Mammographic Classification With Multiple Global Features

by

Richard J. Lee

A thesis submitted to the Faculty of Graduate Studies
in partial fulfilment of the requirements for
the degree of
Doctor of Philosophy

Department of Physics
University of Manitoba
Winnipeg, Manitoba

National Library of Canada

Acquisitions and Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada

Acquisitions et services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

0-612-62650-4

THE UNIVERSITY OF MANITOBA

FACULTY OF GRADUATE STUDIES

*****

COPYRIGHT PERMISSION PAGE

Mammographic Classification with Multiple Global Features

BY

Richard J. Lee

A Thesis/Practicum submitted to the Faculty of Graduate Studies of The University

of Manitoba in partial fulfillment of the requirements of the degree

of

Doctor of Philosophy

Acknowledgements

I would like to express my sincere thanks to all to have stood by me during the course of this degree. I could not have completed this work without your guidance, encouragement and support throughout both the difficult and not so difficult times. I would like to make specific mention of my supervisors, Roger and Norm and the rest of my committee; my parents and sisters for their support; Boyd, Barry and Steve for the many hours of useful discussion and, of course, Marilyn who kept minor inconveniences from becoming major disasters.

Abstract

In this thesis several global mammographic features were examined for their ability to classify the mammograms into

1. classes based on the proportion of dense tissue
2. normal/abnormal groups.

A set of 240 digitised mammograms was obtained from the Digital Database for Screening Mammography from the University of South Florida. The database was composed of mammograms that were digitized using one of three high resolution x-ray digitisers. It was necessary for the images to be corrected for three systematic differences between the x-ray digitisers: the resolution, the slope of the calibration curve and non-linearities in the calibration curve. A simple correction was also made for differences in the mammographic technique by adjusting the histogram of the breast shadow.

The breast shadow was then segmented using a semi-automatic procedure and several mammographic properties were extracted: global moments of the histogram, the average local moments calculated for ~3x3 mm^2 regions covering the breast shadow, subregions of the global histogram, multifractal dimensions and the texture energy, entropy and inertia calculated for the wavelet transform of the image.

The classification accuracy, when considering the density grades, was consistently $\sim 40\%$ correct and independent of the properties used in the classifier. When classifying into normal/abnormal groups, the regional moments, histogram sub-regions and the multifractal dimensions all had approximately the same performance at $\sim 60\%$ correctly classified cases, while the global moments classified $\sim 70\%$ of the cases correctly. The texture energy, entropy and inertia also had approximately the same performance but at $\sim$ 80–85% correct. In addition, the classifiers exhibited no significant change in classification performance for variations in age for any of the examined properties with $p = 0.001$.

The texture features resulted in the highest classification accuracy. The results may show some residual dependence on the x-ray digitiser but the small sample size precluded any definitive conclusions regarding the influence of the scanners. Overall, a classifier using six texture inertia features exhibited the best overall classification accuracy with minimal age dependence.

Contents

Appendices

List of Figures

List of Tables

Chapter 1

Introduction

The prevalence of breast cancer in industrialised nations and its rapid growth in developing nations has it poised as one of the most common malignancies world-wide. In Canada alone, its incidence has risen steadily over the last three decades to an estimated 105 cases per 100 000 [NCSC, 1999]. Simultaneously, the decrease in mortality rates over the same period may be the best testament to the effectiveness of mammographic screening. A secondary effect of the widespread adoption of screening programs is a tremendous growth in the sheer volume of screening mammograms that must be evaluated. This, combined with the low contrast inherent in soft tissue x-ray imaging contributes to making mammographic interpretation difficult and time consuming. There are many options, both emerging and well established, that are intended to augment the specificity of screening mammography, such as positron emission tomography (PET), magnetic resonance imaging (MRI), ultrasound, etc. [Adler and Wahl, 1995, Säbel and Aichinger, 1996, Jones, 1992, Reynolds, 1999]. These modalities would enable better differentiation between malignant disease and a benign condition. However there are few alternatives to mammography itself, although there have been a number of significant developments in this direction such as the use of synchrotron radiation [Burattini et al., 1995], phase imaging [Ingal et al., 1998] and digital mammography [Newman, 1999, Yaffe and Rowlands, 1997, Schmidt and Nishikawa, 1995, Simonetti et al., 1998]. Both phase imaging and the use of synchrotron radiation for diagnostic imaging are relatively new developments and attempt to reduce patient dose while improving image quality by using a nearly monochromatic x-ray source. Unfortunately, both modalities are quite far from clinical use.

A more relevant development is in the area of digital mammography. The approach is more conventional, replacing the film with a solid state detector. Many types of detectors have been used in different systems. For example, some employ a cartridge of amorphous selenium similar to that used in the, now obsolete, Xeromammography units while others use two dimensional arrays of CCD elements or a line of CCDs scanned across the breast. As yet, these systems are not currently in common use and the most sensitive modality at present continues to be conventional screen/film mammography. However, the importance of digital systems will only increase in the future. In addition, the need for both computer manipulation of the images and computer assisted diagnosis will grow with it.

Even without the widespread adoption of fully digital systems, attempts have been made to reduce some of the volume of screening mammograms that require interpretation, using the automated systems as a "second reader". In particular, a group from the University of Chicago has developed a system that has been undergoing clinical trials [Nishikawa et al., 1996]. This system attempts to detect both masses and microcalcification clusters through a fairly complex series of steps. The system requires a large number of parameters, such as threshold levels, and the developers have attempted to adjust these parameters automatically in order to optimise its performance [Anastasio et al., 1998].

Digital mammography will also permit extensive computer processing of mammograms. In anticipation of widespread introduction of such systems, this thesis examines several tasks that could be incorporated into a screening procedure. The primary purpose is to identify mammographic features of interest either for diagnostic purposes or as indicators of risk. For diagnostic applications, the intention is to simply flag a possibly abnormal mammogram for special consideration rather than to attempt to isolate the region where the abnormality is located. While it may be possible to extend the procedures used in this thesis to encompass the more difficult task of abnormality identification, this is beyond the scope of the present work.

A secondary goal is the identification of mammographic properties which may be useful in assessing risk. In general, there are three categories which influence the development of breast cancer,

1. heredity
2. hormonal and reproductive factors
3. environment/lifestyle.

The interactions of these contributing factors are complicated and it is difficult to quantify the factors as well as their interactions in order to assess the risk. While there do exist biological markers which are indicative of breast cancer risk, such as BRCA1 and BRCA2, there are a large number of breast cancer cases where either or both of these genes are normal [Weber, 1998]. Therefore, the assessment of risk can be difficult, in general. However, if a property can be found that is indicative of breast cancer risk, and can be extracted from a mammogram, it would no longer be necessary to attempt to quantify these qualitative factors and their interactions.

A quantifiable measure of the risk using mammographic features could also be useful for applications in the following areas:

Risk assessment A patient categorised as being in a high risk group would likely have a different course of treatment than one at low risk. The different treatment can include a shorter time between mammography screens, preventative drug treatments or even lifestyle changes.

Evaluation of prevention protocols A simple and reliable method to evaluate the effectiveness of an experimental preventative therapy would enable new therapies to be brought into practice much faster than is currently possible. Presently, the effectiveness of a protocol cannot be evaluated without lengthy trials involving large numbers of patients. Indeed, there is already work ongoing in this direction by Boyd *et al.* who examined the effects of dietary fat on mammographic features [Boyd et al., 1997], as well as by Ursin *et al.* [Ursin et al., 1996] and Atkinson *et al.* [Atkinson et al., 1999] both of whom are investigating the impact of Tamoxifen on mammographic features. Of course, the adoption of such a test would require irrefutable evidence to link the factor with breast cancer risk which would not be possible without extensive clinical trials. However, even prior to reaching this stage of research, the test would still be useful for ranking candidate preventative therapies for the more difficult and expensive clinical trials.

In addition, a feature characteristic of risk would also assist in research into the contribution and interaction between the factors listed above and breast cancer.

Clearly, there would be tangible benefits from the development of a mechanism for the reliable classification of mammograms for either assisting in the diagnosis of screening mammograms or in the area of risk prediction.

1.1 Breast Cancer Risk Evaluation

Currently, there are two primary methods used to assess the risk of developing breast cancer from mammographic features:

1. Wolfe grades

2. The fraction of dense parenchymal tissue.

Wolfe grades were introduced by J. Wolfe [Wolfe et al., 1986, Wolfe, 1976a] and is one of the earliest mammographic classification schemes that reflects breast cancer risk. Wolfe grades classify mammograms into four grades with increasing cancer risk. The lowest risk was assigned to mammograms with little parenchymal tissue and the highest for extensive and atypical growth of the duct epithelium (atypical hyperplasia). The two remaining grades were assigned for the amount and appearance of ducts. The relationship between the Wolfe grade and breast cancer risk has already been extensively studied by many others. As a brief overview see [Brisson et al., 1982a, Brisson et al., 1982b, Boyd et al., 1982, Tabár and Dean, 1982, Brisson et al., 1984, Goodwin and Boyd, 1988, Arthur et al., 1990, Salminen et al., 1998][1]. In fact, there has even been some work on correlating Wolfe grades to the histological classification of biopsy samples, with respect to breast cancer [Urbanski et al., 1988].

[1] The work done in this area is extensive and the list is by no means complete.

The primary elements that distinguish the divisions between the Wolfe grades are the mammographic density and the appearance of the duct structure and there has been some work on quantifying both characteristics. For instance, automatic quantification of the duct patterns has been used by Shadagopan [Shadagopan et al., 1982] who employed morphological features, such as the shape and spatial frequency, to distinguish actual ducts from other mammographic features. Alternatively, Wolfe [Wolfe et al., 1986] and Saftlas [Saftlas et al., 1991] a quantitative measure of the mammographic density using a planimeter was compared to the Wolfe grades for correlation to breast cancer risk. Their results indicate that the mammographic density (or simply density) is a more significant risk factor than the appearance of the duct structure.

Since then, many have followed their lead in concentrating on the mammographic density in preference to the duct structure but different groups have used varying numbers of density grades; from as few as four as in the American College of Radiology BiRADs guidelines [ACR, 1993] to a continuous scale from 0–100 as in [Boone et al., 1998]. The range of density classes is a result of a compromise between the need for distinguishing subtle differences and minimization of inter- and intra-observer variations, since the only available standard for the mammographic density is the classification according to an experienced observer. Several large and very significant studies involving 708 cases were reported by Boyd *et al.* [Boyd et al., 1995, Byng et al., 1997] who found a strong correlation between breast cancer risk and a six category classification scheme, SCC, where the density classes were divided as: None, (0,10%), [10,25%), [25,50%), [50,75%) and [75,100%] of dense tissue. The correlation of the risk to three other mammographic features (described in detail below) was also considered by Boyd.

1.1.1 Automated Methods

The difficulty with the conventional approach to assigning either a Wolfe grade or a density class is the subjective nature of the assignment which can have quite a low inter-observer correlation. In fact, Boyd [Boyd et al., 1982] found a 70% agreement between two radiologists when assigning Wolfe grades and 60% when classifying the extent of dysplasia.

There have been several approaches to using some automated characteristic to reduce this subjectivity. For example, Boone [Boone et al., 1998] created a continuous scale (0–100) to categorise breast density and relied on the rank of a set of mammograms when ordered according to the proportion of breast density. This ranking was then used to generate the standard rather than depending upon someone's judgement as to the category in which the mammogram should belong when examining them individually. Once the mammograms were ordered, six unique features that were mostly related to some form of fractal dimension, were extracted and used to form a linear "breast density index" (BDI) so that a numerical value could be generated without further need of a human observer. In addition, Tahoces *et al.* [Tahoces et al., 1995] was able to achieve reasonable classification of Wolfe grades (70–90% correct with the majority of misclassifications offset by only one class) by performing some image enhancement using unsharp mask filtering before extracting some simple textures. The texture features included the RMS power and the mean of the limits of the grey scale range in a region of interest (ROI) for a computer selected ROI.

Alternatively, Karssemeijer *et al.* [Karssemeijer, 1998] used some simple features from the grey level distribution of the lateral views from a set of screening mammograms to classify four density grades ($< 5\%$, 5–25%, 25–75% and $> 75\%$) with $\sim 80\%$ accuracy. The features which were used were quite straightforward, including: the standard deviation, skewness, the difference in means between the grey level distribution in the breast tissue and pectoral muscle and the integrated difference between the two distributions. Each property was found as a function of the distance to the skin surface. Even with compression, the breast thickness changes rapidly close to the surface so that this functional dependence on the distance to the skin can make a significant impact on the results.

Some of the most extensive work has been done by Byng *et al.* who developed three classification properties. Each was assessed for their classification ability in categorizing mammograms into a six grade density classification scheme (SCC). The first method, described in [Byng et al., 1994], used a semi-automated procedure where the user was required to select a pixel that was used as a threshold grey level to distinguish the breast tissue from the background. This also allowed the computer to automatically calculate the area of the actual breast shadow. Then the user selected a second threshold that was representative of the parenchymal tissue. The percent density was calculated from the fraction of the breast tissue above this second threshold normalised by the total segmented image size. Since the exposure conditions can vary from film to film, a fixed threshold to delineate the tissue types could not be employed. While the approach was still subjective, the resulting mammogram classification did not have nearly as much inter-observer variation as for the conventional approach to density classification. Byng was able to achieve high inter-observer correlation even with novice users with minimal training (typically with a Spearman correlation coefficient $\gtrsim 0.9$).

In a recent study by Huo [Huo et al., 2000] the difficulty in relating mammographic density and risk was bypassed by considering a well established risk factor with a strong biological basis, the mutation in two genes, BRCA1 and BRCA2. Again, several features were examined — many extracted from the histogram characteristics such as the average, minimum and maximum grey level in a region of interest as well as the grey level that delineated a given fraction of the total number of pixels in the ROI and several other conventional texture characteristics. When a receiver operating characteristic (ROC) curve was generated for the various features, the area beneath the curve varied from 0.53–0.87 with an average value of 0.72. When four features were considered simultaneously, the area increased to 0.91 which indicates a considerable increase in the probability of correctly identifying the patients with the genetic mutations.

