly darker than the surrounding tissue. Excluding this image, all tuned classifiers had sensitivities of 100% for the type of abnormality they were designed to detect. Though the mass is not easily visible in the image, it appears to be present, and the image was retained in the data set, limiting the sensitivity of all of the classifiers that were tuned to detect masses only.

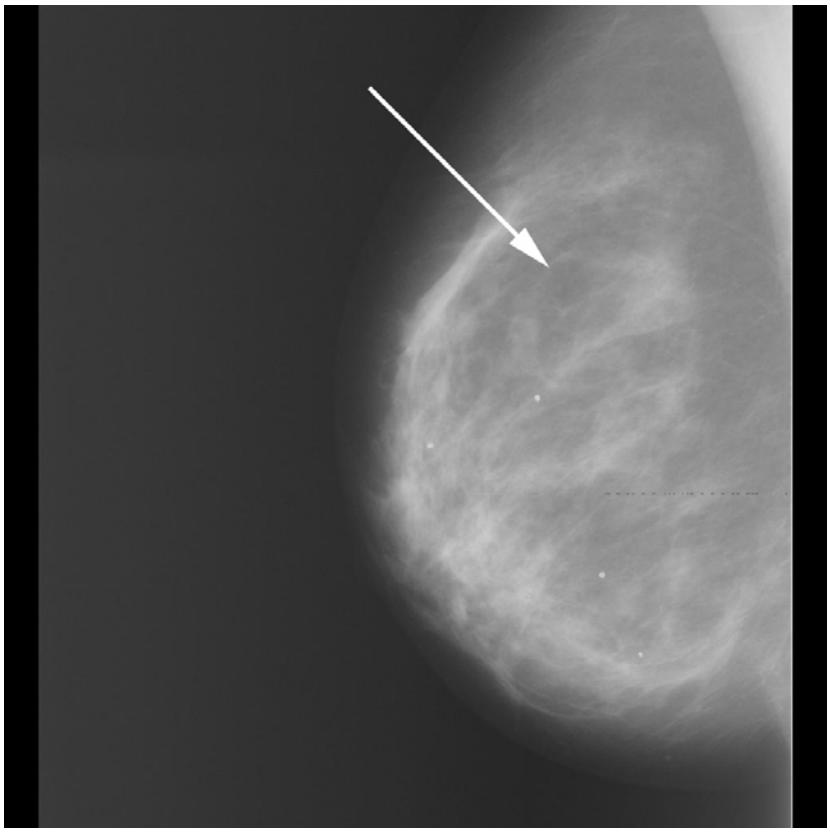


Figure 5.6 – Image showing benign mass missed by all 11 classifiers

5.6.2 Sequential series of classifiers

Using the individual classifiers listed in Table 5.15, a sequential set of classifiers was designed as in Section 5.5.1 to provide confidence levels for classifying images from the DDSM data set. Table 5.18 lists the classifiers in the order they were used in the ensemble and the confidence levels of each of their outputs. The number of images of each type classified as normal or suspicious by each classifier are also provided.

Table 5.18 – Performance of sequential series of classifiers

Basis of individual classifier		Number of images classified as normal		Number of images classified as suspicious		Confidence levels	
		Actually normal	Actually suspicious	Actually normal	Actually suspicious	$C_{even}(N)$ (%)	$C_{even}(S)$ (%)
1	**Haar**	390	5	675	644	97.9	61.0
2	**Bior 3.7**	268	2	407	642	98.8	72.1
3	**Bior 2.2**	103	4	304	638	94.0	77.5
4	**Bior 6.8**	131	7	173	631	91.9	85.7
5	**Db 2**	87	10	86	621	84.1	92.2
6	**Bior 5.5**	20	6	67	615	67.0	93.8
7	**Bior 1.5**	21	8	45	607	61.5	95.7
8	**Db 8**	15	11	30	596	45.4	97.0

Several trends are apparent in these data. The confidence levels for the images in the normal bins is high until the sixth classifier, when the numbers of suspicious and normal images classified as normal become comparable. The confidence that images classified as suspicious and passed on to more classifiers increases after every stage as relatively fewer and fewer normal images remain in the suspicious bins. A recommended cut-off for this sequence of classifiers would be to stop after the fifth classifier, since the final three classifiers provide much lower confidence levels for the images that they classify as normal.

5.6.3 Vote-taking combination of classifiers

Table 5.19 shows the results and confidence levels for the three vote combination of classifiers. The confidence level that an image in a bin is normal is given; the confidence that an image in that bin is suspicious is then one minus this level. The three classifiers used to build the concerted-effort set were those that together found the highest sensitivity when two or more of the three classifiers found an image to be suspicious. This meant that as few suspicious images as possible were missed by more than one of the three classifiers. In the case where two sets of classifiers had the same sensitivity, the classifier set featuring the highest specificity was selected. The classifiers used were chosen from the 11 classifiers listed in Table 5.15.

Table 5.19 – Confidence levels for three vote combination of classifiers

Classification by each classifier (N = normal, S = suspicious)			Normal images in bin	Suspicious images in bin	Confidence $C_{even}(N)$ (%)
Haar	Bior 2.2	Bior 5.5			
N	N	N	64	3	92.9
S	N	N	299	0	100.0
N	S	N	1	0	100.0
N	N	S	65	0	100.0
S	S	N	42	7	78.5
S	N	S	49	5	85.7
N	S	S	260	2	98.8
S	S	S	285	632	21.6

The performance of the vote-taking combination of classifiers is stronger for the DDSM data set than for the MIAS data set. By keeping only images with confidence levels below 90% for being normal, the system provides 64.7% specificity and 99.2% sensitivity; by keeping images with confidence levels below 80% only, the system provides 69.3% specificity and 98.5% sensitivity.

The five vote classifier used a different binning procedure, since 32 different combinations of votes were possible, leading to 32 potential bins. To gain better statistical information in each individual bin, only six bins were used, based on the number of classifiers in the set that classified an image as suspicious, regardless of the classification given by any particular classifier. Table 5.20 collects the results and the confidence levels for this design. The five classifiers used were chosen from those in Table 5.15 so that as few suspicious images as possible were missed by more than two classifiers. The five classifiers used were: Haar, Db2, Db8, Bior2.8 and Bior3.7.

Table 5.20 – Confidence levels for five vote combination of classifiers

Number of classifiers that found image to be suspicious	Normal images in bin	Suspicious images in bin	Confidence $C_{even}(N)$ (%)
0	3	2	47.8
1	1	0	100.0
2	427	1	99.6
3	158	5	95.1
4	290	64	73.4
5	186	577	16.4

Again, the results for the DDSM data set are better than for the MIAS data set due to the better representation of relative binning probabilities that the larger data set provides. By setting a confidence threshold at 95%, this system achieves a sensitivity of 99.1% and a specificity of 55.0%.

5.6.4 Network of classifiers working in tandem

As in Section 5.5.3, two networks were designed using individual classifiers tuned to detect particular types of abnormalities. Figure 5.6 uses the same structure as the network in Figure 5.4, but uses different individual classifiers that are trained using the DDSM data set. The classifiers are chosen from the tuned classifiers in Table 5.16 and Table 5.17. The whole network achieved a sensitivity of 99.85% with a specificity of 76.4%. The only suspicious image misclassified was the same mass image shown in Figure 5.5 above. The network also achieved a strong segmentation between images with calcifications and images with masses, potentially providing even more information about an image processed with this system than just its level of suspiciousness. The four output bins from this network consist of: two bins containing mainly normal images with a very small fraction of mass images; one bin containing mainly calcifications with a small fraction of masses and normal images; and one bin containing mainly masses with a small fraction of normal images.