1.1.2 Fractal Techniques

Byng *et al.* [Byng et al., 1996a, Byng et al., 1997] has also examined two additional features, the regional skewness and a fractal dimension, for a correlation with breast cancer risk. For the regional skewness, the breast tissue was segmented and the segmented region tessellated into small ROI's each 3.12x3.12 mm^2 and the skewness calculated for each ROI prior to averaging all the values to obtain a single overall result[2]. The second property, a fractal dimension, was found by treating the image as a surface in a three dimensional abstract space where the height was proportional to the grey level for the corresponding pixel location. Then the behaviour of the

[2] See Section 2.1.1 and Section 3.3.1 or [Byng et al., 1996a] for further details.

surface area was characterised at different resolutions by finding the slope for the relationship between the log(area) as a function of the log(resolution)[3]. All three properties showed good correlation to the density classification using SCC and were used to find the relative risk in the studies listed above. In principle, all of them represent a continuous scale for the breast density. While the change in relative risk for variations in the regional skewness, fractal dimension and (semi-automated) percent density were not as dramatic as for the SCC approach, the results found by Byng were still significant after adjustment for other well accepted risk factors such as family history or reproductive factors.

A similar method for the calculation of a fractal dimension was used by Caldwell *et al.* [Caldwell et al., 1990] for the purpose of distinguishing Wolfe grades. Caldwell encountered only limited success. In particular, he found that distinguishing the high risk (upper two grades) from the low risk (lower two grades) was more reliable than its ability in distinguishing divisions within either the upper or lower two grades. Another conventional fractal dimension, a box-counting dimension had been successfully used by Velanovich [Velanovich, 1996] to characterise the boundary of a suspicious mass and identify benign from malignant masses.

However, the outcome when using fractal dimensions in medical applications has been varied. For example, Karssemeijer [Karssemeijer, 1998] was not able to reproduce the results of Caldwell [Caldwell et al., 1990] and Byng [Byng et al., 1996a] when using a fractal dimension. It should be noted that Karssemeijer used films obtained over a long period, 1983–1994, and the quality of the films varied significantly during that time. Their results indicated that the techniques they employed

[3] For those unfamiliar with fractals, a description of the concepts for these studies along with the concepts used for the multifractal dimensions used in this thesis can be found in Appendix A.

which were successful in classifying the density performed better when the images were constrained to the mammograms that were obtained more recently.

Veenland *et al.* [Veenland et al., 1996], examined many different fractal dimensions for simulated organs. A comparison was made between the "true" fractal dimension and the fractal dimension which would be obtained from film with known characteristics. Veenland suggested that fractal dimensions are very sensitive to variations in the modulation transfer function (MTF) and noise characteristics of the film and exposure conditions. On the other hand, Caldwell [Caldwell et al., 1990] and Byng [Byng et al., 1996a] both have examined the effects of changes in the Hurter and Driffield (HD) curve typical of their mammography system with little impact on their results. This apparent discrepancy may have been due to

- The extensive QA procedures for modern mammography systems. Some of the films used by Karssemeijer was sufficiently old that the film quality was considerably different compared to their more recent samples.

- Sufficient difference in the characteristics of the fractal dimensions for malignant and normal tissue that the influence of the MTF and noise in the film was not able to mask the differences. Veenland's conclusions were drawn from the change the film makes to the "actual" fractal dimension while Byng and Caldwell used genuine mammograms. Therefore, the effect that the film makes to the fractal dimension of the parenchymal tissue may be smaller than the difference in the fractal dimensions of malignant and normal tissue.

- Differences in the characteristics of the fractal dimensions themselves, i.e. the fractal dimension that Caldwell and Byng used may not have been as susceptible to the MTF and noise of the film[4]. Different methods of calculating fractal dimensions examine different characteristics of the image.

1.1.3 Therapy Evaluation

While the relative risk for those with a BRCA1 or BRCA2 mutation is much greater than for those without the abnormality, the prevalence of cases with the mutations is relatively low [Weber, 1998]. If the mammographic features that were identified are correlated strictly to the gene mutations they may not be useful for applications such as the evaluation of preventative therapies. However, the work of Boyd *et al.* [Boyd et al., 1997] on mammographic features and dietary fat, Ursin *et al.* [Ursin et al., 1996] and Atkinson *et al.* [Atkinson et al., 1999] on mammographic features and Tamoxifen suggest that this may not be the case. In particular, in [Boyd et al., 1997] it was found that the subjects on a low-fat, high-carbohydrate diet showed a reduction in the amount of mammographic density. Further, the reduction in density was greater than could be accredited to weight loss alone. Ursin, on the other hand, [Ursin et al., 1996], examined the changes in the densities in the contralateral breast of patients diagnosed with breast cancer. It was found that patients treated with Tamoxifen (with and without radiation therapy) exhibited a reduction in the mammographic density compared to patients receiving chemotherapy and/or radiation therapy. Similarly, Atkinson, [Atkinson et al., 1999], found a statistically significant change ($p = 0.0001$) in the Wolfe grade classification (toward the lower risk grades) of patients undergoing treatment with Tamoxifen.

[4] Veenland did not consider the specific fractal dimension used in [Caldwell et al., 1990] and [Byng et al., 1996a].

1.1.4 Conclusions

In the studies discussed previously, the extracted features were quite varied but they were selected with the intent of producing a more reproducible and less subjective feature than the current approach to assigning the mammographic density and they all characterise essentially the same mammographic property. Boone created his breast density index expressly to have a very high correlation to the mammogram ranking based on density. Similarly, Byng verified that there was a reliable correlation between their features (percent density, regional skewness and fractal dimension) and the breast cancer risk but the selection of these particular properties was made with consideration of their relationship to the mammographic density. For example, the regional skewness was specifically selected by Byng *et al.* since a mammogram with predominately dense tissue would have a histogram with a proportionally greater fraction of pixels with higher grey level values thus producing a negative skewness. As well, for a predominately dense breast, the contrast for large portions of the mammogram will be lower than for a breast consisting primarily of fatty tissue. Therefore, if the image of the dense mammogram were viewed as a surface in an abstract three dimensional space, where the pixel intensities represent the third dimension, it would appear smoother than a mammogram with a lower mammographic density, thus producing a lower fractal dimension. The mammographic density and the features discussed above, which were related to the density, represent a basic examination of the information that was contained even within the global characteristics of the mammogram.

For this thesis we considered several systematic and more comprehensive analyses of the mammogram. Since the mammographic density is widely accepted as a risk factor we examined the relationship of our extracted features with the density. Features were extracted that were less obviously connected to the density and features were identified, both individually and in combination, that can be used to classify the mammographic density. Fortunately, the database of images which were used for this project contained the density classification according to the BiRADs guidelines. However, the data were not ideally suited for either determination of density classes or risk assessment[5]. In particular, the number of cases in each density grade was quite variable and the details for a number of other important risk factors were not available for the patients, such as age of menarche, nulliparity, etc. The complete evaluation of either of these would require a full study in itself but the identification of the important properties and some of the basic procedures which are necessary for such a study are provided in the present work. We also selected the set of features which were most closely correlated with the appearance of breast cancer rather than the density classification in order to identify an independent risk factor or a property indicative of breast cancer itself.

1.2 Computer Aided Diagnosis

This section provides background on several image properties that were used in this work. In the following section some of the important studies that employed useful conventional techniques are described. The method typically involves the calculation of various textures which quantify characteristics of the image such as the contrast or the homogeneity. In general, a large number of textures was found and a subset that was most useful for the given problem was selected using a method such as

[5] See Chapter 3.

stepwise refinement or a genetic algorithm. Regardless of the method used for the feature selection, a technique such as linear discriminant analysis was used in order to evaluate the selected set of features. Several of the studies relevant in mammography which used a genetic algorithm are described in Section 1.2.2.

Finally, approaches for investigating the scale which was the most significant for a particular problem was discussed in Section 1.2.3. If too large or too small a scale is used to examine the texture, the structure relevant to the problem may not be reflected in the extracted features. Therefore, examining many scales to identify the most useful is important.

1.2.1 Texture Methods

Many textures can be calculated from what is known as a spatial grey level dependence (SGLD) matrix. The SGLD matrix is a two dimensional array which is a function of two variables, d and θ. Each entry in the matrix, (i, j), contains the frequency of occurrence for a pair of pixels with grey levels i and j separated by a distance d and with an orientation characterised by an angle θ. A similar array, a spatial grey level *difference* matrix has an extra parameter, the difference in the grey levels for the pixel pair. In other words it is the SGLD matrix for only those pixel pairs with a specific value for $\mid i - j \mid$. The majority of studies described in this chapter utilised textures calculated using either of these matrices. However, while many textures that appear in the literature are quite common, many more have been developed for the specific purpose of their respective studies. Therefore, a complete list of textures used in each study as well as an explanation of how to calculate the individual textures (except for the ones used in this thesis) was left to the literature.

Two of the earliest works using texture measures in mammography are Taylor *et al.* [Taylor et al., 1990] and Magnin *et al.* [Magnin et al., 1986]. Taylor used various texture measures such as the skewness, fractal dimension and Laws energy to identify the "easy to interpret" mammograms (fatty) from "difficult" ones (dense) in addition to identifying Wolfe grades. Magnin attempted to distinguish Wolfe grades using the examination of several common texture features (eg. energy, inertia and others) that were derived from a SGLD matrix for horizonal pixel pairs separated by 10 pixels or features extracted from the grey level difference matrix.

Chan *et al.* [Chan et al., 1995] used an approach similar to that of Magnin to classify tissue regions into abnormal masses and normal tissue. After preprocessing to remove the effects of the background on the texture values, eight texture measures were calculated (eg. energy, entropy, correlation, inverse difference moment, etc.). Each was derived from the SGLD matrix for four different directions (θ) and several pixel separation distances (d). This approach generated a pool of texture features from which a subset that best distinguish the normal tissue from abnormal masses can be extracted. Chan *et al.* also used a stepwise refinement procedure for the feature selection and linear discriminant analysis for the feature evaluation.

Linear discriminant analysis is a standard statistical procedure to create a function that is linear in the variables and minimizes the number of incorrectly classified cases. The procedure can be viewed as the projection of the feature vectors onto a one dimensional axis and the linear discriminant procedure changes the orientation of the axis to maximise the difference between the classes in the sample. The particular axis orientation or linear combination of variables that is found is generally referred to as the linear discriminant function.

The feature selection process, stepwise refinement, requires the selection of two thresholds and a statistic of significance. Only if the inclusion of a variable changes the statistic by an amount greater than the inclusion threshold is the variable used in the discriminant function. Once all the variables have been tested for inclusion, each selected variable is then tested for removal. If the removal of a variable changes the statistic by less than the second threshold then it is removed. The procedure is repeated until the set of features is stable. Both the stepwise refinement and linear discriminant analysis are conventional and widely used approaches for feature selection and classification.

Several general aspects on the use of textures for medical applications can also be found in [Veenland et al., 1998]. Veenland investigated the effect of the MTF and noise characteristics common in general anatomical radiographs on a large number of texture measures, features from the power spectrum and morphological properties. In addition, Veenland *et al.* [Veenland et al., 1998] also studied the effect of the MTF and noise on several fractal dimensions.

1.2.2 Genetic Algorithmic Methods

An application that has received much attention in computer aided diagnosis (CAD) of mammograms is in differentiating benign and malignant microcalcification clusters. For example, the approach of Chan *et al.* [Chan et al., 1998] for this problem involved extracting textures as well as certain morphological characteristics of the microcalcification clusters and selecting a subset of features using stepwise refinement or a random optimization technique: a genetic algorithm.

A genetic algorithm is a method of randomly exploring a large feature space. The overall method is straightforward but there are many subtle variations to the technique and a general discussion of the procedure can be found in Appendix B, while the variations used specifically for the program that implemented the genetic algorithmic approach for this thesis, **ga_ors**, can be found in Section 2.3. The approach to the exploration was inspired by genetics and evolution so that the ideas are couched in those terms. One common approach utilises a "chromosome", represented by a string of bits, and an encoding scheme, so that each bit position represents a different feature. Initially, a large number of chromosomes (a "population") is created with random features selected and each member of the population evaluated relative to a fitness function. In [Chan et al., 1998] the fitness function was related to the area beneath the ROC curve for their test data[6]. Next, a new generation was formed by "reproducing" the chromosomes and the probability of a particular chromosome taking part in reproduction is determined by a function of its fitness. There are two common methods of reproduction, the first is a crossover technique which uses two chromosomes selected at random and a part of each chromosome is interchanged with the other. The second method involves a random alteration of the genes in each chromosome (mutation). Ideally, after a fixed number of generations the chromosomes have evolved to a small set of the best features.

Chan [Chan et al., 1998] found the genetic algorithm selected a set of features that was consistently better than those found using the conventional stepwise refinement technique. As well, an earlier study performed by Sahiner *et al.* [Sahiner et al., 1996] made a comparison between the performance using features selected using stepwise refinement, a genetic algorithm and a neural network. When they calculated the area beneath the ROC curve for their test data, the genetic algorithm outperformed both

[6] Recall the area beneath a ROC curve is proportional to the probability of a correct classification.

alternative techniques.

1.2.3 Wavelet Methods

A wavelet transform is an integral transform, like a Fourier transform, that has a basis with specific properties[7]. It can be viewed as the result of a signal after filtration through a series of high and low pass filters that are arranged in a specific order. The transformed signal can then be divided into several regions that contains either

1. a representation of the signal at different resolutions
2. components that are lost when the signal is examined at the different resolutions (multi-resolution analysis).

The applications for a wavelet transform are growing rapidly and one widespread use has been in the area of image enhancement. The enhancement tends to work particularly well for high frequency regions such as those that contain edges [Giger and MacMahon, 1996]. For example, within mammographic applications they have been used to increase the conspicuity of objects that can be difficult to locate, such as microcalcification clusters. The transform and variations of the procedure has also been used for in the automatic identification of microcalcification clusters as in [Zhang et al., 1998] who took the wavelet transform of the image and weighted the components before reconstruction. Similarly, Clarke and Qian *et al.* used a procedure resembling a wavelet transform to enhance microcalcifications [Clarke et al., 1994, Qian et al., 1995] and masses [Qian et al., 1999]. Their modified transform eliminated the need for empirically chosen weights for the enhancement of the specific objects in which they were interested[8]. Further, Lado *et al.*

[7] See Appendix C for a more detailed description.

[Lado et al., 1995] used a wavelet transform not only to enhance microcalcifications but extracted various features from the transformed image for the purpose of identifying clusters with malignant characteristics.

In an application similar to Chan [Chan et al., 1998], Wei *et al.* [Wei et al., 1995, Wei et al., 1997] created a method for the reduction of normal tissue that was mistaken for abnormal masses in a CAD system. The basic methods were the same as the studies described in Section 1.2.1 but in [Chan et al., 1998] a multi-scale texture measure was achieved using various pixel separations in the SGLD matrix, the d parameter. However, in [Wei et al., 1995], Wei *et al.* compared multi-scale texture analyses by using various values of d in the SGLD matrix created from the original image to the use of a wavelet transform of the image while constraining d to be 1. Wei found that the results using features formed from the wavelet transformed images were comparable or better than the results using textures formed from changing d. A later study can also be found, [Wei et al., 1997], where a more sophisticated preprocessing method and a larger number of textures was considered.

1.3 Overview

For this thesis, we explored several global mammographic characteristics for their ability to classify mammograms into density grade categories as well as into normal/abnormal groups. We began by extracting several global mammographic features and used either an exhaustive search or a genetic algorithm to select the subset of the best features to categorise the cases using either classification scheme.

[8] In [Qian et al., 1999] the image was enhanced to improve the performance of several textural, morphological and grey level properties in their CAD system.

The features that were extracted from the mammograms fell into two categories

1. Spectral features which can be obtained from the grey level histogram.
2. Multiscale texture features which were extracted using multifractal models and wavelet transforms of the images.

For the spectral features, we considered generalisations of the features in the studies given previously. In particular, extensions to the features used by Byng and Boyd *et al.*, such as combinations of regional moments, combinations of global moments and sub-regions of the global histogram itself were considered. The intent was to identify less obvious features or combinations of features that may have better classification ability than those found in the literature.

In addition to these spectral features we investigated several texture features: a multi-fractal dimension and three texture measures applied to the wavelet transformed images. The difficulty with the use of a fractal dimension in the work cited earlier was the property was selected with the intent of emulating the behaviour of the mammographic density. Such an approach is sufficient if simply a property which is less susceptible to intra- and inter-observer variation was desired. However, it does not consider new properties that may be independent of the density but still correlated to breast cancer risk or incidence. The technique of extracting textures from the wavelet coefficients of a mammogram has also been used previously but the applications were in reducing the false positive rate in a CAD system. Therefore, the textures were selected to distinguish a property characteristic of a malignant mass, such as a spiculated border. In this thesis all features were selected to characterise a global property related to cancer or cancer risk and these features may not be currently known to be correlated with malignancy.

In addition, the use of a few, or even a single quantity, to characterise the properties of a mammogram was common in the studies discussed previously. This would be sufficient in an application such as identifying a malignant mass where the border can possess characteristics quite different from a benign lesion. However, a more comprehensive characterisation of the mammographic properties would likely require a collection of properties to be calculated from the mammograms. There are different approaches that can be used for this purpose but we considered only properties that examined the image at different resolutions or scales. One feature in particular that was investigated was a generalisation to a fractal dimension that treats the object as a collection of fractals, possibly with different dimensions, that were intricately intertwined with each other, i.e. a multifractal. We also examined three conventional texture measures, the energy, entropy and inertia. All these textures had been used for segmentation of masses and microcalcification clusters in the studies listed above but we examined these textures for other purposes — their ability to distinguish density classes and to distinguish normal from abnormal groups. A wavelet transform of the image was also performed prior to extracting the texture measures after which the property was calculated directly from the wavelet coefficients. This procedure was simply to collect a pool of features and again we applied either an exhaustive search or a genetic algorithm to select a manageable subset of the features for either density grades and normal/abnormal classifications.

Chapter 2 describes the conceptual basis for the choice of spectral features which were employed — global and local moments, as well as subsets of the histogram, and the texture features — multifractal dimensions, wavelet transforms and the texture energy, entropy and inertia. However, one of the essential components needed for the work is the program used to select the essential properties. It is always difficult to identify the optimum components from a large pool of possible parameters and conventional systematic approaches tend to have difficulties with becoming trapped in local extrema. Therefore, we have employed a method of randomly searching the feature space through a genetic algorithm. The modifications to the basic approach necessary to use genetic algorithms was also briefly discussed in Chapter 2.

We proceed in Chapter 3 to discuss the details of the procedures needed to extract the desired features. This includes the normalization procedure applied to the images in order to remove systematic dependencies such as exposure and processor differences or characteristics specific to a particular x-ray digitiser. The extracted properties were evaluated for the ability to classify the mammograms into different classes and a number of datasets were needed for this. Several were selected to examine different goals and to evaluate the effects of the sample selection. In particular, many different classifications were possible, such as dividing the images on the basis of density grade, on the mammogram diagnosis or the patient diagnosis (where the left and right mammogram were regarded as having an "abnormal" outcome if the malignancy was in either breast). We describe the procedure used to select the various cohorts as well as presenting the specific details of the methods used to extract the various spectral and texture properties.

The results are presented in Chapter 4. There were many aspects to examine and the results are organised primarily along the lines of classification categories, eg. classification of density grades, classification of diagnostic outcome, etc. Additionally, the performance of the classifier when using each property, both individually and in combination, is described for the various classifiers. We also investigated some special situations such as x-ray digitiser dependencies and age dependencies and their impact on the classification performance.

Finally, a summary of the results is given in Chapter 5 along with potential future directions which are important to consider but beyond the scope of the current thesis. It should be noted that no attempt was made to investigate the specific *visual* features in the mammogram that correspond to selected abstract features that were extracted. The primary purpose of the thesis was exploratory. Therefore, the work reduces the problem from the selection of a single set of features from, say, millions of options to selecting one set from, say, tens. The correlation between the selected features and the visual properties was beyond the scope of this thesis.

A note on the terminology that appears in the remainder of the work should be given as well. The term "property" is used to refer to the mammographic characteristics extracted using different computational procedures, namely:

- global moments of the histogram
- regional moments of the histogram
- subregions of the histogram
- multifractal dimensions
- texture energy of the wavelet transformed image
- texture entropy of the wavelet transformed image
- texture inertia of the wavelet transformed image

However, several of these properties contain arbitrary parameters and the term "feature" or "feature set" refers to a specific combination of values for the parameters of a particular property. For example, the texture energy is a property while the texture energy for $d = 5$ and $\theta = 0^o$ is a feature.

Finally, unless explicitly stated otherwise, all programs used were created in house. This includes, but is not limited to: the procedures to segment the breast tissue in the mammograms from the background and the various programs to extract the properties.

Chapter 2

Theory

The primary methods of estimating the risk for developing breast cancer solely from its mammographic appearance, have revolved around the classification scheme developed by Wolfe [Wolfe, 1976a, Wolfe, 1976b, Wolfe et al., 1986] or the mammographic density [Saftlas et al., 1991, Boyd et al., 1995, Byng et al., 1996a]. Often these characteristics were evaluated by inspection of the mammograms and generally required an experienced radiologist, although some quantitative measurements of the mammographic density have employed planimeters [Wolfe et al., 1986, Saftlas et al., 1991]. However, with the increase in the accessibility of high performance computer systems combined with high quality x-ray scanners, more attention has been given to automated approaches. This trend has been encouraged by the results of Boyd *et al.* [Boyd et al., 1995] who showed, in a large case-control study, that the mammographic density is a significant breast cancer risk factor independent of the more commonly accepted risk factors such as family history, age of first live birth, etc. His results indicate that the general appearance of a mammogram contains significant information aside from the presence and location of abnormalities, the density being one simple characterization of the appearance.

There have been many approaches for the extraction of the additional information in a mammogram through, for example, the use of various texture properties [Magnin et al., 1986, Taylor et al., 1990], fractal dimensions [Boone et al., 1998, Caldwell et al., 1990], spectral properties [Tahoces et al., 1995, Karssemeijer, 1998] and several unique approaches as in [Shadagopan et al., 1982] (identification and quantification of ducts) and [Breitenstein and Shaw, 1998] (quantitative measurement of dense tissue). Regardless of the property which was considered, the investigators all attempted to classify the images into Wolfe grades or density categories. A drawback with this approach is the subjective nature of the various categories and the natural variability due to intra- and inter-observer differences. In this work, we forego the use of mammographic density classes, for the most part, and attempt to identify characteristics of the mammographic appearance that are indicative of disease. Due to the limitations imposed by the image database (see Section 3.1) combined with a relatively small data set, a quantitative estimate of the relative risk, as in [Boyd et al., 1995], is beyond the scope of this study. Rather, we confine ourselves to the identification of features which could be investigated further.

The mammographic features used in the thesis fall into two broad categories: what we will call "spectral"[1] features and "texture" features. Spectral features are generally simple methods of describing the global properties of the mammogram by characterising the distribution of the grey levels in the segmented image without regard to their spatial location. In general, these features can be extracted from a histogram of the frequency of appearance for each grey level. On the other hand,

[1] The term "spectral" features is somewhat nonstandard and was selected simply due to the resemblance of the grey level histogram to a intensity spectrum.

texture features are more complex and are attempts to quantify the appearance of the image. Texture features generally combine the grey level with some aspect of its spatial position. The remainder of this chapter is devoted to a more detailed description of the spectral and texture features which were employed as well as the method used for the selection of the most significant properties.

Many previous approaches to using texture features resulted in a single value for each texture and as a result many different textures were needed to distinguish mammograms which belong to different classes. Although Wei *et al.* was more concerned with distinguishing malignant masses from normal tissue, the general approach for the texture properties used in this study was similar to that found in [Wei et al., 1995] and [Wei et al., 1997] where a set of values was extracted which can describe the characteristics of the image in a straightforward and natural way. Specifically we employed a few simple texture properties, such as the contrast, but each image was transformed in such a way as to generate a collection of images viewed at different length scales. The texture features were then applied to the set of images. This vector of multi-scale values can be used to characterise an image more completely than does the same property when applied to only a single scale image.

We also examined a feature which was inspired by a multifractal dimension. It has been found that many objects encountered in nature with a fractal character behave as though they were composed of a collection of intricately intertwined single fractals. For these objects a continuum of fractal dimensions is needed to fully describe the object. A more detailed discussion of the multifractal dimensions is given in Section 2.2.1 while the texture features are described in Section 2.2.2 and the spectral features below.

2.1 Spectral Properties

All of the spectral properties which were used can be generated from the function that quantifies the frequency of appearance for each grey level in an image (histogram or grey level histogram). The histogram was found for the region that was segmented to contain just the breast shadow[2]. The mammographic density is an example of a spectral feature. The greater the density the more radio-opaque the tissue and the brighter the region appears on the radiograph. Therefore, the density or proportion of dense tissue can be viewed as the proportion of bright pixels in the segmented region. Unfortunately, the differences in exposure for different patients changes the threshold grey level that delineates the majority of the dense parenchymal tissue from the "dark" fatty tissue. While a human can readily compensate for the differences in exposure, attempting to give a computer program a comparable facility is quite difficult. For this reason, the density was not employed for this study. The spectral features which were actually used consisted of:

1. The global moments of the histogram. This is a method of describing specific properties of the grey level distribution. Global moments were calculated from the entire segmented breast image. Some commonly used moments are the mean (first moment), variance (second moment) and skewness (third moment).

2. The regional moments. These are the averaged moments calculated from histograms generated from subregions of the segmented breast tissue and were calculated as in [Boyd et al., 1995, Byng et al., 1996a, Byng et al., 1996b]

3. The mean of subregions of the global histogram which were the most significant

[2] The segmentation procedure is described in Chapter 3.

in classifying the images.

One advantage of both the global and regional moments over the density is that both are extensible, that is, it is easy to generate a large number of moments, each of which examines a different characteristic of the histogram. In addition, when using the density, a single threshold is desirable to signify the presence of a pixel containing parenchymal tissue for all images. However, the use of a single threshold would make the density sensitive to both the shape and position of the non-zero parts of the histogram whereas the moments generally isolate these characteristics into separate moments.

The remaining spectral feature, the subregions of the histogram, would also be sensitive to the same type of changes in the histogram, but this property has the potential of providing considerably more information than the density could provide. The potential information that could be extracted was more than sufficient to justify the additional difficulty in compensating for the exposure differences and, hopefully, the procedure for sub-region selection can identify regions that were relatively insensitive to these systematic changes. All these spectral features were also very straightforward to evaluate and are described below.

2.1.1 Moments

The global moments of the image were calculated from a histogram of the entire segmented region, while the regional or local moments were calculated using the histogram for many small regions lying within the segmented breast shadow. In either case the raw moment, m'_i, was obtained using the usual definition

$$m'_i = \begin{cases} \frac{1}{N}\sum_{j=0}^{N} P_j, & i=1 \\ \frac{1}{N}\sum_{j=0}^{N} (P_j - \overline{P})^i, & i>1 \end{cases} \tag{2.1}$$

where P_j was the grey level for the j^{th} pixel, $\overline{P}(\equiv m'_1)$ the average pixel value and N the total number of pixels. For the higher moments, $i > 2$, it was more convenient to employ a unit-less quantity by normalizing m'_i to the standard deviation raised to the appropriate power. Hence we now have

$$m_i = \begin{cases} m'_i\,, & i \leq 2 \\ \frac{m'_i}{m_2^{i/2}}, & i > 2 \end{cases} \tag{2.2}$$

The drawback with describing the characteristics of the histogram through simple hierarchial properties, such as the moments, was that they tend to be most useful for relatively simple problems. A difficult classification problem would likely be dependent on more subtle characteristics of the distribution which is manifest in the higher moments. Unfortunately, these same moments tend to be extremely sensitive to small differences in the distribution and may cause problems when evaluated numerically because of the high value of the exponent. Therefore, what is desired is the smallest collection of the lowest moments necessary to classify the images.

When images have different values in the low moments of their histograms, the images tend to have obvious differences. For example, an image with a large value for the mean was brighter overall than one with a low value. This is particularly important since the purpose of the automatic exposure control (AEC) of the mammographic unit was to make the films have the same overall optical density. Hence, the mean was unlikely to be useful for classification. However, as described in [Byng et al., 1999], a slightly higher moment, m_3 (the skewness), reflects the relative contribution of bright to dark pixels in the image and would be useful for density classification. On the other hand, very high moments are quite sensitive to variations in the distributions and the natural variation in the appearance of the parenchyma from patient to patient would make the range of possible values in each class so broad that it would not be possible to resolve the different classes.

For a difficult classification problem it is often not obvious which set of moments that would give the most accurate classification. Hence, more moments than are likely to be useful were intentionally calculated and various combinations of the available moments were tested for the best subset.