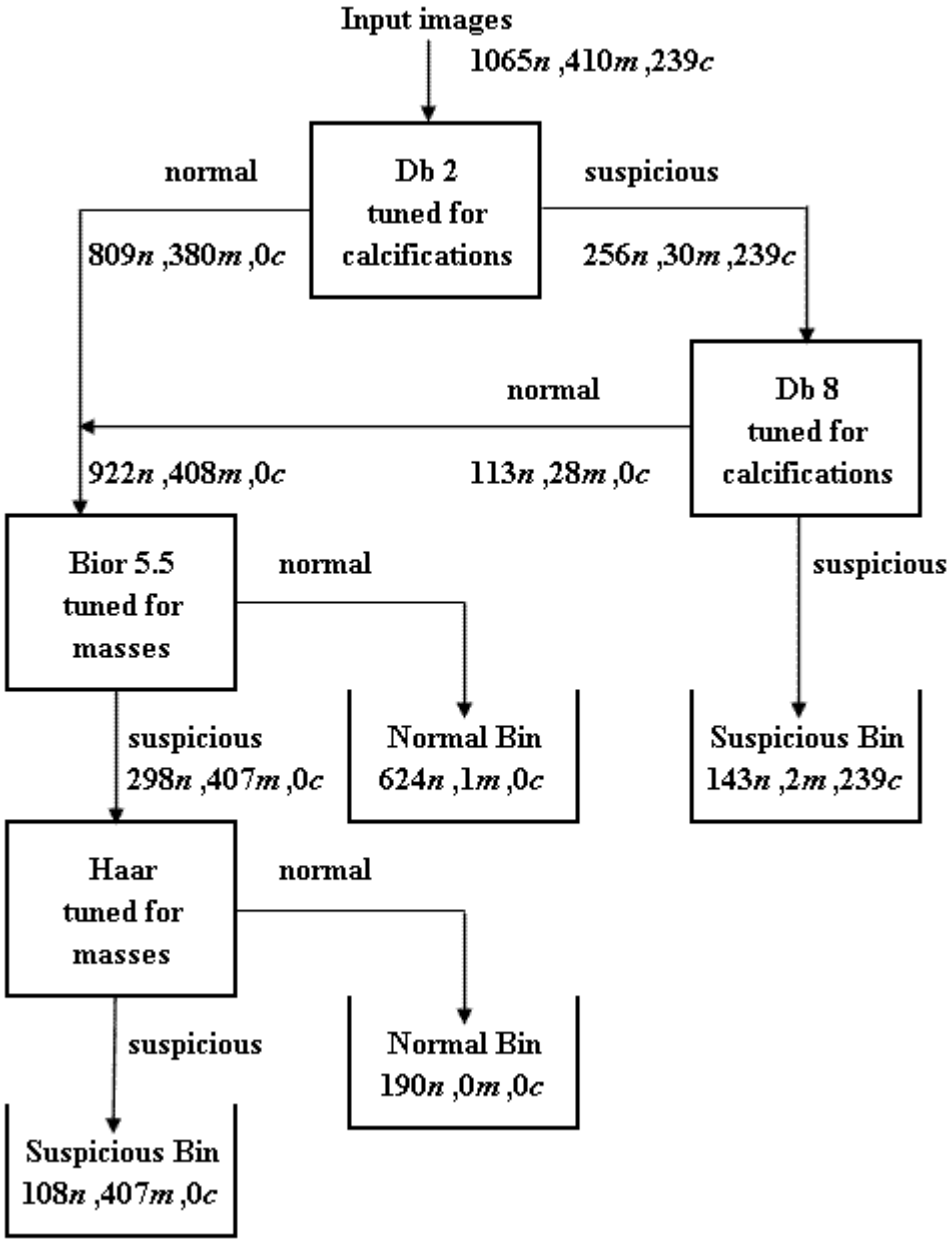


Figure 5.6 – Network for detecting abnormalities, detects calcifications first

The other network design tested was analogous to the design in Figure 5.4 but using classifiers trained on the DDSM data set. The first classifier was the one which best segmented masses and calcifications into two separate pools, then two classifiers were sequentially run on each pool to remove as many normal images as possible. The

network achieved a sensitivity of 99.85% and a specificity of 78.8%. The only misclassified suspicious image was the mass image shown in Figure 5.5. The network design has six output bins for images: three bins contain only normal images; one bin contains mainly normal images with a very small fraction of masses; one bin contains mainly masses with a fraction of normal images; and one bin contains calcifications with a fraction of normal images and a small fraction of mass images.

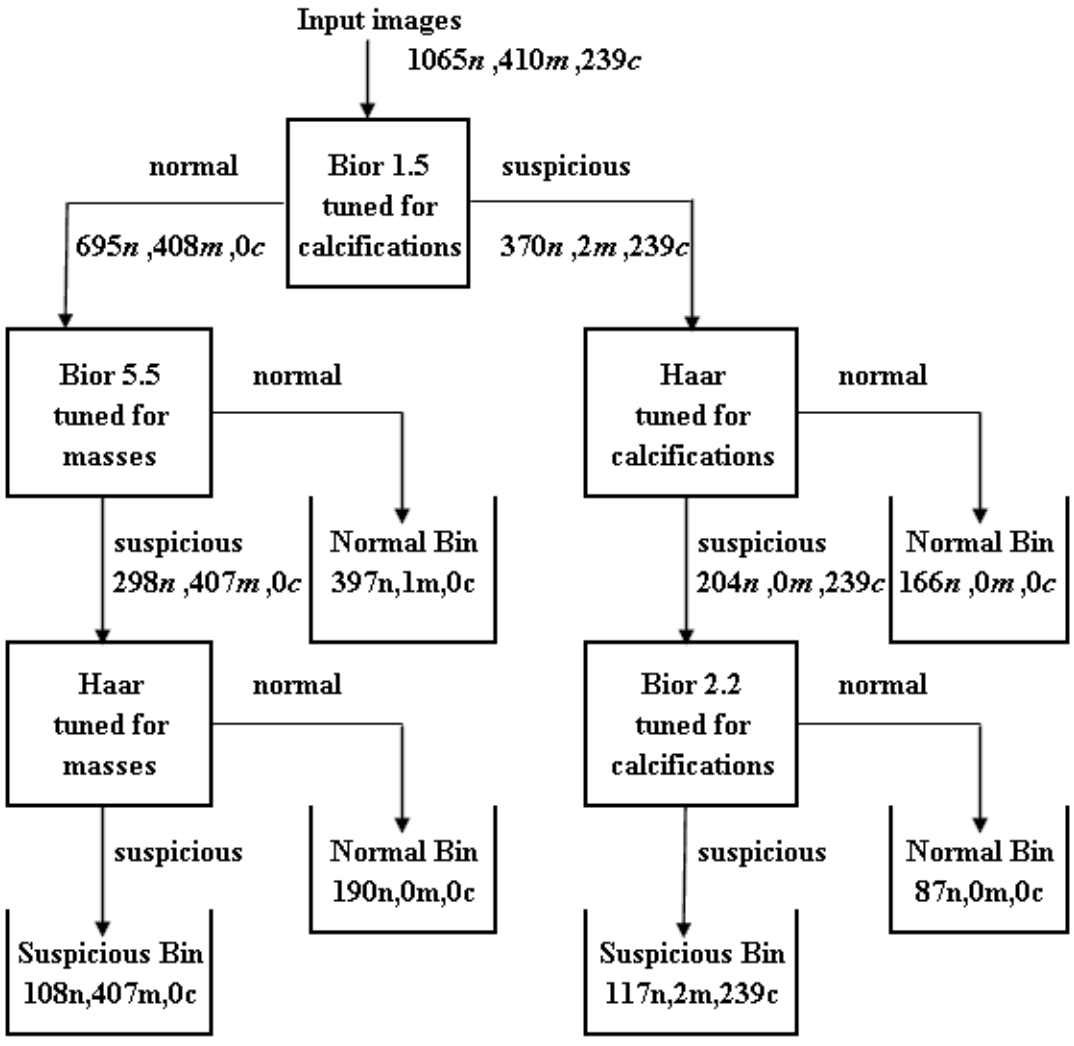


Figure 5.7 – Network for detecting abnormalities, detects masses first

CHAPTER 6 – DISCUSSION

The system developed in this work was designed to classify x-ray mammography images as either normal or suspicious and provide a confidence level for this classification. The system was built from individual classifiers; the performance of these classifiers is discussed in Section 6.1. Several variations were tested for combining the individual classifiers; their relative performances are described in Section 6.2. The system was developed and tested on the MIAS data set containing 303 images with verified diagnoses. A larger independent second image set, the DDSM data set, consisted of 1714 images and was used to confirm that the system's performance was not dependent on the data set used and to examine the effect of a larger training set on system performance, as discussed below. The two data sets were kept separate throughout the analysis, ensuring that the DDSM data set could always be used as an independent check of the system's behaviour once it had been developed using the MIAS data set.