The regional moments are generated in a similar fashion but use a much smaller region of the segmented breast shadow. The procedure basically followed that described by Boyd *et al.* and Byng *et al.* in [Boyd et al., 1995, Byng et al., 1996a, Byng et al., 1996b]. Here, only a brief overview of the procedure is described. Greater detail can be found in the references and in Chapter 3.

It is expected that moments calculated using more local information will be better able to deal with inhomogeneities in the tissue type [Byng et al., 1996a]. Additionally, while the thickness of the compressed breast was fairly uniform over the middle region, toward the skin surface the thickness changes rapidly and the amount of glandular tissue was more pronounced toward the chest wall. These effects may obscure the differences we wish to identify in a histogram found using the entire segmented breast tissue but local moments, calculated strictly from small regions, can give a better reflection of the tissue composition.

Some studies, such as [Magnin et al., 1986, Tahoces et al., 1995], examine only a constrained region of the mammogram for their respective properties. However, the size and position of the region that would optimize the performance of each extracted feature is unclear. Therefore, rather than choose a single subregion, the entire segmented breast tissue was divided into many regions. The moments of the histogram for each region was found and the corresponding moments were then averaged together. Since it is less likely that the compressed breast thickness varies as dramatically over the smaller region, an averaged regional moment is less susceptible to confounding factors such as variations in thickness and more closely reflects differences in the proportion of tissue types.

2.1.2 Histogram Regions

An alternative to using the moments with their accompanying drawbacks, was to utilise simple statistics calculated from a small portion of the entire histogram. The expectation was that the amplitudes of the histogram have significant classification ability. Using properties from only portions of the histogram also has the advantage of being quick to calculate and allows regions with little classification ability to be ignored. The difficult task was then to identify the regions of the histogram that were the most useful for separating the images into the desired classes. If up to, say, 10 regions of varying widths were to be chosen from a total pool of 4096 grey levels, it was clearly not possible to do an exhaustive search of all the possible combinations.

Fortunately a program from the Institute for Biodiagnostics, NRC, was developed for this type of classification problem. The program, "**ga_ors**", utilised a genetic algorithm which is a very powerful method of randomly searching a large feature space for an optimal or near optimal configuration relative to a fitness function [Nikouline, 1998], in a reasonable amount of time. It has been used for a similar purpose in classifying infrared (IR) and magnetic resonance (MR) spectra into normal and abnormal groups. For our purposes, it was used to select a predefined number of regions that were the most useful in discriminating the patient classes. The mean was taken as the property to characterise each selected region. A further discussion on genetic algorithms in general can be found in Appendix B and in Section 2.3.

2.2 Textures

Whereas spectral features characterise only the frequency of the appearance of each grey level, textures characterise a specific aspect of the spatial relationship between the grey levels as well as their frequency of appearance. There is a large number of possible textures which can be utilised, each of which considers a slightly different characteristic. For instance, Byng *et al.* [Byng et al., 1996a] have used a fractal dimension for analysis of mammographic densities and Magnin *et al.* [Magnin et al., 1986] have examined the classification ability of a number of conventional textures for a similar purpose. Many textures were constructed to quantify a specific aspect of an image, such as the apparent roughness or the proportion of vertical lines, and tend to be moderately simple to evaluate with an intuitive interpretation but there are many others which are extremely complex and have no easily discernable physical interpretation.

A few textures that belong to both categories had been selected for the particular application described in this thesis. Multifractal dimensions (Section 2.2.1), which are an extension to a fractal dimension, were considered along with three simple texture measures applied to a wavelet transform of the images (Section 2.2.2). The textures were selected to produce a set of values that characterised the images at different scale lengths.

2.2.1 Multifractal Dimensions

Many earlier studies had shown that a fractal dimension was useful in texture classification in a variety of different applications. For example, the fractal dimension used by Byng *et al.* [Byng et al., 1996a] to distinguish mammographic density classes was calculated by treating the image as a surface where the height was represented by the pixel value and the variation in the area of the surface was examined as a function of scale. For their case the intended texture feature was the roughness of the surface and a rougher, more convoluted surface, would produce a fractal dimension closer to three (characteristic of a volume) than to two (characteristic of a surface). Although a fractal dimension is inherently a multi-resolution characteristic, many natural objects with a fractal character are often actually multifractal and require a continuum of values to fully characterise the object. Further details of both conventional fractal geometry and multifractals can be found in Appendix A and the references.

Conventional fractal objects, such as a Sierpinski gasket, a Koch curve or a Peano curve (Figures A.2–A.3) are examples of strictly self similar objects. That is, portions of the object can be made to appear identical to the original, if the portion is rescaled by the appropriate factor. However, many physical objects with fractal-like behaviour are created by random processes and the resulting objects are statistically rather than strictly self-similar. Common examples of random fractals are coastlines and mountain ranges. For these objects it is not possible, in general, to make any sub-region exactly correspond to the original but the general character of the sub-regions does resemble the full object. Indeed, if an image of the sub-region is viewed without reference to the original it is difficult to judge whether it is a sub-region or the full object.

Multifractals are generally random fractals and can be thought of as consisting of many random fractals, with possibly different dimensions, which are intricately intertwined. Then, when different approaches to calculate the fractal dimension are applied, a different dimension may result depending on the "fractal component" to which the method is most sensitive. Because of this, when the fractal dimension is applied to any natural object the method of calculation for the fractal dimension is critically important.

The difference between the conventional fractal dimension and multifractal dimensions is more apparent in a specific example. Consider the situation of several fields with different types of ore visible over its surface. The fields are approximately the same size but of vastly different composition and value. It would be desirable to identify the most valuable field, but it is too difficult to estimate the total value of the ore for all fields. In that case we may be interested in the distribution of ore over a relatively small sample of each field and assume it is typical for the entire region. It is likely that the distribution has a fractal character and one approach which is often used to evaluate the dimension is to use what is frequently called the box counting dimension (or Hausdorff mesh). In this approach, a regular grid with a side length of ε is superimposed over the field and the number of cells, N_ε, which contain any type of ore are counted. The process is then repeated with many different sized meshes.

The value of the fractal dimension, d, is then related[3] to the slope of the regression fit of $\log N_\varepsilon$ as a function of $\log \varepsilon$. A dimension closer to two indicates a greater amount of ore but this process ignores the type of ore in each cell. Further, if the net value of a collection of ore is desired, the composition of the samples in each cell is very important.

The distribution of ore is more likely multifractal and the multifractal dimensions can be found following a method similar to that used for the box counting dimension. The process of calculating the dimensions starts with the same regular grid but we assign a weight to each cell, μ_{ij}, where ij specify a location within the mesh. In this case, the total value of the ore in the cell may be used for this purpose. The distribution can then be characterised by the set of fractal dimensions for the various collections of cells with the same μ_{ij}.

From this point, there are different approaches which can be applied. In this work, the technique known as the method of moments was used. What follows is a brief overview of the approach. A detailed description of the method can also be found in [Peitgen et al., 1992]. This approach was pioneered by Rényi and employs what is known as a partition function[4], $\chi_q(\varepsilon)$, for the q^{th} moment where

$$\chi_q(\varepsilon) = \sum_{i,j}^{N_\varepsilon} \mu_{ij}^q \ , \quad q \in \mathbf{R} \tag{2.3}$$

The partition function is analogous to the number of cells needed to cover the object, N_ε, in the box counting dimension. Therefore, for a fractal object, χ_q scales with the

[3] When this procedure is applied to an image, the slope is exactly the fractal dimension but for other methods this may not be true.

[4] The term was coined due to the parallels between how χ_q is used and a partition function of statistical mechanics [Schroeder, 1991].

characteristic length and the generalised fractal dimension, D_q. However, the method of moments is not identical to the Hausdoff mesh approach and an additional factor of $q-1$ is required. We now have

$$\begin{aligned} \chi_q(\varepsilon) &\propto \varepsilon^{(q-1)D_q} & (2.4) \\ \chi_q(\varepsilon) &\propto \varepsilon^{\tau} & (2.5) \\ \tau &= (q-1)D_q & (2.6) \end{aligned}$$

The D_q is known as the generalised fractal dimension since specific values of q correspond to more commonly known dimensions. For example $q = 0$ gives the usual box counting dimension while $q \to 1$ corresponds to the information dimension [Peitgen et al., 1992, Schroeder, 1991]. The calculation then proceeds similarly to the box counting dimension[5] with χ_q substituted for $N(\varepsilon)$. A property has been extracted from the images based on this procedure. The precise details of the method are given in Chapter 3.

2.2.2 Texture Measures

Texture measures have been used in the past for a wide range of applications including many in digital mammography. This includes segmentation of suspicious masses, [Gupta and Undrill, 1995], and the separation of masses into benign and malignant classes [Chan et al., 1995]. As well, Magnin *et al.* [Magnin et al., 1986] applied similar textures as the ones selected of this work to automatically classify mammograms into Wolfe grades. However, to fully characterise the image a large number of texture

[5] The primary difference is that, for a multifractal, the calculation must be repeated as q changes.

measures is frequently required and many texture measures do not have any apparent physical interpretation. Studies have also been done in using a wavelet transform combined with texture properties for the segmentation of masses [Qian et al., 1995] and microcalcifications [Qian et al., 1999].

In the present work, and following the approach of Qian *et al.* [Qian et al., 1995, Qian et al., 1999], a set of textures to characterise the image was extracted in order to classify the set of mammograms into several categories. However, the texture measures were constrained only to simple textures with an intuitive physical interpretation. Further a multi-scale decomposition of the mammograms was performed in order to characterise the images more fully. The textures that were selected can be found in, for example, [Haralick et al., 1973, Magnin et al., 1986, Wei et al., 1995] and are sometimes referred to as the energy, H, entropy, S, and the inertia, I. All three can be calculated from the SGLD matrix and are given by Equations (2.7)–(2.9).

$$H(d,\theta) = \sum_{i,j} p_{ij}^2(d,\theta) \tag{2.7}$$

$$S(d,\theta) = \sum_{i,j} p_{ij}(d,\theta)\log p_{ij}(d,\theta) \tag{2.8}$$

$$I(d,\theta) = \sum_{i,j} (i-j)^2 p_{ij}(d,\theta) \tag{2.9}$$

where p_{ij} is the entry in the SGLD matrix for pixels with grey levels of i and j. The d and θ are arbitrary parameters. (See Chapter 3.)

The colourful names are derived from the form of the equations which resemble their physical counterparts. Both H and S quantify the homogeneity of the image; summing two different functions of the probabilities, p_{ij}, over all possible combinations of grey levels. I, on the other hand, is a measure of the contrast obtained by considering a function of the difference in grey levels, $(i - j)$, but weighted by the probabilities, p_{ij}. The previous studies cited required a considerable number of textures for reasonable classification performance including many that were much more complex than those chosen here. Rather than emulating this approach, the texture measures that were selected were constrained to the energy, entropy and inertia. In order to obtain a more complete (textural) description of the image these textures were applied to the images at multiple scales.

The required multi-scale decomposition of the images was performed through the use of a wavelet transform. At this time, the use of the wavelet transform is quite widespread but not a part of most standard curricula, therefore, a brief presentation of method is given.

Historically, the development of wavelet transforms had its origins in many diverse fields of study and one of the results of this was that there are two explanations pervasive in the literature. The first is quite mathematically intensive and puts the transform on a rigourous basis while the second is more relevant for creating efficient implementations. Some of the mathematical basis of the transform along with the connection between the two interpretations is given in Appendix C while a brief overview of the technique is described below.

The transform can be viewed as a general integral transform, $\mathcal{T}$, of a function, $f(x)$, with a form

$$(\mathcal{T}_{\psi}f)(a,b) = C\int_{-\infty}^{\infty} f(x)\psi(a,b)dx \tag{2.10}$$

where C is a normalization constant and $\psi(a,b)$ a basis of functions of position (characterised by a) and a scaling factor (characterised by b). The choice of basis functions determines the overall properties of the transform and a Fourier transform becomes a special case where the basis $\{\psi(\cdot,\omega) \equiv e^{i\omega t}\}$ is used. Some of the important properties resulting from this choice of basis is that the contribution of each frequency for the original signal is found but the transform contains no spatial information. In addition, a large number of terms is needed to represent a signal with sharp transitions, like edges or boundaries.

Many of the typical bases chosen for a wavelet transform attempt to reduce the extreme properties of a Fourier transform. For example, some spatial *and* frequency information can be extracted from the wavelet transformed signal. The exact choice of bases that was used for the transform was a bi-orthogonal wavelet described by Sweldens [Sweldens, 1994]. An example of two typical functions in the basis is shown in Figure 2.1.

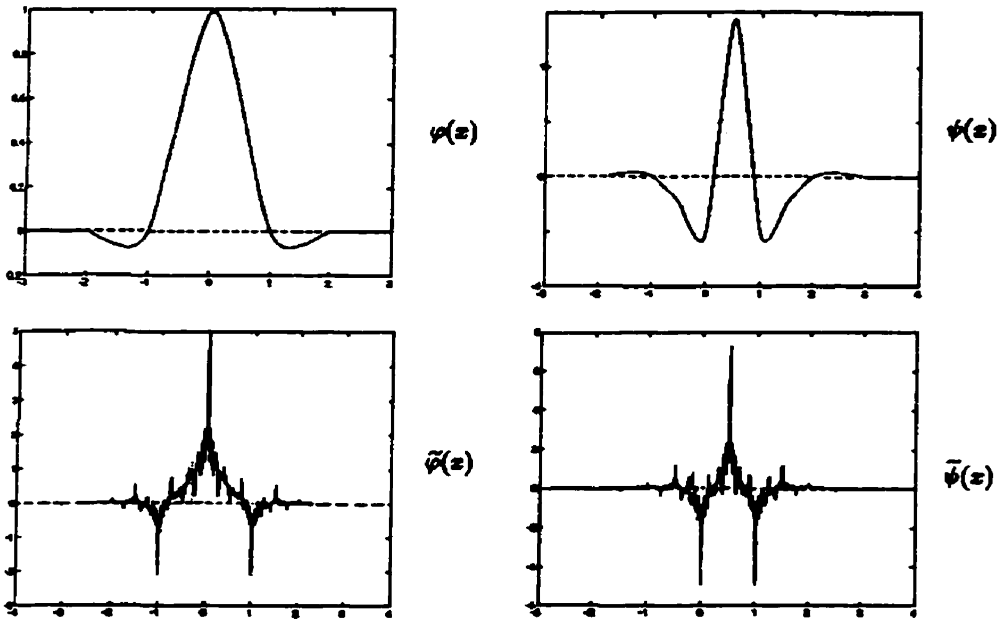


Figure 2.1: Example of the mother wavelet and scaling function, along with their duals, for a bi-orthogonal wavelet as described by Sweldens. ([Sweldens, 1994], page 21)

An additional, and important, feature of a wavelet transform is that it is possible to generate a multi-resolution analysis of the input signal. The typical technique to achieve this result is to perform an iterative decomposition on the input with the resolution at each successive level as one half that of the preceding level. Then, the transform can be performed by sending a copy of the signal through both a high pass filter and a low pass filter followed by sampling the output by two, for a discrete signal. For the next level, the output of the low pass filter, only, is subjected to a second pair of high and low pass filters and down-sampled by 2. The output of each high pass filter is sent directly to the output and the process repeated until the desired number of levels is obtained. (See Figure 2.2.) For this work, a maximum of five levels of the decomposition was utilised. A two dimensional image can be transformed by applying a one dimensional transform to the rows and columns successively and this produces a different result for each quadrant. Clearly, there is some ambiguity in the order of application of the transform and the most common arrangement, due to Mallat [Mallat, 1989a, Mallat, 1989b], is shown in Figure 2.3 where $XYXYXY\ldots$, $X, Y \in \{\mathrm{H}, \mathrm{L}\}$ represents the coefficients after the sub-image was subjected to a high pass filter (H) to the columns (X) then a low pass filter (L) on the rows (Y)(and down-

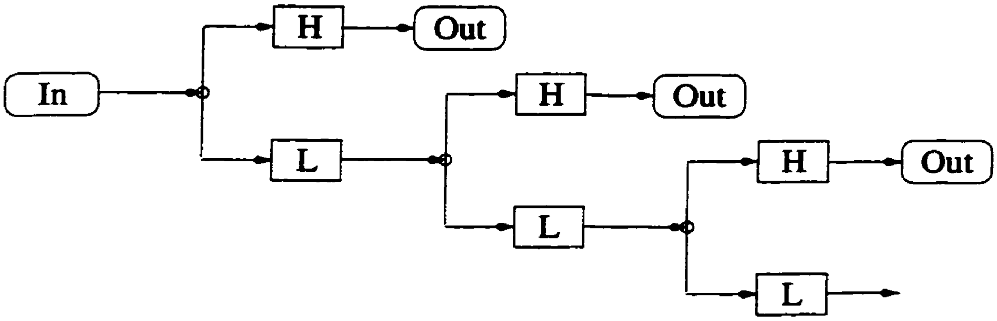


Figure 2.2: Discrete Wavelet Transform as a filter bank cascade

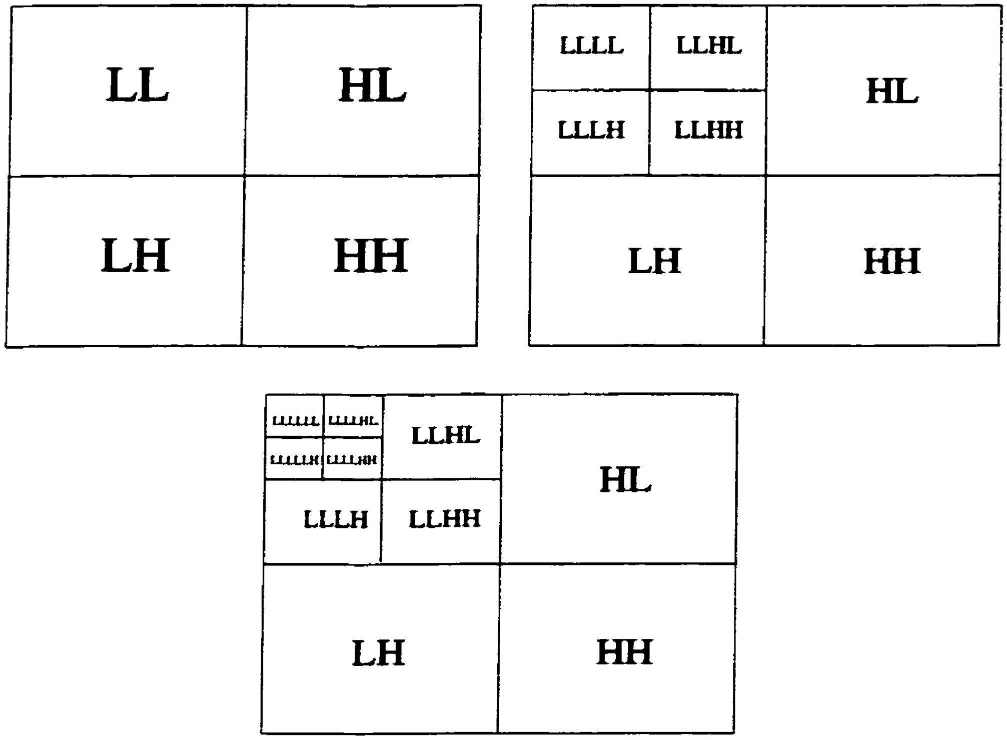


Figure 2.3: Mallat format for three levels in the two dimensional wavelet transform, showing the band pass filter order over a image and where L and H represent low and high band pass filters respectively.

sampled by 2). Once the transform was performed, all three textures (H, S and I) were calculated from the images in the quadrants after filtering by HL, HH and LH at each level of the transform[6]. With five iterative levels of the transform retained, combined with three textures obtained from three quadrants at each level and 20 different choices of d and θ for the SGLD matrices results in a total of 900 different textures. The large number of variables made it necessary to use a sophisticated mechanism which will find the most important textures and reduce the number which were actually used to classify the images to a manageable level.

2.3 Classification Methods

There were two considerably different operations needed for our work.

1. The selection of a small subset of features that have the greatest discriminatory power or classification ability for our given problem.

2. The evaluation of the classification performance of those same features.

Indeed, within the realm of feature selection a method of evaluating a set of features was necessary in order to identify the most promising subset. Linear discriminant analysis, described in Section 2.3.2, was used for the evaluation of the subsets of features. The selection of the subset of the best features was made using either an exhaustive search of all possible subsets of features, if the number of combinations is small enough, or using a genetic algorithm otherwise. A general description of genetic algorithms is given in Appendix B while the details of the technique that are specific to the program **ga_ors** are described in Section 2.3.1.

[6] Note that the remaining quadrant that was subjected to the LL filter combination is used for the input image for the next level of the transform.

2.3.1 Parameters for **ga_ors**

As can be seen in the literature, (see Chapter 1), the image properties which can be obtained from a mammogram are numerous and there is practically an unlimited number of additional features that can be employed. A daunting aspect of this study is the selection of a set of features that provides reliable classification while being sufficiently few in number to be manageable. Additionally, the nature of the extracted image properties required that several considerably different methods of feature selection be employed. For some features, such as the global or regional moments, the pool of properties from which the subset of features were to be selected is small enough that an exhaustive search is practical. For other features, most notably the histogram regions and the textures, the number of features was large enough that an exhaustive search was computationally too expensive. Another complication was that the number of features under consideration (> 900 in one case) greatly exceed the number of cases in the sample (max. 240), therefore even if an exhaustive search were possible it might well result in an overfitted solution. For these properties, we turned to a program developed at the Institute for Biodiagnostics, **ga_ors**, to perform the feature selection. The program uses a popular technique to randomly explore a very large feature space, a genetic algorithm. The program used in this work, **ga_ors** finds a user selected number of "best" features. The procedure used by **ga_ors** is an extension to the conventional approach and details of the differences can be found in [Nikouline, 1998].

Many aspects of the genetic algorithm are usually problem dependent, such as the map between the histogram regions to genes on the chromosome. For **ga_ors**, the histogram was treated as a collection of subregions and the mapping designates the bins in the histogram that were to be taken as part of the same subregion. The chromosomes were represented by bit strings and a 1 in the i^{th} bit represents the inclusion of the subregion containing the i^{th} to $(i+1)^{\text{th}}$ grey level but a 0 in the chromosome indicates those grey levels should *not* be included in the feature set for the chromosome. The length of the chromosome was therefore the same as the number of bins in the histogram, 4095 for the smallest bin size, and there were at most a user selected number of contiguous regions containing 1's, say 5, for example.

Another important choice for the genetic algorithm lay in the creation of the objective function. **ga_ors** uses an objective function based on the squared difference, or error, between the classification results from a linear discriminant procedure and the known classification for each case[7]. The total squared error for each chromosome was then used to rank the population and any repeated chromosomes were removed.

To enable the population to reproduce, i.e. explore the solution space, **ga_ors** used the genetic operators, mutation and crossover. The crossover operator was quite conventional, see Appendix B, but the operation of mutation was somewhat unusual in that a single gene was not necessarily changed at a time. A block of k genes was changed with each mutation and k varied as the population evolved. Initially, k was $\frac{1}{64}$ of the range of possible values so for a histogram of 4096 grey levels, initially $k = \frac{4096}{64} = 64$ and decreased with each generation. This allowed the mutation to have a noticeable influence throughout the entire procedure[8].

The final detail to be described lies in the creation of the next generation. For ga_ors, once the chromosomes had been evaluated, ranked and the repeated chromosomes removed, the best N_E chromosomes were immediately transferred to the

[7] The classes are enumerated in order to be able to find the squared difference.

[8] The conventional approach of changing a single gene for each mutation makes the effects of the mutation operator significant primarily during the later generations [Nikouline, 1998].

population for the next generation (elite population) and the remainder formed by reproduction in the full, current population. The chromosomes did not have an equal likelihood of reproducing, rather the chromosomes with a higher rank (lower classification error) were more likely to be selected and the probability decreased in proportion with the rank. The selected chromosomes were then mutated with probability p_m and the operation of crossover performed with probability p_c. After the genetic operations had been performed the resulting chromosomes were placed in the new population. The parameters used in **ga_ors** for this work were:

$$p_m = 0.001 \tag{2.11}$$

$$p_c = 0.66 \tag{2.12}$$

$$N_p = 300 \tag{2.13}$$

$$N_g = 50 \tag{2.14}$$

$$N_E = 10 \tag{2.15}$$

where N_p is the size of the population or number of chromosomes and N_g the number of generations to allow the population to evolve.

2.3.2 Linear Discriminant Analysis

Regardless of the method used to explore the feature space, a technique was needed to evaluate the classification performance for each candidate feature set. Linear discriminant analysis (LDA) was used for this purpose. See, for example, [McLachlan, 1992]. This is a conventional statistical technique to form a function that can be used to distinguish the various classes. The linear discriminant function, Z, has the form

$$Z = \sum_{i=0}^{m-1} a_i X_i \tag{2.16}$$

for a set of m features, X_i, and the m coefficients, a_i. The precise form for the set of constants, a_i, can be calculated from maximizing an F statistic. In this case the F statistic is defined using the ratio of the mean square variance between classes to the mean square within class variance so that maximizing this quantity produces the tightest groups with the largest separation. The evaluation of the statistic itself can then be found following [Manly, 1986] or [Bernstein et al., 1988], for example.

At this point, the precise combination of a_i's that maximises F needs to be found. Fortunately, Fisher described the approach in 1936 [Fisher, 1936]. If a numeric value is assigned to each case in the sample depending upon the class to which the case belongs, then the necessary mathematical procedure is identical to finding the least square coefficients for linear multivariate regression [Flury and Riedwyl, 1988]. There are different approaches for the conversion of a categorical group label to a numeric one but the method given by Fisher for a two group problem is to replace the original label: abnormal/normal, for example, by: c_1/c_2 where

$$c_1 = \frac{n_2}{\sum_{i=1}^{k} n_i} \tag{2.17}$$

$$c_2 = \frac{-n_1}{\sum_{i=1}^{k} n_i} \tag{2.18}$$

and n_i is the number of cases in the i^{th} class out of a total of k classes[9]. The necessary procedure can be found, in detail, in [Bernstein et al., 1988, Manly, 1986,

[9] $k = 2$ for the abnormal/normal classification.

McLachlan, 1992] while an outline of the equations alone in [Zwillinger, 1996].

The final step in the classification was to evaluate the performance of the discriminant function. Clearly there is little point in using the same dataset which was used in the creation of the discriminant function, since the function was created with the intention of optimizing the classification accuracy for that sample. A small number of the images ($\frac{1}{3}$ of the total) were always reserved for a test set and the remaining images ($\frac{2}{3}$) used for the "training" set. The results given in Chapter 4 were exclusively from the selected test groups.

Chapter 3

Materials and Methods

The features used for this work were extracted from a database of digitised mammograms which were made publicly available from the University of South Florida. The actual mammograms themselves originated from several different centres and were digitised using several different x-ray scanners. The database and the cases used for this study are described further in Section 3.1.

Due to variations in the film type and exposure conditions as well as variations in the performance of the x-ray scanners, it was necessary to normalise the images, Section 3.1.2. The images were transformed to remove the effects of the different pixel sizes and differing grey level/optical density calibration from the various scanners. A simple correction was also made for the exposure differences using a characteristic of the grey level histogram (also described in Section 3.1.2).

After the normalizations were conducted, several features based on the spectral features and texture properties were extracted from the corrected images, Section 3.3. From the full collection of extracted properties, the most significant were selected using a genetic algorithm (**ga_ors**) or an exhaustive search through all possible features.

The results of the analysis are reserved for Chapter 4.

3.1 Images

The most obvious method to obtain the mammograms needed for this work would be to turn to the local breast screening centres. This way a data set could be created that exactly meets the criteria for any desired aspect in our study. Unfortunately, this posed some local study difficulties.

1. Much of the needed patient screening information is not stored electronically so that even creating a list of patients that meet a particular set of criteria was a tedious and time consuming process.

2. Many of the screening films and patient files are not kept at the Health Sciences Centre, Winnipeg, Manitoba and there can be a significant delay for their delivery.

3. Our departmental x-ray scanner (from Vision Ten Inc.) was designed for general radiology and not intended for mammography. The scanner was sensitive to an optical density range from 0 to 2.5 and was roughly linear from 0–2.0 which is inadequate for use in mammography where an optical density of 3 or more on some parts of the mammogram is not uncommon. There were also limitations imposed by the computer system driving the scanner that made the image acquisition process more difficult than necessary. Additionally, the quality of the scanned image itself was less than desirable, containing various scanning artifacts that made segmentation of the breast tissue from the background unusually difficult.