6.1 Performance of single naïve Bayesian classifiers

Each individual classifier was constructed by selecting 3 features from a particular wavelet basis' decomposition. For the MIAS data set, the most effective features were skewness and kurtosis features, likely due to their sensitivity to pixels with

intensities much different from the norm, such as those showing bright calcifications. For the larger DDSM data set, mean intensity and skewness features performed best. For every choice of feature type, the DDSM data set showed higher sensitivities to abnormal images; this was most likely due to the larger training set size that better represented the distributions of normal and suspicious images that each classifier used classify a new, unknown image.

For the MIAS data set, the strongest classifier used the Biorthogonal 2.8 basis to achieve a sensitivity of 94.9% and a specificity of 27.3%. The different wavelet bases tested all performed similarly, with sensitivities ranging from 90.8% to 94.9%, suggesting that the classifiers' performances were not strongly dependent on the type of wavelet used in the decomposition. The specificities for the MIAS data set were relatively low, ranging from 9.3% to 32.7%. The low specificity of the individual classifiers was compensated for in the final system by combining the outputs of multiple classifiers.

For the DDSM data set, the strongest classifier used the Haar wavelet basis to achieve a sensitivity of 99.2% and a specificity of 36.6%. The different wavelet bases tested all performed well, with sensitivities ranging from 92.9% to 99.2%, again suggesting that the classifiers' performances were not strongly dependent of the choice of wavelet basis. The specificities of the classifiers were higher than for the MIAS data set, ranging from 20.8% to 46.9%, likely due to the larger training set size.

Classifiers were also designed that detected just one type of abnormality, either masses or calcifications. These classifiers showed extremely high sensitivities as well

as strong specificities, allowing for extremely sensitive concerted-effort sets of classifiers to be designed.

For the MIAS data set, 6 of the 11 wavelet bases generated classifiers that detected 100% of the 23 calcification images and the other 5 bases missed only one calcification; further, each classifier maintained a specificity of 50.2% to 82.9%. As well, all 11 wavelet bases generated classifiers that detected 100% of the 75 mass images while maintaining specificities between 23.9% and 43.4%.

For the DDSM data set, all 11 wavelet bases generated classifiers that detected 100% of the 239 calcification images; further, each maintained a specificity of 65.3% to 80.8%. As well, all 11 wavelet bases generated classifiers that detected all but one of the 410 mass images while maintaining specificities between 56.8% and 72.5%. The one mass image missed was the same in all cases, and may be an erroneous image, though it showed no obvious artefacts or defects.

The generally lower specificities for classifiers tuned to detect masses is likely due to the more subtle appearance of masses as compared to calcifications in images. This discrepancy in specificities was especially pronounced in the MAIS data set, which contains 19 masses that are only visible through architectural distortions and are notoriously difficult to detect using CAD techniques [4]. Achieving 100% sensitivity when architectural distortion images were present is very encouraging, as most CAD systems in literature are not sensitive to these abnormalities.

6.2 Performance of concerted-effort sets of classifiers

Three techniques were developed and tested for combining individual classifiers into a system capable of providing confidence levels for image classification: a sequential series of classifiers, a vote-taking scheme of classifiers, and a network of classifiers tuned to detect particular types of abnormalities.

The sequential series of classifiers passed images on to the next stage of the classifier if they were found to be suspicious by the previous stage. Using this method, the normal bin from each stage had a slightly lower confidence level than the previous stage's bin, providing a natural progression for applying a thresholding procedure. After four stages for the MIAS data set, the series retained 88.8% of the suspicious images while removing 69.8% of the normal images as being unsuspicious, meaning that the system was 88.8% sensitive and 69.8% specific after four stages. After five stages for the DDSM data set, the system was 95.7% sensitive and 91.9% specific.

These results are highly encouraging, especially for the DDSM data set, and exceed the performance of many current approaches. For example, the CAD systems in current clinical use generate approximately 0.5 false negative regions per image [4]; though their sensitivities are close to that of the system in this work, their low specificities limit their effectiveness as a pre-screening tool. This system's performance values also compare well with the performance of human readers: though there is a large range in the reported classification rates of human readers, their performance values tend to be between 75% and 90% for both sensitivity and specificity [27].

The vote-taking scheme of classifiers gave confidence levels based on which classifiers found an image to be suspicious. This method did not work well in practice,

especially for the relatively smaller MIAS data set, since it was difficult to develop sufficient counts in each output bin to calculate useful confidence levels. Using the MIAS data set with a confidence level cut-off of 80%, that is, marking every image with a likelihood of normalcy below 80% as suspicious, a three vote classifier scheme achieved a sensitivity of 93.9% and a specificity of 35.4%. For a five vote classifier scheme, a confidence level cut-off of 80% gave a sensitivity of 91.8% and a specificity of 59.5%.

Using the DDSM data set for training offered slightly better results. A three vote classifier scheme with a confidence level cut-off of 80% gave a sensitivity of 98.5% and a specificity of 69.3%. A five vote classifier scheme with a confidence level cut-off of 80% gave a sensitivity of 99.1% and a specificity of 55.0%.

For both three and five vote classifier schemes using either the MIAS or the DDSM data sets, the complete systems perform little better than individual classifiers trained on the same data sets and are not as powerful as the sequential classifiers or the networks of tuned classifiers. Using more than five classifiers would have required too many output bins, many of which were under-populated even for the three and five vote classifier schemes, while using less than three classifiers would have produced too few output bins to sort images into an effective number of possible confidence levels.

The network of classifiers tuned to detect particular types of abnormalities offered extremely high sensitivity, making it a useful candidate for isolating normal images and removing them from a population without losing suspicious images. Several representative networks were designed that isolated one type of abnormality and then the other.

For the MIAS data set, a network that first removed images with calcifications (Figure 5.3) offered a sensitivity of 100% and a specificity of 46.4%. A network that first segmented images with calcifications and images with masses into two pools (Figure 5.4) offered a sensitivity of 100% and a specificity of 65.4%. Because the networks were perfectly sensitive, confidence levels did not add to the information produced by the classifier network: effectively, an image classified into a normal bin had a 100% confidence of being normal.

For the DDSM data set, a network that first removed images with calcifications (Figure 5.6) offered a sensitivity of 99.85% and a specificity of 76.4%. A network that first segmented images with calcifications and images with masses into two pools (Figure 5.7) offered a sensitivity of 99.85% and a specificity of 78.8%. In both cases, the only misclassified suspicious image was the mass image that every individual classifier missed, as mentioned above.

The results for both data sets were promising using the networks of tuned classifiers. The systems all had sensitivities above 99.8%; thus, any images identified as normal were unlikely to be false negatives. Further, the networks, particularly those trained using the DDSM data set, had acceptably high specificities of above 75%, making these networks capable of identifying a significant fraction of normal images without missing many images that merit further analysis.

The concerted-effort sets of classifiers performed as well as or better than other classifier ensembles in literature, though direct comparisons between the current work and other works are difficult to make. This work was novel in classifying whole images, making it fundamentally different from the variety of methods that locate

suspicious regions and do not classify images as a whole. One example that can be compared to and that also uses multiple classifiers is the Adaboosting method [40]: it achieved 95% sensitivity with 0.591 false positives per image for a single classifier or 0.327 false positives per image for an ensemble of classifiers. The sequential series of classifiers in this work offered a higher specificity than this at a comparable sensitivity, while the network of tuned classifiers achieved a higher sensitivity with a comparable specificity. Again, the Adaboosting method identified suspicious regions in images and did not identify and isolate normal images from the input image pool in the way that the method in this work did.

The primary advantage of the use of whole image features extracted from wavelet maps of mammography images was in the classification of whole images rather than of image regions. Each wavelet map examined a different scale and type of detail, making it possible to search for abnormalities of different sizes and shapes; in contrast, examining a raw image alone limits the range of abnormalities that can be detected, since structures of different sizes and shapes would require different detection procedures. The skewness and kurtosis features were especially sensitive to abnormalities, likely due to their strong dependence on pixels with intensities much different from the norm, such as those generated by the sharp boundaries of calcifications or the brightened regions of masses.