Initially, this was the approach used to obtain data. However, the combination of the effects described above resulted in a total of $\sim$ 60 patient exams which were scanned, preprocessed and the necessary features extracted in $\sim$ 8 months of effort. To obtain a reasonable number of cases, say 250, approximately 2.5 years would have been necessary for the data acquisition process alone. Therefore, in order to obtain a statistically significant sample of images an external source of screening mammograms were used for the feature selection and analysis. The locally obtained images were used strictly for the formulation and evaluation of the segmentation and feature extraction procedures prior to the processing of the images from the external source.

In particular, a set of cases from the Digital Database for Screening Mammography (DDSM) was used. The images are available from the University of South Florida and consists of the digitised screening mammograms from a large number of women from several different centres [Heath and Bowyer, 1998]. The database consists of the digitised mammograms, the diagnosis and some basic information for each patient such as the age, date of the study, density classification, etc. The mammograms themselves were obtained using conventional techniques and then digitised using one of three high performance x-ray scanners (DBA M2100, Howtek MultiRad 850 or Lumisys 200). The characteristics of these scanners were sufficiently different from the departmental Vision Ten scanner that it was more straightforward to obtain additional images than to attempt to incorporate the locally obtained images into the sample from the DDSM.

The patient cases were also provided with three classifications:

1. normal with at least 5 years of follow-up

2. abnormal with a biopsy confirmed malignancy
3. abnormal but benign changes.

Only the first two cases were considered and the mammograms from a total of 240 patients were obtained from strictly the first two categories[1].

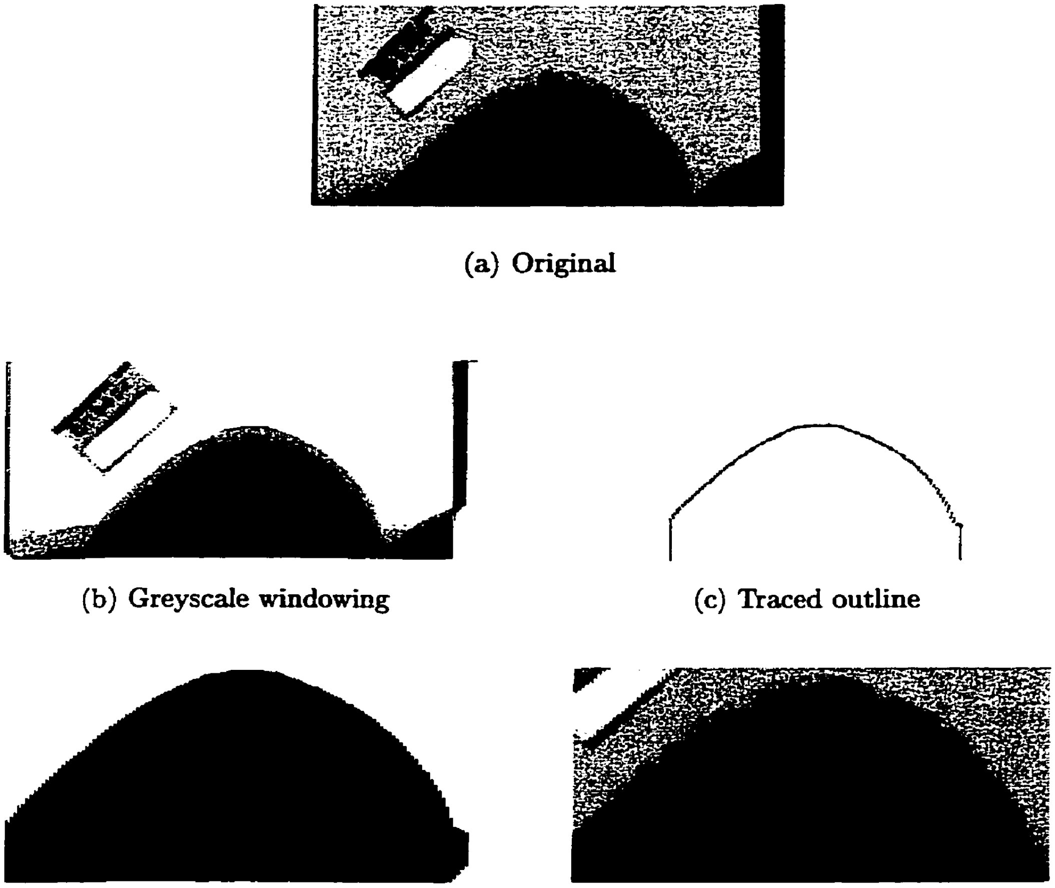
The drawback with this organisation is that the sample is less than ideal for applications that do not utilise these categories. For example, the examination of breast cancer risk can be performed through the mammographic density. The density grade classification, according to the BiRads guidelines [ACR, 1993], was provided with the patient information but the cases were not selected for a uniform distribution of cases in each grade. In particular, there were few examples of mammograms in the lowest density grade in our sample. The risk can also be evaluated directly as was done by Boyd *et al.* [Boyd et al., 1995]. A similar analysis could not be performed with this dataset due to a lack of patient information. An evaluation of the relative risk requires the selection of similar cases and controls who are matched for breast cancer risk factors that are beyond our control, such as the age, nulliparity, age of menarche, etc. The effects of these limitations are further discussed in the context of the results which were obtained in Section 4.1.1. However, it was possible to analyse the extracted features for a correlation with the appearance of cancer directly. That is, the features that were useful for classifying normal/abnormal were identified without relying on the density grades as an indicator of risk.

Two further pre-processing steps were performed prior to the extraction of mammographic features: the segmentation of the breast shadow in the images and normalization for systematic variations. Both these steps are described below.

[1] It was felt that the behaviour of the classification system for the third class of images would be beyond the scope of the current study.

3.1.1 Segmentation

The cranial-caudal (CC) views were employed and transformations were applied to the images to give them a uniform orientation, chest wall to the bottom, mid-line to the right (Figure 3.1(a)), and the segmentation was performed using a semi-automatic procedure. The first step involved smoothing the image by averaging over a 5x5 pixel^2



(a) Original

(b) Greyscale windowing

(c) Traced outline

(d) Cropped filled and smoothed mask

(e) Cropped final image

Figure 3.1: Example of segmentation procedure. The rectangular object in the upper left is a tag that provides patient and study information. The scale for (d) and (e) are different from (a) to (c).

kernel and retaining only the resulting values which lay between some maximum and minimum (Figure 3.1(b)). The limits were chosen by inspection on a case by case basis such that the majority of the region outside the breast shadow and most of the regions obscured by muscle tissue were outside the selected grey level range. Next, a routine which traced the outline of the breast was applied (Figure 3.1(c)). The procedure required two points to be chosen manually as the endpoints. The delineated region was then filled and the edge smoothed by repeatedly applying a dilation and erosion operator with a 11x11 pixel^2 square kernel to form the image mask. The mask and image were also cropped to remove most of the unnecessary regions outside the segmented tissue (Figure 3.1(d)–3.1(e)). Note that other than cropping the image, no modification was made to the image itself during this part of the preprocessing.

For the majority of the images this was sufficient to produce a mask which isolated the breast shadow from the remaining part of the mammogram. A number of the images required a substantially greater amount of custom editing. For example, an insufficient amount of the visible muscle tissue may have been removed automatically and some of the images had information tags very close to the breast shadow itself, which could mistakenly be included in the mask. As well, any radio-opaque markers (mostly beads) indicating regions containing suspicious tissue or the position of the nipple were also removed.

3.1.2 Normalization

Since the average overall optical density was maintained at a uniform level by the automatic exposure control of the mammography unit, the differences in the distribution of tissue types between patients may require that different regions of the films' sensitometric curve be used in their respective mammograms. Further, the variation in the characteristics of the x-ray scanners also introduce differences in the images which must be corrected. Some of the details of the scanners' performance were provided with the database which allowed the images to be normalised for these variations, in principle. However, insufficient data was provided on the exposure conditions and film characteristics prior to digitization of the film. Therefore, it was necessary to employ the characteristics of the grey level distribution of the images themselves and perform an elementary correction for these film variations.

Scanner Differences

The x-ray digitisers which were used to obtain the images were from Lumisys (LS), DBA or Howtek (HT), all of which had a similar dynamic range but their detailed characteristics were considerably different. For example, the scanners all had different resolutions (LS: 50μm/pixel, DBA: 42μm/pixel and HT: 43.5μm/pixel); LS and HT had a linear response curve while the DBA did not.

To compensate for the various resolutions, the images were re-sampled using linear interpolation between pixels along the rows and columns in succession. This procedure was performed in order to provide a uniform pixel size of $\sim 110\mu$m. The reduced resolution was comparable to that used in [Byng et al., 1994], [Byng et al., 1996b], [Chan et al., 1995], [Karssemeijer, 1998], or [te Brake et al., 1998] and since we were not attempting to identify the location of high detail characteristics, such as microcalcifications, the low resolution had little impact on the performance of the system. In addition, a reduced resolution removed some of the high spatial frequency MTF and noise differences. Finally, the re-sampling considerably reduced the storage requirements for the data as well as improving the computational speed of the various algorithms.

It was also necessary for the images to have a single response curve so that the pixel values correspond to a unique optical density. Since the calibration data was supplied with the DDSM, the response of the various scanners for a given optical density was known. It was also apparent that the Lumisys digitiser, LS, had the simplest response curve. This was expected since the LS digitiser was constructed as a scanning densitometer which made it an ideal choice for the standard calibration curve. Therefore, the grey levels in the images obtained with the remaining two scanners were converted to the grey level that would be expected if they were scanned with the LS scanner. The response curve itself for the LS scanner was given by

$$P_{\mathrm{LS}} = m_{\mathrm{LS}}(\mathrm{OD}_{\mathrm{Max}} - \mathrm{OD}) + b_{\mathrm{LS}} \tag{3.1}$$

where P_{LS} is the pixel grey level for the images (obtained with the Lumisys scanner), m_{LS} and b_{LS} the slope and intercept of the regression fit for the calibration data and $\mathrm{OD}_{\mathrm{Max}}$ is the darkest film the scanner is capable of identifying. This "darkness" of the film is quantified by the optical density and given by

$$\mathrm{OD} = \frac{\log_{10} I_{in}}{\log_{10} I_{tr}} \tag{3.2}$$

where I_{in} and I_{tr} is the incident and transmitted light intensity, respectively. Similarly, the calibration data and regression fits for the remaining two scanners, also supplied with the database, had the form

$$\begin{aligned} P_{\mathrm{HT}} &= \frac{\mathrm{OD} - b_{\mathrm{HT}}}{m_{\mathrm{HT}}} & (3.3) \\ P_{\mathrm{DBA}} &= 10^{m_{\mathrm{DBA}}\mathrm{OD}+b_{\mathrm{DBA}}} & (3.4) \end{aligned}$$

Therefore, the curves used for the conversion of the images from the HT and DBA format to the standard (Std or LS) format was given by

$$P_{\mathrm{Std}} = \begin{cases} P_{\mathrm{LS}}, & \mathrm{LS} \\ m_{\mathrm{LS}}\left[\mathrm{OD}_{\mathrm{Max}} - (m_{\mathrm{HT}}P_{\mathrm{HT}} + b_{\mathrm{HT}})\right] + b_{\mathrm{LS}}, & \mathrm{HT} \\ m_{\mathrm{LS}}\left\{\mathrm{OD}_{\mathrm{Max}} - \frac{1}{m_{\mathrm{DBA}}}\left[\log_{10}(P_{\mathrm{DBA}}) - b_{\mathrm{DBA}}\right]\right\} + b_{\mathrm{LS}}, & \mathrm{DBA} \end{cases} \quad (3.5)$$

and the values for the constants are given in Table 3.1

m_{Std}	=	1000	m_{HT}	=	0.00094568	m_{DBA}	=	1.07553
$\mathrm{OD}_{\mathrm{Max}}$	=	3.6	b_{HT}	=	3.789	b_{DBA}	=	4.80662
b_{Std}	=	495						

Table 3.1: Parameters for linear regression of x-ray digitisers, LS, HT and DBA

The calibration curves were taken as linear but this was not valid over the entire range, the most obvious non-linear regions appearing near the extremes of their dynamic ranges. The simplest correction for this effect was performed: any pixels falling outside the linear region were ignored. The linear part was delineated by the last points in the calibration data that lay within the 95% confidence limits of the regression fit. The interval common for all three scanners with a linear response curve occurred for optical densities in the range 0.5–3.0.

An unfortunate limitation of the data set was that all the normal cases were digitised using one specific scanner (DBA) while the majority of the abnormal cases were digitised on a different scanner (LS). The available scanner data allowed for our normalization procedure to compensate for differences in the scanner resolution (rebinning), differences in the contrast due to the *scanner* (using a standard calibration curve) and differences in the size of the optical density range for each increment in the grey level (ignoring non-linear part of response curve) and represents the extent of the scanner behaviour that can be corrected under typical working conditions. In addition, there was a limited amount of data in our study which can be used to test this assumption. A more rigorous test would likely require the acquisition of a set of images specifically for such a purpose.

Quantities such as the MTF and Wiener spectrum[2] are useful for comparisons between different systems and it may be conceivable to use this form of information to alter or correct an image to conform to the characteristics of a different system but it would be difficult and such a correction was not attempted. However, if the computer system was able to detect the scanner dependencies after all practical normalisations, then this would have serious implications for any automated system with a similar purpose. Essentially, a program would need to be tuned for each specific equipment configuration. Additionally, QC procedures, testing after maintenance, etc. would be necessary to ensure that even routine changes to the hardware would not interfere with the algorithms. It also makes comparisons between studies using different mammographic properties difficult if they were carried out at different institutions, with differing hardware systems. Further, a system which is capable of detecting minor variations due to the digitization would also be sensitive to variations in film and processor performance or changes in exposure conditions. These variations would

[2] Noise power spectrum

not, in general, be correlated to, say, normal and abnormal cases but if a system can detect these differences using properties calculated over the entire segmented image, measures to address this issue for any other studies of this nature would likely be necessary. Of course, these difficulties may be alleviated through the use of digital mammography but a detailed examination of the sensitivity of spectral and texture properties to scanner, processor and exposure conditions would still be valuable since film/screen mammography is likely to be the primary modality for mammographic screening for some time.

It is possible that some digitization effects remain in the images in spite of the efforts to remove them. However, the primary goal was to remove enough of the gross dependencies so that the discriminant methods would utilise the variation inherent to the imaged objects over any differences due to the x-ray scanner or exposure technique. In addition, more detailed normalization for local or anisotropic characteristics requires precise information that is often difficult to obtain. Therefore, the majority of the remaining work assumes that the scanner characteristics are not exhibited in the extracted properties but some testing of this assumption is given with the limited data that is applicable for this purpose.

Exposure Differences

Standard mammography units have a number of features which enable the technician to consistently produce high quality films with similar contrast characteristics. The goal is to produce the same average optical density film for varying breast thickness and x-ray quality. Therefore, it is difficult to accurately infer the tissue type from the optical density on a mammogram. To deduce the tissue type would require more detail on exposure conditions[3].

In correcting for these exposure effects a portion of the histogram at either extreme was ignored. Both the upper and lower regions that were ignored consisted of two parts, a fixed part and a part with a size that varied with the segmented image size. For the lower end of the histogram, the fixed grey level range was taken as 0–1095 and the variable region from 1096 to 1096 + 0.5% of the total number of segmented pixels. For the upper end of the histogram the corresponding regions were 3595–4096 and 3594 - 0.05% of the total number of segmented pixels to 3594. All these limit values were chosen empirically from an examination of a small number of images. The values were chosen so that for the high pixel values, contributions from noise, dust etc. were not included and for the low pixel values, any pixels outside the breast region were ignored. There were considerably more pixels belonging to the latter group than the former and this was reflected in the difference in the proportions that were ignored. The remaining values were then re-mapped to occupy the full range of possible pixel intensities. This procedure should make the most radiographically dense tissue in the thickest part of the breast and the least radiographically dense tissue in the thinnest region have consistent grey level values across different films with varying exposures.

In summary, the original images had a resolution of 42–50 μm/pixel and 4096 grey levels to give a typical file size of 25-30 Mb per image while the normalised images had a resolution of $\sim$ 110 μm/pixel and 4096 grey levels. After segmentation and cropping of the images, a typical image size was $\sim$ 1500x800 pixels and a file size from 5-10 Mb. In addition, the conversion of the pixel values,P, to optical density,

[3] This may also require the use of a calibration step wedge in each image. However, one difficulty with this method is that the step wedge must be placed close to (or in) the penumbra. It should also be noted that details such as the kVp, mAs, and breast thickness, was not provided in the database.

OD, is given by

$$\mathrm{OD} = \mathrm{OD}_{\mathrm{Max}} - \frac{P - b_{\mathrm{Std}}}{m_{\mathrm{Std}}} \tag{3.6}$$

3.2 Group Selection

The various properties calculated for the images were tested for their ability to differentiate between groups under two different classification schemes. The first used a recognised method of predicting cancer risk, the mammographic density (**Den** classification). The density grades were assigned by experienced radiologists following the guidelines outlined by the American College of Radiology Bi-RADS specification [ACR, 1993] and were provided along with the DDSM image database.

In order to preclude the possibility of detecting characteristics unrelated to the density classification, such as a characteristic indicative of a visible lesion, only the cases with both breasts diagnosed as normal were used. In addition, an arbitrary but consistent choice was made to use only the CC views of the mammograms for the left breast of the normal cases. The images were divided into a training and test set at random. Further, the effects of the distribution of the cases on the classification accuracy was explored by randomly re-assigning the full set of cases into a training and test set five times. This forms the dataset for classifying the density grades or the **Den** group.

The alternative classification, “diagnosis classification” (**Diag** classification) divided the patients into normal and abnormal groups. In this case, only the patients with both breasts evaluated as normal were assigned to the normal class since a normal breast contralateral to one with a malignancy may have a malignancy without any clinical signs at the time of the exam. In addition, the presence of a malignancy may produce some subtle influence on the appearance of both mammograms. The abnormal class consisted exclusively of those mammograms where a malignancy was diagnosed (i.e. the mammogram for the contralateral breast was ignored if it was cancer free.) The selection of the cohorts for this group was done at random from the collected pool of images. The images were divided into a training set and test set (at random) with the constraint that the number of cases in each category was roughly equal. It was found that some of the feature selection algorithms performed best when there were roughly equal numbers in each category and while it was possible to weight some categories more heavily than others this had little impact on the final result and balancing the number of cases in each category was significantly more effective[4].

The case selection procedure was then repeated 4 additional times to form a total of 5 training and test groups. The constraint on balancing the number of cases in each category combined with the random selection of cases resulted in few cases where mammograms from both the left and right breasts of the same patient appeared in the same sample. However, the restricted size of the entire pool of images resulted in considerable overlap in the cases between samples. The multiple samples were used to provide some insight into the amount of variation that can be expected due to redistribution of cases into the training and test groups.

There were additional cohorts selected for two specific purposes. Specifically, the results of the classification procedure on the normal contralateral mammograms (**Contra**) from the breast where a malignancy was diagnosed as well as those mammograms for the breasts contralateral to the "normal" cases were examined. For the

[4] It should be noted that the category equalisation was not applied to the **Den** class as it would have resulted in an unacceptably small sample.

normal cases these contralateral mammograms were also classified as normal, however, the contralateral breasts for the abnormal cases were clinically normal but had a considerably higher risk of developing breast cancer [Gajalakshmi et al., 1998].

The second cohort which was created was used to examine the age dependence in the classification results. Since the risk of breast cancer is influenced by the age of the patient, an examination of the age dependence in the mammographic properties which were selected was also performed. For this part of the study the images from the **Diag** classification were divided into sub-groups of patients with ages from 40–54, 42–56,· · ·, 54–68. Again, within each sub-group 5 random selections of a training and test group were made.

3.3 Feature Extraction

Three spectral properties were considered: regional moments, global moments and the mean for subregions of the histogram of the segmented breast tissue. For the cases that had a small total number of variables, as was the case with the global and regional moments, an exhaustive search of all possible combinations of variables was performed and the best selected, based on the results from a linear discriminant analysis. Indeed, this approach was carried out for all cases where it could be performed in a reasonable amount of time. However, an exhaustive search was not possible when the total number of variables exceeded $\sim$ 50. Therefore, for these situations, a method of performing a random search of a large feature space was used to identify the most significant properties, a genetic algorithm, as implemented in **ga_ors**. This approach was used to select the sub-regions of the histogram and in the selection of textures.

3.3.1 Moments

The global moments of the image were found by calculating the histogram of the entire segmented region and applying Equation (2.2). On the other hand, the calculation of the regional moments followed the procedure described in [Byng et al., 1996a]. Briefly, the segmented region was tessellated with square 29x29 pixel^2 ($\sim$ 3x3 mm^2) regions of interest (ROI's), ignoring any ROI's which were not entirely contained in the segmented region. The moments were then calculated from the histogram for each ROI and the corresponding moments averaged together. Byng *et al.* found a specific regional moment, the third or regional skewness, to be useful in the classification of density grades. Therefore the use of the regional skewness alone was considered as a separate case in the analysis.

3.3.2 Histogram Regions

The selection of the regions of the histogram with the greatest discriminatory power was conducted primarily through the program **ga_ors** from the Institute for Biodiagnostics. The program selects a number of "best" regions up to a user-defined maximum. If too many regions were used, many may be positioned in areas which have little discriminatory power thus producing a more complex discriminant function with little improvement in classification performance. Another concern was that an increase in the number of variables in the discriminant function also increases the likelihood of tailoring the discriminant function to characteristics that were discriminatory only in the sample used as the training set (overfitting). Naturally, if the number of regions was too small, an insufficient number of characteristics were being used and it would not be possible to accurately reflect the structure in the data.

Since it is not possible to know *a priori* the ideal number of features, the analysis was repeated using several numbers of regions: 2–7, 10 and 15.

Some preprocessing was performed on the histogram data prior to submitting it to `ga_ors`. First, the pixel counts in the histograms were rank ordered, that is the values were replaced by their ranking in the total histogram. Hence, the maximum count was replaced by 4095, the total number of grey levels, the next highest by 4094, and so on to 0. For grey levels with equal pixel counts, the rank was decided at random. The rank ordering had two effects, first it removed a non-zero background and second it tended to provide some protection from biasing the region selection with only large peaks, which may or may not be discriminatory. For this work, whether the rank ordering was or was not used did not have any significant impact on the resulting accuracy of the method but it improved the stability of the selected regions with respect to random fluctuations. Since a genetic algorithm was driven by a random number generator, it was not surprising that the regions selected varied even with identical input[5]. However, applying the rank ordering reduced the variation in the selected regions over multiple trials and the stability was retained for a higher number of selected regions.

One difficulty that must be dealt with was that the full histogram contained 4096 discrete grey levels and the program can select a region with a minimum of two channels. The information contained in such a small region may be dominated by noise. In order to combat this effect, the size of the histogram was reduced by dividing the full range of grey levels into small intervals and taking the median of the histogram data for each interval. The width of the intervals was selected as a

[5] This may not be true if there was a single well defined optimum solution to the problem but that was not the case here. There appeared to be multiple solutions of comparable effectiveness.

compromise between a number of competing factors. For example, if the intervals were too small the noise in each interval may be too severe and lead to poor classification. On the other hand, if the interval was too large, the regions which were actually significant may be removed from the pool of available variables. Additionally, when the reduced data was then processed by **ga_ors** for the selection of subregions in this rebinned histogram and if the window was too large, the program frequently chose portions of the histogram which were as small as possible (two rebinned grey levels). The performance of the system was evaluated after combining: 1, 2, 4, 8, 16, 32 and 64 grey levels and selecting the best 5 regions when the **Diag** cohort was used. As shown in Figure 3.2, a reasonable choice for the window size seemed to be 16 as it is the largest bin size with comparable classification accuracy to the raw histogram data.

3.3.3 Multifractal Features

In order to evaluate the generalised dimensions, the partition function given in Equation (2.3)

$$\chi_q(\varepsilon) = \sum_{i,j}^{N(\varepsilon)} \mu_{ij}^q \,, \quad q \in \mathbf{R} \tag{3.7}$$

was evaluated for multiple values of q. This required superimposing a regular grid with a characteristic length ε over the image and calculating μ_{ij} for each cell. To calculate μ_{ij}, an average grey level for each cell was found and truncated to its integral value. μ_{ij} was then taken as the fraction of the segmented image which contained pixels with that truncated, average grey level. In other words, for a mesh size of ε x ε pixels^2 the average for each cell was found and a histogram was calculated for just the segmented region with the reduced resolution. Hence, if an average grey level, $\overline{P}_{ij}$, was found for

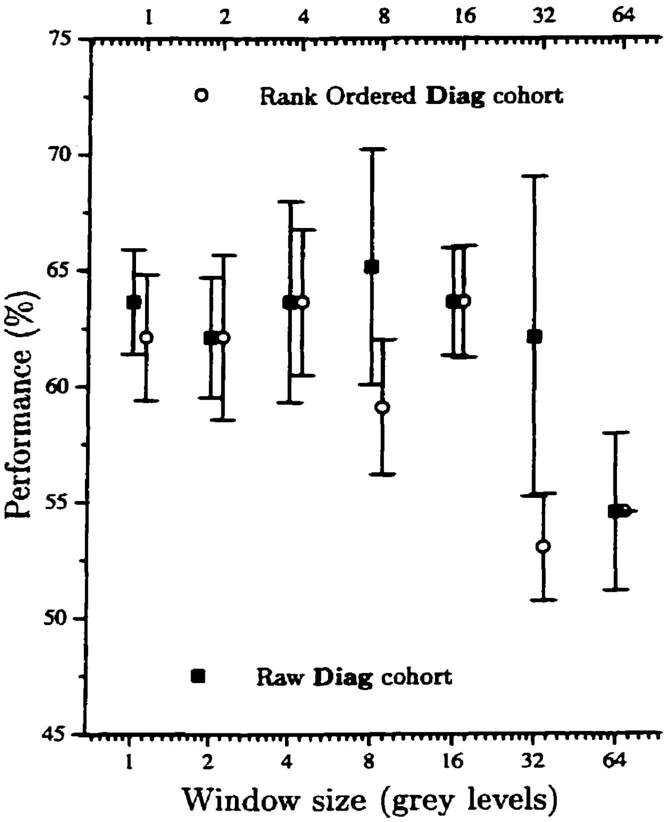


Figure 3.2: Percentage of correctly classified cases with 5 GA selected histogram subregions after rebinning the histogram and applying rank ordering. The upper horizontal axis refers to the rank order data and the lower axis to the data without rank ordering. The error bars reflect the standard deviation of the five redistribution trials of the cohort.

cell (i, j), μ_{ij} was taken to be the number of cells with an average grey level of $\overline{P}_{ij}$, normalised by the total number of cells contained in the segmented breast shadow[6]. The histogram was recalculated for every change in scale ε. In addition, for every choice of q, χ_q was found for $\varepsilon = \{1, 3, 5, 9, 17\}$ and the linear regression fit calculated for $\log \chi_q$ as a function of $\log \varepsilon$. The slope of the fit gives τ from which D_q can be easily calculated from Equation (2.6). A range for ε was selected after applying the procedure to a small subset of images and selecting a region such that $\log \chi_q \sim C \log \varepsilon$, for some constant C, which indicates that the fractal model is valid.

It was found that the multifractal model grew progressively less applicable as $q \gtrsim 1$. This may be due to the limitations of the double precision libraries that were used for the calculation of χ and μ^q and the rapid growth of μ^q as q increased.

It was possible to accommodate the wide range of values through the use of infinite precision math libraries but these routines were computationally very intensive. Further, the purpose of the study was not to demonstrate the multifractal nature of a mammogram. Rather, a set of parameters which may be used to characterise the texture of the image was desired. For this purpose it was not significant that the calculated quantities exactly reflected the value of the "true" generalised fractal dimensions for the images and simply a fixed range of ε values was used along with the standard double precision arithmetic for the calculations. Therefore, 20 evenly spaced values were selected for q from -5 to 0.7. i.e. $q \in \{-5, -4.7, -4.4, \cdots, 0.1, 0.4, 0.7\}$ and combinations of features which consisted primarily of D_q far from $q = 0$ were rejected[7].

[6] This is also the histogram entry for the $\overline{P}_{ij}^{\text{th}}$ grey level normalised by the total area under the histogram curve.