The individual classifiers were limited to have no more than three features, though this requirement could be relaxed in future applications. The limitation was imposed to prevent over-learning: if too many features were used in a concerted-effort set of classifiers, the system would be more likely to become biased towards the training

data and lose the flexibility necessary to apply the system to new images acquired in a clinical setting.

CHAPTER 7 – CONCLUSIONS

Breast cancer is the most commonly diagnosed cancer among Canadian women and the second-leading cause of cancer-related death behind lung cancer. X-ray mammography uses low energy x-ray absorption imaging of the soft tissues of the breast to screen for cancer and is the leading method for breast cancer screening. As the population ages and as policies call for increased screening frequencies, the volume of mammography images taken will continue to increase. Automated techniques are being introduced to increase screening sensitivity and manage the increasing data volume. This work introduced a method for pre-screening images to rate their degree of suspiciousness and help determine the need to further analyze particular images.

The first objective of this work was to develop pre-processing procedures for regularizing the appearance of images to isolate only salient differences between normal and suspicious images. This was successfully done through a series of masking and intensity normalization steps. The pre-processing of images needed to be done slightly differently for each data set used; for this system to be viable in broad use, some set of standards would need to be developed for the appearance of images presented to the algorithm.

The second objective of this work was to apply the discrete wavelet transform to parse the images and extract statistical features that characterize an image's content.

This was successfully done using 11 different wavelet bases, all of which generated useful features. The features extracted from the wavelet map decomposition were whole-image scalar values: the mean intensity, the standard deviation of the intensity, and the skewness and kurtosis of the intensity. The use of wavelets was effective, as it decomposed the information content of each image into different scales, isolating features of different sizes, such as large masses and small calcifications. The extraction of scalar features from the wavelet maps also gave encouraging results: the distinction between the normal and suspicious images was most pronounced with the higher order skewness and kurtosis features, though all four types showed reasonable sensitivity. It is possible that higher order features or other scalar parameters could be even more effective, given further study.

The third objective of this work was to use the measured features to classify images as either normal or suspicious with a corresponding confidence level. A naïve Bayesian classifier was employed to classify the images, and did so with sensitivities as high as 94.9% for images in the MIAS data set and 99.5% for images in the DDSM data set. The higher sensitivity for the DDSM data set was likely due to its larger size, as this offered a larger training set using the leave-one-out training methodology. A larger training set should better model the realistic distribution of features from the normal and suspicious classes; applying this system in a large-scale clinic should offer high sensitivity as well, if a large set of images with confirmed diagnoses are used as a training set for the classifier system.

To generate confidence levels, multiple classifiers were combined into concerted-effort sets of classifiers in three different ways. A sequential series of

classifiers removed all images classified as normal after each classifier, paring down the pool of suspicious images and increasing specificity significantly. The sequential series of classifiers appeared promising and would be a good candidate for further use. A vote-taking scheme of classifiers based the confidence that an image was normal on which classifiers found the image to be either normal or suspicious; this method did not offer significant gains from the individual classifiers, and would be a poor candidate for further use. Finally, a network of classifiers, each tuned to detect a particular type of abnormality, was designed; this method offered sensitivity greater than 99.8% while maintaining a specificity above 60%, making it the strongest candidate method for pre-screening images. Because of its high sensitivity, however, it was not possible to apply meaningful confidence levels to its outputs, as the confidence levels for the normal bins were all 100 % that the image was normal.

The confidence levels were measured using two methods, the second of which was uniquely designed for this work: first, a real confidence level was given that measured the true probability that an image in a particular bin was suspicious; and second, a normalized confidence level was given that assumed that normal and suspicious images were equally likely to occur. Since suspicious images are much less common than normal images in a screening protocol such as x-ray mammography, the true probabilities strongly favour any given image as being normal; the adjusted probabilities remove this dependence on population incidence rates and measure an image's raw degree of suspiciousness. This technique allows for more flexibility in providing confidence levels, since other factors could be incorporated into its result: for example, the rate of suspicious results varies with age and family history for a patient;

these factors could be applied to the raw confidence level to give a better true confidence level than one that assumes the rate of suspicious images to be the same for all patients of varying backgrounds.

The system designed in this work was successful. It achieved sensitivities exceeding current approaches in literature and practice while maintaining reasonable specificity, especially for the sequential series of classifiers and for the network of tuned classifiers. This method was novel in that it classified whole images as normal or suspicious, rather than the common approach of identifying suspicious regions in most images; this approach was possible only because the wavelet decomposition provided features with such high sensitivity to suspicious images. This whole-image approach made it difficult to directly compare the developed system's performance against others in literature, though typical rates of at least one false positive region per image corresponds to a specificity significantly lower than the current system's, since few images would be analyzed and found to contain no suspicious regions.

Potentially, using this system to pre-screen images could significantly increase the rate at which x-ray mammography images could be screened for signs of cancer. By identifying and removing images with a low degree of suspiciousness, more analysis could be applied to images that merit it. Increasing throughput in this way allows for two gains in screening procedures: screening could be done more frequently for each patient, and screening could begin at an earlier age. For example, the screening recommendation could be changed to an annual cycle beginning at age 40 from the current recommendation in Canada of a mammogram every two years from age 50. Both changes to screening protocols would increase the chance of detecting a cancer at

an earlier stage when it is more treatable while limiting the increased risk to the patient from the radiation dose received from the additional imaging procedures. The challenge to using this approach in practice is to ensure a sensitivity high enough to allow a medical professional to make an informed decision based on the classification and confidence level provided by the system.

An extension of this work could be to attempt to differentiate between benign abnormalities and malignant ones; since only 5% of the images showing abnormalities are positive for cancer, the system developed here still leaves a large volume of images negative for cancer, even when its specificity is maximized. Given the high sensitivities achieved using scalar features measured from wavelet maps, it is possible that this technique could be applied to the pool of suspicious images to segment it into cancerous and non-cancerous classes and again provide confidence levels for this secondary classification.

REFERENCES

[1] Abrami, A. *et al. Medical applications of synchrotron radiation at the SYRMEP beamline of ELETTRA.* Nuclear Instruments and Methods in Physics Research A. **548**: 221 (2005).

[2] Amendolia, S. R. *et al. The CALMA project: a CAD tool in breast radiography.* Nuclear Instruments and Methods in Physics Research A. **460**: 107 (2001).

[3] Astley, Susan M. *Computer-Aided Detection for Screening Mammography.* Academic Radiology. **11**: 1139 (2004).

[4] Astley, S. M. and F. J. Gilbert. *Computer-aided detection in mammography.* Clinical Radiology. **59**: 390 (2004).

[5] Bocchi, L. *et al. Detection of single and clustered microcalcifications in mammograms using fractals models and neural networks.* Medical Engineering & Physics. **26**: 303 (2004).

[6] Boyle, Peter. *Breast cancer control: Signs of progress, but more work required.* The Breast. **14**: 429 (2005).

[7] Chan, Heang-Ping *et al. Computerised Classification of Malignant and Benign Microcalcifications on Mammograms: Texture Analysis Using an Artificial Neural Network.* Physics in Medicine and Biology. **42**: 549 (1997).

[8] Chan, Heang-Ping *et al. Improvement of Radiologists' Characterization of Mammographic Masses by Using Computer-aided Diagnosis: An ROC Study.* Radiology. **212**: 817 (1999).

[9] Cheng, H. D. *et al. Computer-aided detection and classification of microcalcifications in mammograms: a survey.* Pattern Recognition. **36**: 2967 (2003).

[10] Cheng, H. D. *et al. Approaches for automated detection and classification of masses in mammograms.* Pattern Recognition. **39**: 646 (2006).

[11] Cheung, Kan-Cheung *et al.* *First test pictures from X-ray diffraction enhanced imaging camera for high contrast medical imaging at SRS.* Nuclear Instruments and Methods in Physics Research A. **513**: 32 (2003).

[12] Ciatto, S. *et al.* *Computer-aided detection (CAD) of cancers detected on double reading by one reader only.* The Breast. *In Press* (2006).

[13] Chytyk, K., E. Kendall & C. Erickson. *Semi-automatic classification of breast images.* Poster (2004).

[14] Coldman, Andrew J. *et al.* *Organized Breast Cancer Screening Programs in Canada: Effect of Radiologist Reading Volumes on Outcomes.* Radiology. **238**: 809 (2006).

[15] Coulam, Craig M. *et al.* *The Physical Basis of Medical Imaging.* New York, New York: Appleton-Century-Crofts (1981).

[16] Doi, Kunio. *Overview on research and development of computer-aided diagnostic schemes.* Seminars in Ultrasound, CT, and MRI. **25**: 404 (2004).

[17] Domingos, Pedro & Michael Pazzani. *On the optimality of the simple Bayesian classifier under zero-one loss.* Machine Learning **29**: 103 (1997).

[18] Elmore, Joann G. *et al.* *Screening for Breast Cancer.* JAMA. **293**: 1245 (2005).

[19] Erickson, Carissa. *Automated Detection of Breast Cancer Using SAXS Data and Wavelet Features.* M. Sc. Thesis. University of Saskatchewan (2004).

[20] Giger, Maryellen L. *Computerized Analysis of Images in the Detection and Diagnosis of Breast Cancer.* Seminars in Ultrasound, CT and MRI. **25**: 411 (2004).

[21] Hand, D. J. & K. Yu. *Idiot's Bayes – not so stupid after all?.* International Statistical Review. **69**: 385 (2001).

[22] Haralick, R. M., K. Shanmugam & I. Dinstein. *Texture features for image classification.* IEEE Transactions on Systems, Man and Cybernetics. **3**: 610 (1973).

[23] Heath M., K.W. Bowyer, D. Kopans *et al*. *Current status of the Digital Database for Screening Mammography*. Digital Mammography. Kluwer Academic Publishers. pp. 457-460 (1998).

[24] Hernandez, Eugenio and Guido Weiss. *A First Course on Wavelets.* New York: CRC Press. 1996.

[25] Hubbard, Barbara Burke. *The World According to Wavelets.* Wellesley, Massachusetts: A K Peters. 1996.

[26] Jackson, John David. *Classical Electrodynamics*. 3rd Edition. New York: Wiley (1999).

[27] Jiang, Yulei *et al. Comparison of Independent Double Readings and Computer-Aided Diagnosis (CAD) for the Diagnosis of Breast Calcifications.* Academic Radiology. **13**: 84 (2006).

[28] Keyrilainen, Jani *et al. Visualisation of calcifications and thin collagen strands in human breast tissue specimens by the diffraction-enhanced imaging technique: a comparison with conventional mammography and histology.* European Journal of Radiology. **53**: 226 (2005).

[29] Lauria, A. *et al. The CALMA system: an artificial neural network method for detecting masses and microcalcifications in digitized mammograms.* Nuclear Instruments and Methods in Physics Research A. **518**: 391 (2004).

[30] Lee, San-Kan *et al. A computer-aided design mammography screening system for detection and classification of microcalcifications.* International Journal of Medical Informatics. **60**: 29 (2000).

[31] Lehman, Constance D. *Screening MRI for women at high risk for breast cancer.* Seminars in Ultrasound, CT and MRI. **27**: 333 (2006).

[32] Li, Lihua *et al. Digital mammography: Computer-assisted diagnosis method for mass detection with multiorientation and multiresolution wavelet transforms.* Academic Radiology. **4**: 724 (1997).

[33] Liu, Sheng, Charles F. Babbs and Edward J. Delp. *Multiresolution Detection of Spiculated Lesions in Digital Mammograms.* IEEE Transactions on Image Processing. **10**: 874 (2001).

[34] Mackovski. *Medical Imaging Systems.* New York: Prentice-Hall. (1983).

[35] Malich, Ansgar *et al. Reproducibility – an important factor determining the quality of computer aided detection (CAD) systems.* European Journal of Radiology. **36**: 170 (2000).

[36] Matsubara, T. *et al. Novel method for detecting mammographic architectural distortion based on concentration of mammary gland.* International Congress Series. **1268**: 867 (2004).

[37] Mayo Clinic. www.mayoclinic.com/health/breast-cancer/HQ00348. Accessed July 9, 2006.

[38] Mousa, Rafayah, Qutaishat Munib and Abdallah Moussa. *Breast cancer diagnosis system based on wavelet analysis and fuzzy-neural.* Expert Systems with Applications. **28**: 713 (2005).

[39] National Cancer Institute of Canada. *Canadian Cancer Statistics 2004.* Toronto, Canada. 2004.

[40] Nemoto, Mitsutaka *et al. Classifier ensemble for mammography CAD system combining feature selection with ensemble learning.* International Congress Series. **1281**: 1047 (2005).

[41] Olivo, A. *Towards the exploitation of phase effects in clinical synchrotron radiation radiology.* Nuclear Instruments and Methods in Physics Research A. **548**: 194 (2005).

[42] Oregon Health & Science University. www.ohsuhealth.com/images/gs/ei_0385.gif. Accessed July 9, 2006.

[43] Popli, M. *Pictorial Essay: Breast Calcification.* Ind J Radiol Imaging. **12**:1 (2002).

[44] Rickard, Mary *et al. Cancer detection and mammogram volume of radiologists in a population-based screening programme.* The Breast. **15**: 39 (2006).

[45] Sakurai, J. J. *Modern Quantum Mechanics.* Revised Edition. Reading, Massachusetts: Addison-Wesley Publishing (1994).

[46] Saskatchewan Cancer Agency. www.saskcancer.ca/Default.aspx?DN=80e4bb14-ed7d-4f79-852f-3316d42de414. Accessed October 6, 2006.

[47] Sentelle, S., C. Sentelle and M. A. Sutton. *Multiresolution-Based Segmentation of Calcifications for the Early Detection of Breast Cancer.* Real-Time Imaging. **8**: 237 (2002).

[48] Sonka, M., V. Hlavac & R. Boyle. *Image Processing, Analysis, and Machine Vision, 2nd Ed.* Pacific Grove, California: Brooks/Cole Publishing (1999).

[49] Strang, Gilbert and Truong Nguyen. *Wavelets and Filter Banks.* Wellesley, Massachusetts: Wellesley-Cambridge Press. 1996.

[50] Sturges, Herbert A. *The Choice of a Class Interval.* Journal of the American Statistical Association. **21**: 65 (1926).

[51] Suckling, J., *et al. The mammographic images analysis society digital mammogram database.* Exerpta Medica. International Congress Series **1069**: 375 (1994).

[52] Taylor *et al. Assessing the impact of CAD on the sensitivity and specificity of film readers.* Clinical Radiology. **59**:1099 (2004).

[53] Wright, Heather *et al. Magnetic resonance imaging as a diagnostic tool for breast cancer in premenopausal women.* The Americal Journal of Surgery. **190**: 572 (2005).

[54] Yang, Ying & Geoffrey I. Webb. *A comparative study of discretization methods for naïve-Bayes classifiers.* Proceedings, Pacific Rim Knowledge Acquisition Workshop, pp. 159-173 (2002).

APPENDIX A – PEFORMANCE OF DIFFERENT FEATURE TYPES

Individual classifiers tested in Section 5.3 were developed using different combinations of scalar feature types. This appendix lists the performance of the most sensitive classifiers possible for each wavelet basis using any one or any two of the four available types of scalar features: mean intensity, standard deviation of intensity, skewness of intensity, and kurtosis of intensity.

The training data used in this appendix are taken from the MIAS data base. There are 303 images, including 205 normal images and 98 images showing some type of abnormality.

Table A.1 - Relative performance of different wavelet bases using only mean intensity feature type

Wavelet Basis	Best Feature Triplet			Sensitivity (%)	Specificity (%)	Overall Classification Rate (%)
Haar	*M-v6*	*M-d8*	---	84.7	29.8	47.5
Db 2	*M-a4*	---	---	81.6	24.9	43.2
Db 4	*M-h2*	*M-h3*	*M-a7*	82.7	22.9	42.2
Db 8	*M-a6*	*M-a7*	---	90.8	16.1	40.3
Bior 1.5	*M-v1*	*M-d6*	*M-a8*	86.7	34.6	51.5
Bior 2.2	*M-a4*	---	---	83.7	25.4	44.2
Bior 2.8	*M-h3*	*M-h5*	*M-a6*	86.7	30.2	48.5
Bior 3.7	*M-v7*	---	---	89.8	11.2	36.6
Bior 4.4	*M-v2*	---	---	92.9	12.2	38.3
Bior 5.5	*M-d3*	*M-a6*	*M-a7*	85.7	29.3	47.5
Bior 6.8	*M-h3*	*M-h5*	*M-a6*	89.8	28.3	48.2

Table A.2 - Relative performance of different wavelet bases using only standard deviation feature type

Wavelet Basis	Best Feature Triplet			Sensitivity (%)	Specificity (%)	Overall Classification Rate (%)
Haar	*σ-d3*	*σ-d7*	*σ-a4*	82.7	42.4	55.4
Db 2	*σ-h3*	---	---	93.9	9.8	37.0
Db 4	*σ-h4*	---	---	94.9	10.7	38.0
Db 8	*σ-h7*	*σ-v7*	*σ-d7*	82.7	40.0	53.8
Bior 1.5	*σ-a8*	---	---	80.6	29.8	46.2
Bior 2.2	*σ-v7*	*σ-a8*	---	84.7	37.1	52.5
Bior 2.8	*σ-h5*	*σ-h8*	*σ-v8*	82.7	37.6	52.1
Bior 3.7	*σ-h7*	---	---	87.8	29.3	48.2
Bior 4.4	*σ-h5*	*σ-v1*	*σ-a7*	88.8	25.9	46.2
Bior 5.5	*σ-a8*	---	---	89.8	22.0	43.9
Bior 6.8	*σ-a8*	---	---	88.8	20.0	42.2

Table A.3 - Relative performance of different wavelet bases using only skewness feature type

Wavelet Basis	Best Feature Triplet			Sensitivity (%)	Specificity (%)	Overall Classification Rate (%)
Haar	*S-d5*	*S-d6*	*S-d7*	87.8	40.0	55.4
Db 2	*S-v5*	*S-a5*	*S-a6*	90.8	26.8	47.5
Db 4	*S-v5*	---	---	92.9	13.2	38.9
Db 8	*S-a4*	---	---	91.8	18.0	41.9
Bior 1.5	*S-a3*	---	---	93.9	14.6	40.3
Bior 2.2	*S-h4*	*S-v3*	---	91.8	14.1	39.3
Bior 2.8	*S-d5*	*S-a2*	*S-a7*	92.9	23.9	46.2
Bior 3.7	*S-d6*	*S-a2*	*S-a6*	92.9	22.9	45.5
Bior 4.4	*S-a7*	---	---	92.9	19.0	42.9
Bior 5.5	*S-h7*	---	---	89.8	19.0	41.9
Bior 6.8	*S-a3*	---	---	92.9	15.1	40.3

Table A.4 - Relative performance of different wavelet bases using only kurtosis feature type

Wavelet Basis	Best Feature Triplet			Sensitivity (%)	Specificity (%)	Overall Classification Rate (%)
Haar	*K-a6*	---	---	90.8	7.3	34.3
Db 2	*K-a3*	---	---	93.9	14.1	39.9
Db 4	*K-h5*	*K-h6*	*K-a3*	93.9	19.0	43.2
Db 8	*K-a1*	*K-a8*	---	90.8	28.8	48.8
Bior 1.5	*K-h3*	*K-a1*	*K-a8*	94.9	13.7	39.9
Bior 2.2	*K-a2*	---	---	94.9	14.1	40.3
Bior 2.8	*K-a7*	---	---	93.9	17.6	42.2
Bior 3.7	*K-a5*	*K-a7*	---	93.9	19.0	43.2
Bior 4.4	*K-h6*	*K-a2*	*K-a5*	93.9	16.1	41.3
Bior 5.5	*K-a1*	---	---	93.9	14.1	39.9
Bior 6.8	*K-h7*	*K-a5*	---	93.9	21.5	44.9

Table A.5 - Relative performance of different wavelet bases using mean and standard deviation feature types

Wavelet Basis	Best Feature Triplet			Sensitivity (%)	Specificity (%)	Overall Classification Rate (%)
Haar	*M-d8*	*σ-h2*	*σ-d7*	85.7	34.1	50.8
Db 2	*σ-h3*	---	---	93.9	9.8	37.0
Db 4	*σ-h4*	---	---	94.9	10.7	38.0
Db 8	*M-a6*	*M-a7*	---	90.8	16.1	40.3
Bior 1.5	*M-h3*	*M-a8*	*σ-h2*	88.8	25.9	46.2
Bior 2.2	*M-d6*	*σ-d8*	*σ-a8*	84.7	47.3	59.4
Bior 2.8	*M-h5*	*M-a6*	*σ-v8*	86.7	42.9	57.1
Bior 3.7	*M-v7*	---	---	89.8	11.2	36.6
Bior 4.4	*M-v2*	---	---	92.9	12.2	38.3
Bior 5.5	*σ-a8*	---	---	89.8	22.0	43.9
Bior 6.8	*M-h3*	*M-h5*	*M-a6*	88.8	28.3	47.9

Table A.6 - Relative performance of different wavelet bases using mean and skewness feature types

Wavelet Basis	Best Feature Triplet			Sensitivity (%)	Specificity (%)	Overall Classification Rate (%)
Haar	*M-h7*	*M-d8*	*S-h6*	88.8	35.6	52.8
Db 2	*S-v5*	*S-a5*	*S-a6*	90.8	27.3	47.9
Db 4	*S-v5*	---	---	92.9	13.2	38.9
Db 8	*M-a7*	*S-h5*	*S-a4*	91.8	38.0	55.4
Bior 1.5	*S-a3*	---	---	93.9	14.6	40.3
Bior 2.2	*S-h4*	*S-v3*	---	91.8	14.1	39.3
Bior 2.8	*S-d5*	*S-a2*	*S-a7*	92.9	23.9	46.2
Bior 3.7	*M-h6*	*M-v3*	*S-h3*	92.9	29.8	50.2
Bior 4.4	*M-v4*	*S-a1*	*S-a3*	92.9	45.9	61.1
Bior 5.5	*S-h7*	---	---	89.8	19.0	41.9
Bior 6.8	*S-a3*	---	---	92.9	15.1	40.3

Table A.7 - Relative performance of different wavelet bases using mean and kurtosis feature types

Wavelet Basis	Best Feature Triplet			Sensitivity (%)	Specificity (%)	Overall Classification Rate (%)
Haar	*K-a6*	---	---	90.8	7.3	34.3
Db 2	*K-a3*	---	---	93.9	14.1	39.9
Db 4	*K-h5*	---	---	94.9	9.3	37.0
Db 8	*M-a7*	*K-d4*	*K-a5*	91.8	45.4	60.4
Bior 1.5	*K-h3*	*K-a1*	*K-a8*	94.9	13.7	39.9
Bior 2.2	*M-h3*	*K-a1*	*K-a2*	94.9	22.9	46.2
Bior 2.8	*M-h4*	*K-v8*	*K-d5*	94.9	27.3	49.2
Bior 3.7	*M-a8*	*K-h6*	*K-a7*	95.9	19.0	43.9
Bior 4.4	*M-v8*	*K-a1*	*K-a2*	93.9	16.6	41.6
Bior 5.5	*K-a1*	---	---	93.9	14.1	39.9
Bior 6.8	*M-d3*	*K-a1*	*K-a7*	93.9	25.4	47.5

Table A.8 - Relative performance of different wavelet bases using standard deviation and skewness feature types

Wavelet Basis	Best Feature Triplet			Sensitivity (%)	Specificity (%)	Overall Classification Rate (%)
Haar	*σ-h3*	*S-d5*	*S-d6*	92.9	18.0	42.2
Db 2	*σ-h3*	---	---	93.9	9.8	37.0
Db 4	*σ-h4*	---	---	94.9	10.7	38.0
Db 8	*σ-d2*	*S-a5*	*S-a7*	93.9	51.2	65.0
Bior 1.5	*S-a3*	---	---	93.9	14.6	40.3
Bior 2.2	*S-h4*	*S-v3*	---	91.8	14.1	39.3
Bior 2.8	*S-d5*	*S-a2*	*S-a7*	92.9	23.9	46.2
Bior 3.7	*σ-h5*	*σ-h6*	*S-a7*	92.9	45.9	61.1
Bior 4.4	*S-a7*	---	---	92.9	19.0	42.9
Bior 5.5	*σ-a1*	*σ-a8*	*S-a1*	91.8	32.2	51.5
Bior 6.8	*S-a3*	---	---	92.9	15.1	40.3

Table A.9 - Relative performance of different wavelet bases using standard deviation and kurtosis feature types

Wavelet Basis	Best Feature Triplet			Sensitivity (%)	Specificity (%)	Overall Classification Rate (%)
Haar	*K-a6*	---	---	90.8	7.3	34.3
Db 2	*K-a3*	---	---	93.9	14.1	39.9
Db 4	*σ-h4*	---	---	94.9	10.7	38.0
Db 8	*σ-d2*	*K-h7*	*K-a6*	95.9	51.7	66.0
Bior 1.5	*K-h3*	*K-a1*	*K-a8*	94.9	13.7	39.9
Bior 2.2	*K-a2*	---	---	94.9	14.1	40.3
Bior 2.8	*K-a7*	---	---	93.9	17.6	42.2
Bior 3.7	*σ-h6*	*K-d6*	*K-a7*	93.9	22.0	45.2
Bior 4.4	*K-h6*	*K-a2*	*K-a5*	93.9	16.1	41.3
Bior 5.5	*K-a1*	---	---	93.9	14.1	39.9
Bior 6.8	*σ-d5*	*K-a7*	*K-a8*	95.9	16.1	41.9

Table A.10 - Relative performance of different wavelet bases using skewness and kurtosis feature types

Wavelet Basis	Best Feature Triplet			Sensitivity (%)	Specificity (%)	Overall Classification Rate (%)
Haar	*S-h1*	*K-a2*	*K-a8*	90.8	32.7	51.5
Db 2	*K-a3*	---	---	93.9	14.1	39.9
Db 4	*K-h5*	---	---	94.9	9.3	37.0
Db 8	*S-h3*	*S-a4*	*K-d4*	91.8	27.3	48.2
Bior 1.5	*K-h3*	*K-a1*	*K-a8*	94.9	13.7	39.9
Bior 2.2	*K-a2*	---	---	94.9	14.1	40.3
Bior 2.8	*S-d5*	*K-a4*	---	94.9	27.3	49.2
Bior 3.7	*S-d6*	*K-d8*	*K-a7*	93.9	23.9	46.5
Bior 4.4	*K-h6*	*K-a2*	*K-a5*	93.9	16.1	41.3
Bior 5.5	*K-a1*	---	---	93.9	14.1	39.9
Bior 6.8	*S-h3*	*K-h7*	*K-a3*	94.9	22.0	45.5

APPENDIX B – PEFORMANCE OF DIFFERENT FEATURE TYPES FOR DDSM DATA SET

Individual classifiers tested in Section 5.3 were developed using different combinations of scalar feature types. This appendix lists the performance of the most sensitive classifiers possible for each wavelet basis using any one or any two of the four available types of scalar features: mean intensity, standard deviation of intensity, skewness of intensity, and kurtosis of intensity.

The training data used in this appendix are taken from the DDSM data base. There are 1714 images, including 1065 normal images and 649 images showing some type of abnormality.

Table B.1 - Relative performance of different wavelet bases using only mean intensity feature type, DDSM data set

Wavelet Basis	Best Feature Triplet			Sensitivity (%)	Specificity (%)	Overall Classification Rate (%)
Haar	*M-v4*	---	---	85.7	16.9	42.9
Db 2	*M-h2*	*M-h3*	*M-a8*	87.5	47.2	62.5
Db 4	*M-h7*	*M-v3*	*M-v7*	93.5	34.1	56.6
Db 8	*M-h5*	*M-h7*	*M-a8*	91.7	24.3	49.8
Bior 1.5	*M-h7*	*M-v5*	*M-v6*	87.7	22.3	47.0
Bior 2.2	*M-h7*	*M-d5*	*M-a3*	89.1	27.2	50.6
Bior 2.8	*M-h8*	*M-a8*	---	87.7	19.3	45.2
Bior 3.7	*M-v5*	*M-d5*	*M-a8*	88.3	43.2	60.3
Bior 4.4	*M-v3*	---	---	91.4	11.8	41.9
Bior 5.5	*M-d4*	*M-d8*	*M-a5*	85.5	37.4	55.6
Bior 6.8	*M-a8*	---	---	93.2	9.3	41.1

Table B.2 - Relative performance of different wavelet bases using only standard deviation feature type, DDSM data set

Wavelet Basis	Best Feature Triplet			Sensitivity (%)	Specificity (%)	Overall Classification Rate (%)
Haar	*σ-h5*	*σ-d2*	*σ-d6*	93.5	25.1	51.0
Db 2	*σ-d5*	*σ-d6*	*σ-a8*	90.1	38.7	58.2
Db 4	*σ-h5*	*σ-d5*	*σ-a1*	95.1	23.2	50.4
Db 8	*σ-h2*	*σ-h8*	*σ-d3*	96.6	30.9	55.8
Bior 1.5	*σ-h4*	*σ-h7*	*σ-d2*	94.9	20.0	48.4
Bior 2.2	*σ-d1*	*σ-d6*	---	94.8	19.0	47.7
Bior 2.8	*σ-h1*	*σ-h6*	*σ-d4*	91.7	41.8	60.7
Bior 3.7	*σ-d1*	*σ-d6*	---	94.5	24.8	51.2
Bior 4.4	*σ-d4*	*σ-a4*	---	95.2	31.9	55.9
Bior 5.5	*σ-h4*	*σ-d2*	*σ-d6*	91.1	15.1	43.9
Bior 6.8	*σ-h1*	*σ-h6*	*σ-d4*	96.5	33.2	57.2

Table B.3 - Relative performance of different wavelet bases using only skewness feature type, DDSM data set

Wavelet Basis	Best Feature Triplet			Sensitivity (%)	Specificity (%)	Overall Classification Rate (%)
Haar	*S-v2*	*S-v8*	*S-d6*	85.8	21.7	46.0
Db 2	*S-h8*	*S-v1*	*S-d5*	93.4	33.7	56.3
Db 4	*S-d8*	---	---	92.3	12.7	42.8
Db 8	*S-v8*	*S-d3*	*S-a7*	94.0	39.2	59.9
Bior 1.5	*S-v5*	*S-v8*	---	92.0	11.9	42.2
Bior 2.2	*S-v5*	*S-v7*	*S-d4*	86.9	28.9	50.9
Bior 2.8	*S-v1*	*S-d5*	---	87.8	44.5	60.9
Bior 3.7	*S-v7*	*S-v8*	*S-d5*	91.2	43.1	61.3
Bior 4.4	*S-h6*	*S-v1*	*S-d7*	91.2	14.4	43.5
Bior 5.5	*S-h4*	*S-v2*	*S-a8*	92.3	39.1	59.2
Bior 6.8	*S-v1*	*S-v5*	*S-d2*	92.3	34.4	56.3

Table B.4 - Relative performance of different wavelet bases using only kurtosis feature type, DDSM data set

Wavelet Basis	Best Feature Triplet			Sensitivity (%)	Specificity (%)	Overall Classification Rate (%)
Haar	*K-h4*	*K-a2*	*K-a3*	95.1	22.1	49.7
Db 2	*K-v1*	*K-v4*	*K-d7*	90.9	15.6	44.1
Db 4	*K-h7*	*K-v3*	*K-v7*	89.8	29.0	52.0
Db 8	*K-h4*	*K-h5*	*K-a8*	95.2	27.9	53.4
Bior 1.5	*K-v7*	*K-d6*	*K-a8*	95.1	29.5	54.3
Bior 2.2	*K-v1*	*K-d6*	*K-a8*	92.3	28.2	52.5
Bior 2.8	*K-d1*	*K-a7*	*K-a8*	91.4	36.5	57.3
Bior 3.7	*K-v5*	*K-d7*	---	96.3	14.5	45.4
Bior 4.4	*K-v6*	*K-d4*	*K-d8*	89.8	27.6	51.2
Bior 5.5	*K-v2*	*K-v5*	*K-d3*	91.8	19.4	46.8
Bior 6.8	*K-v2*	*K-d7*	---	93.1	9.9	41.4

Table B.5 - Relative performance of different wavelet bases using mean and standard deviation feature types, DDSM data set

Wavelet Basis	Best Feature Triplet			Sensitivity (%)	Specificity (%)	Overall Classification Rate (%)
Haar	*M-d1*	*σ-h4*	*σ-a7*	98.3	35.0	59.0
Db 2	*M-h3*	*M-d8*	*σ-a6*	96.8	39.4	61.1
Db 4	*M-d3*	*σ-h3*	---	99.5	32.1	57.6
Db 8	*σ-h2*	*σ-h8*	*σ-d3*	96.6	31.2	56.0
Bior 1.5	*M-d1*	*M-d5*	*σ-h2*	97.1	39.3	61.2
Bior 2.2	*M-h3*	*σ-h2*	*σ-h5*	98.6	36.2	59.8
Bior 2.8	*M-h7*	*M-v1*	*σ-d1*	98.3	26.2	53.5
Bior 3.7	*M-d4*	*σ-h3*	*σ-v2*	96.8	35.9	58.9
Bior 4.4	*σ-d4*	*σ-a4*	---	95.2	31.9	55.9
Bior 5.5	*M-h7*	*M-v2*	*σ-d2*	95.8	22.3	50.1
Bior 6.8	*M-d4*	*σ-h3*	*σ-d4*	98.0	43.6	64.2

Table B.6 - Relative performance of different wavelet bases using mean and skewness feature types, DDSM data set

Wavelet Basis	Best Feature Triplet			Sensitivity (%)	Specificity (%)	Overall Classification Rate (%)
Haar	*M-h1*	*M-d1*	*S-h3*	99.2	36.6	60.3
Db 2	*M-h3*	*M-d8*	*S-h5*	97.4	42.7	63.4
Db 4	*M-h8*	*M-d1*	*S-h5*	95.2	20.8	49.0
Db 8	*M-h6*	*S-v8*	*S-d3*	97.5	40.4	62.0
Bior 1.5	*M-d4*	*S-h6*	---	96.9	38.8	60.8
Bior 2.2	*M-h5*	*M-v2*	*S-d2*	98.8	44.8	65.2
Bior 2.8	*M-d4*	*S-d2*	*S-a5*	92.9	46.9	64.4
Bior 3.7	*M-d4*	*S-h4*	*S-d4*	98.9	28.1	54.9
Bior 4.4	*M-h1*	*M-d4*	*S-d2*	96.1	43.0	63.1
Bior 5.5	*M-h6*	*M-d5*	*S-d2*	98.5	38.1	61.0
Bior 6.8	*M-v3*	*M-d4*	*S-d2*	98.0	39.0	61.3

Table B.7 - Relative performance of different wavelet bases using mean and kurtosis feature types, DDSM data set

Wavelet Basis	Best Feature Triplet			Sensitivity (%)	Specificity (%)	Overall Classification Rate (%)
Haar	*M-d5*	*K-d2*	---	98.0	40.5	62.3
Db 2	*M-h3*	*M-d8*	*K-d3*	95.7	41.0	61.7
Db 4	*M-v7*	*M-d7*	*K-v8*	96.3	18.8	48.1
Db 8	*M-h4*	*M-a7*	*K-a8*	95.7	30.0	54.8
Bior 1.5	*M-d3*	*K-h1*	*K-h4*	97.1	39.0	61.0
Bior 2.2	*M-d6*	*K-v4*	*K-d1*	97.4	38.2	60.6
Bior 2.8	*M-d3*	*K-h7*	*K-d7*	94.9	34.6	57.5
Bior 3.7	*M-h1*	*M-d3*	*K-d5*	97.2	44.4	64.4
Bior 4.4	*M-d3*	*K-h5*	---	94.5	27.2	52.7
Bior 5.5	*M-h5*	*M-d4*	*K-d2*	95.4	35.3	58.1
Bior 6.8	*M-d4*	*K-d5*	*K-a6*	95.4	42.7	62.7

Table B.8 - Relative performance of different wavelet bases using standard deviation and skewness feature types, DDSM data set

Wavelet Basis	Best Feature Triplet			Sensitivity (%)	Specificity (%)	Overall Classification Rate (%)
Haar	σ-h6	σ-d4	S-h5	93.7	27.9	52.8
Db 2	σ-d3	S-h8	S-d5	94.0	33.0	56.1
Db 4	σ-h5	σ-d5	σ-a1	95.1	23.2	50.4
Db 8	σ-h2	σ-h8	S-a7	97.5	36.0	59.3
Bior 1.5	σ-h4	σ-h7	σ-d2	94.9	20.0	48.4
Bior 2.2	σ-h1	σ-h8	S-v1	96.3	25.4	52.2
Bior 2.8	σ-h1	σ-h8	S-v1	95.8	27.7	53.5
Bior 3.7	σ-d4	S-h1	---	98.2	21.2	50.4
Bior 4.4	σ-d4	σ-a4	---	95.2	31.9	55.9
Bior 5.5	σ-h3	S-h2	S-h3	94.8	40.4	61.0
Bior 6.8	σ-h1	σ-h6	σ-d4	96.5	33.2	57.2

Table B.9 - Relative performance of different wavelet bases using standard deviation and kurtosis feature types, DDSM data set

Wavelet Basis	Best Feature Triplet			Sensitivity (%)	Specificity (%)	Overall Classification Rate (%)
Haar	*σ-v8*	*K-h5*	*K-v7*	96.1	25.4	52.2
Db 2	*σ-d4*	*K-h5*	*K-a8*	95.5	36.7	59.0
Db 4	*σ-h5*	*σ-d5*	*σ-a1*	95.1	23.2	50.4
Db 8	*σ-h2*	*σ-h8*	*σ-d3*	96.6	31.2	56.0
Bior 1.5	*σ-v2*	*σ-d2*	*K-v7*	98.6	18.3	48.7
Bior 2.2	*σ-h1*	*σ-d5*	*K-a7*	94.8	35.3	57.8
Bior 2.8	*σ-h6*	*σ-d1*	*K-a2*	95.4	30.5	55.1
Bior 3.7	*σ-h1*	*K-h2*	*K-v5*	97.5	17.5	47.8
Bior 4.4	*σ-h1*	*σ-h5*	*K-a8*	97.5	40.0	61.8
Bior 5.5	*σ-d2*	*K-h3*	*K-v7*	95.5	21.7	49.6
Bior 6.8	*σ-h1*	*σ-h6*	*σ-d4*	96.5	33.2	57.2

Table B.10 - Relative performance of different wavelet bases using skewness and kurtosis feature types, DDSM data set

Wavelet Basis	Best Feature Triplet			Sensitivity (%)	Specificity (%)	Overall Classification Rate (%)
Haar	*K-h4*	*K-a2*	*K-a3*	95.1	22.1	49.7
Db 2	*S-h8*	*S-v1*	*S-d5*	93.4	33.7	56.3
Db 4	*S-h7*	*S-v7*	*K-v7*	92.6	21.8	48.6
Db 8	*S-d5*	*K-a2*	*K-a8*	97.2	36.2	59.3
Bior 1.5	*K-v7*	*K-d6*	*K-a8*	95.1	29.5	54.3
Bior 2.2	*K-v1*	*K-d6*	*K-a8*	92.3	28.2	52.5
Bior 2.8	*S-a8*	*K-d5*	---	92.9	37.1	58.2
Bior 3.7	*K-v5*	*K-d7*	---	96.3	14.5	45.4
Bior 4.4	*S-d5*	*S-a8*	*K-h2*	92.9	50.7	66.7
Bior 5.5	*S-h4*	*K-a8*	---	95.7	38.5	60.2
Bior 6.8	*S-v1*	*S-d5*	*K-a7*	93.2	42.0	61.4