[7] These dimensions were the most likely to be far from the "true" value for the generalised fractal dimension.

3.3.4 Texture Features

For this work, the wavelet transform of the image was performed using a publicly available library: liftpack [Sweldens, 1994, Sweldens, 1995, Fernández et al., 1996]. The library was obtained from a web site[8] and used a method called lifting which enabled a bi-orthogonal wavelet basis to be generated after specifying a few desired properties. In particular, a basis described by Sweldens [Sweldens, 1994] was used. The wavelet coefficients were retained for 5 levels in the decomposition. The textures were then calculated for the areas which corresponded to the segmented breast region in the three "high pass" filtered quadrants[9] at each level of the decomposition[10].

For each quadrant the texture energy, entropy and inertia were calculated. All three texture measures can be found using the SGLD matrix, with elements, $p_{ij}(d, \theta)$ (as described in Section 2.2.2). Recall $p_{ij}(d, \theta)$ gives the probability of finding a pair of pixel values i and j separated by a distance d and with an orientation characterised by an angle θ. The SGLD matrix was found directly from its definition — examining each pair of pixels in the segmented image with the desired separation distance and orientation. It should be noted that any pair of pixels that did not have both points within the segmented tissue was ignored.

The choice of d and θ were arbitrary and 20 different combinations were selected corresponding to the distance and direction for the cartesian vectors given by:

$$\{(0,l),(l,0),(l,-l),(l,l)\}, l = \{1,2,4,8,16\} \tag{3.8}$$

[8] `http://www.cs.sc.edu/~fernande/liftpack/index.html`

[9] For example, the three high pass filtered quadrants at the highest resolution level were labelled as HL, HH and LH in Figure 2.3.

[10] Each "level" of the wavelet decomposition is identified by a different number of letters in Figure 2.3. Therefore {HL, HH, LH} and {LLHL, LLHH, LLLH} represent different levels of the decomposition.

Using different values for d and different levels in the wavelet transform may appear to be redundant. However, in [Wei et al., 1995] a comparison was made between changing texture resolutions by using different levels in a wavelet transform and by changing the values for d. In their work, the wavelet transform method had comparable or better performance. Since the study by Wei was designed to reduce the false positive rate for a CAD system and considered different features than those used in this thesis, there was no justification to choose one method to the exclusion of the other. Hence, both methods of examining multi-resolution texture features were utilised.

The different levels in the transform combined with the three quadrants in each level and the various combinations of d and θ resulted in 300 features per texture that were available for the discriminant function. The large number of values prohibited the use of an exhaustive search of all possible texture combinations. Therefore, the genetic algorithm was used for the selection of the best textures as well as the best subregions of the histogram. The program requires a spectrum as input hence a method is needed to map the collection of texture values into some form of "spectrum". The mapping was arbitrary and the texture values were placed sequentially for the textures corresponding to the directions given in Equation (3.8) to form groups for each quadrant of the wavelet transformed image. These groups were then collected by the order of the quadrants from which the textures originated, starting from the quadrant in the lower left and proceeding counterclockwise for each level (Figure 2.3). Next, each level group of features was abutted in the "spectrum" from the lowest resolution to the highest. Finally, each texture value was copied 10 times prior to inserting the next texture feature. The program, **ga_ors**, identifies regions with a mean that was useful for the classification problem and it is unlikely that an average of several texture features would correspond to a physical quantity. In principle, the repeated sequence of values, which was quite large relative to the minimum region size of two bins, would inhibit the program from selecting a region spanning several textures.

Clearly, the order of the textures in the artificial spectrum is hierarchical. As an example consider a texture such as the energy. The texture value for the highest level of the wavelet transform occupied the bins numbered 2401–3000, the next highest level 1801–2400, etc. Within each level, say 2401–3000, 2801–3000 was used for the textures in the quadrant labelled HL in Figure 2.3, then 2601–2800 for HH and 2401–2600 for LH. Then within each quadrant, eg. 2801–3000, 2961–3000 contained the textures for the five different combinations of d and θ corresponding to $l = 16$ in Equation 3.8, 2921–2960 for $l = 8$, etc. Within each group of bins, eg. 2961–3000, 2991–3000 contained textures for the specific combination of (d, θ) corresponding to the vector[11] (16,16), 2981–2990 for (16,-16), 2971–2980 for (16,0), and so on. Finally, within each of these smallest groups, eg. 2991–3000, the same texture value was repeated for each bin. See Figure 3.3.

[11] $(l, l), l = 16$ from Equation 3.8

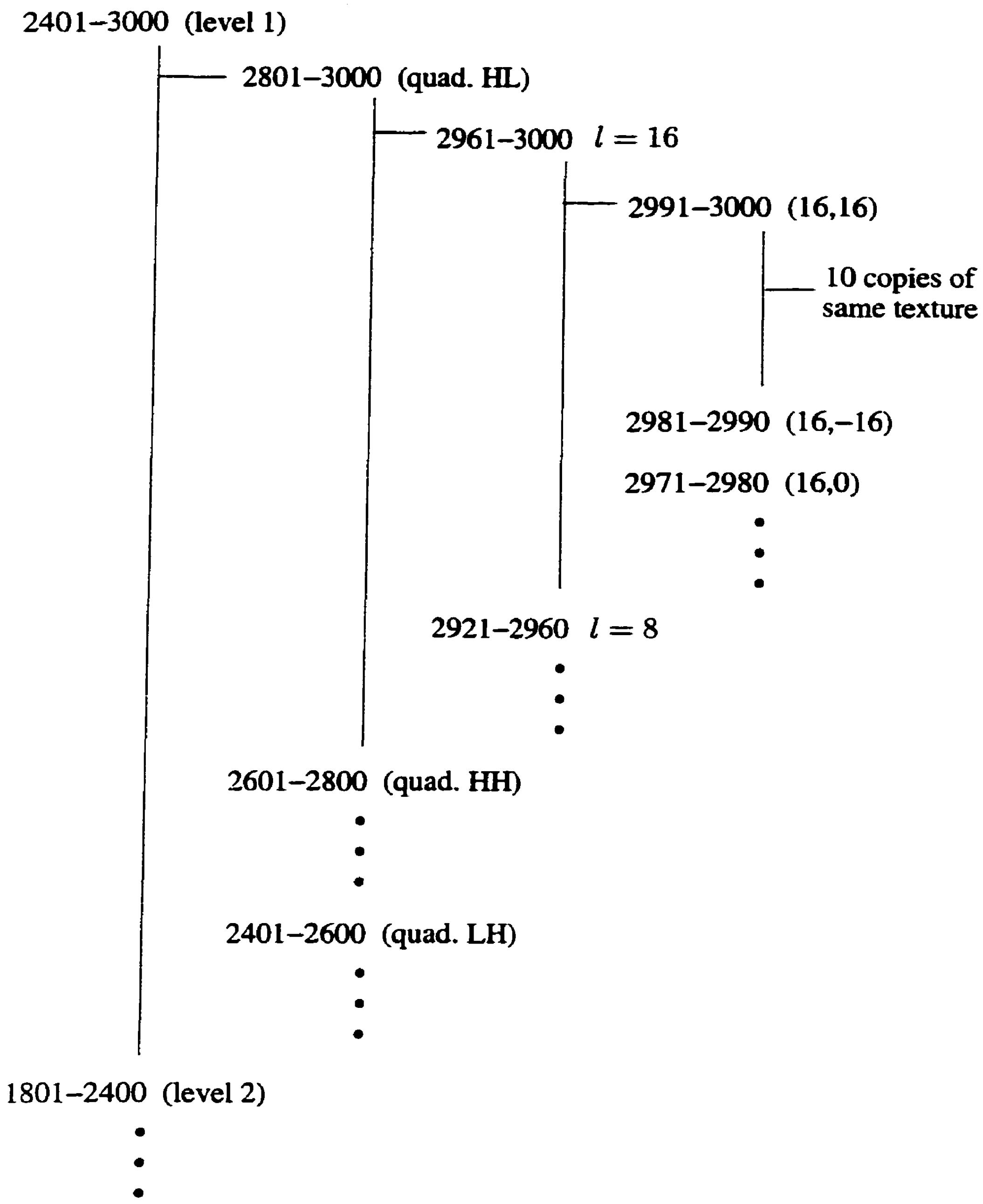


Figure 3.3: Hierarchical mapping of texture features of the wavelet transformed image to a spectrum for the genetic algorithm. "level" refers to the scaling level in the wavelet transform. "quad." refers to the quadrant HL, HH or LH in each level. (See Figure 2.3.) "l" refers to Equation 3.8. Finally, the vectors refer to Equation 3.8 with particular instances of l.

Chapter 4

Results

The results were organised into two main divisions. The first part describes the features selected that maximise the classification performance for the primary cohorts, datasets based on mammogram density classes (**Den**) and datasets based on the diagnosis (**Diag**), along with any conclusions which can be drawn from the nature of the selected properties. The second part evaluated the selected properties on various subsets of the data. In particular, we examined:

1. The performance of the selected features to classify the normal images contralateral to the breast with a diagnosed malignancy and the mammograms contralateral to the normal mammograms used for the **Diag** cohorts.

2. The correlation between the patient age and the selected features.

3. The correlation between the scanner used to digitise the mammogram and the selected features.

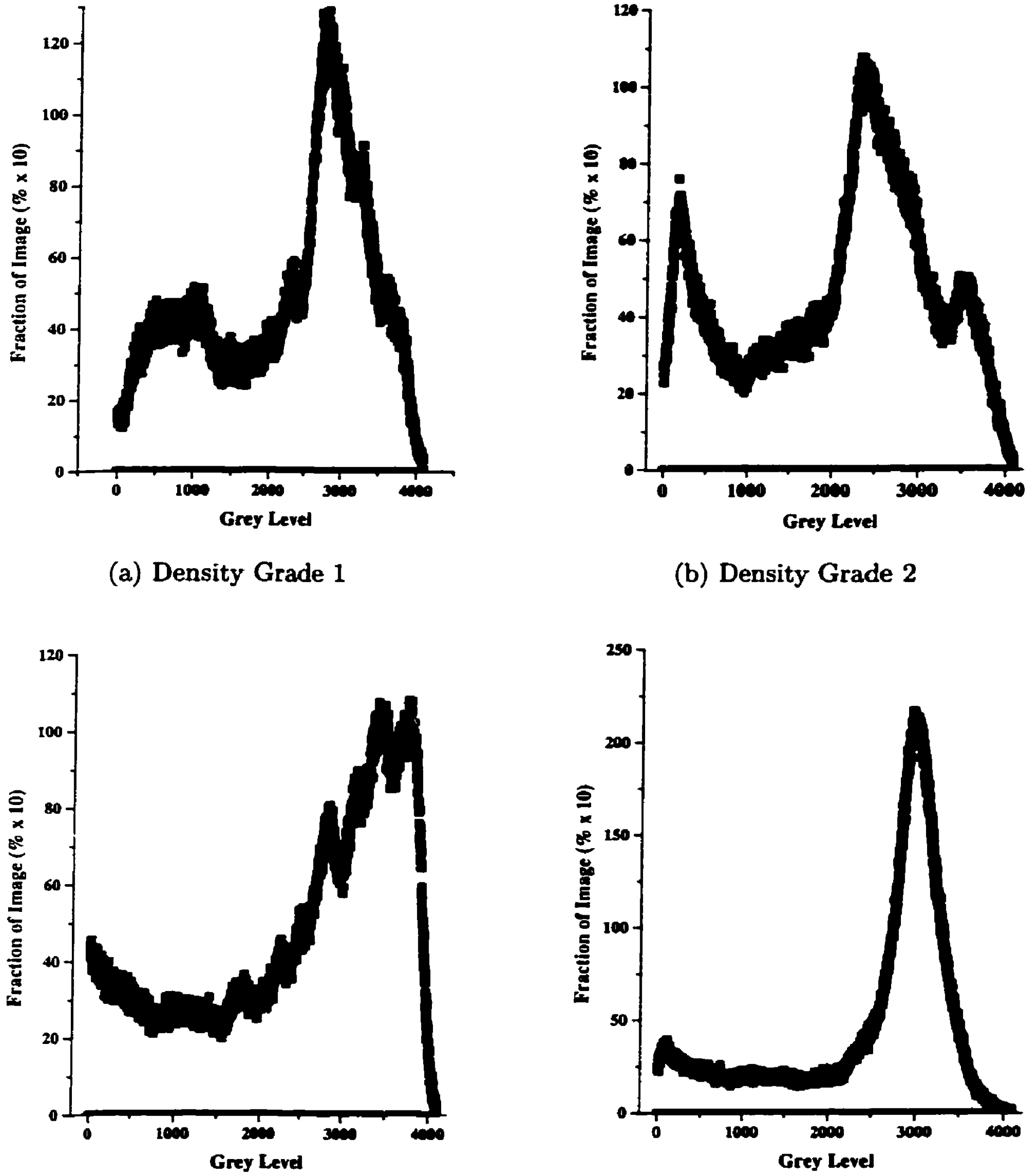
4.1 Feature Selection and Classification

The features which were selected as the most significant varied considerably depending on the classes in the cohort which were to be distinguished. In this work, two main classification divisions were considered: classification into density grades (**Den**) and classification into diagnosis classes (**Diag**).

4.1.1 Density Grade Classification

For this classification scheme, the mammograms were evaluated into four density grades by experienced clinicians following the guidelines specified by the ACR (Bi-RADS). This was performed at the centre where the mammogram was taken and supplied along with the patient information in the DDSM. The database was organised along the lines of *confirmed malignancy*, *normal* or *abnormal but benign* classes. Therefore, the distribution of cases in each density grade reflected the distribution of the general population and few cases were present with the lowest density grade. A typical histogram for each density grade is shown in Figure 4.1. Most of the normalisations have already been applied prior to the extraction of the histograms, including the normalisation for variations in image size. Indeed, the large number of 0 values is due to the extension of the grey levels to occupy the full 4096 range. The normalisations that were excluded were used exclusively for the sub-regions of the histogram, i.e. rank ordering of the histogram and smoothing by only retaining the median value of each 16 successive grey levels.

For the study of density grades, it was advisable to organise the analysis so as to guard against detection of features which were characteristic of a malignancy. Therefore mammograms were used for only those patients who were free of cancer.



(a) Density Grade 1

(b) Density Grade 2

(c) Density Grade 3

(d) Density Grade 4

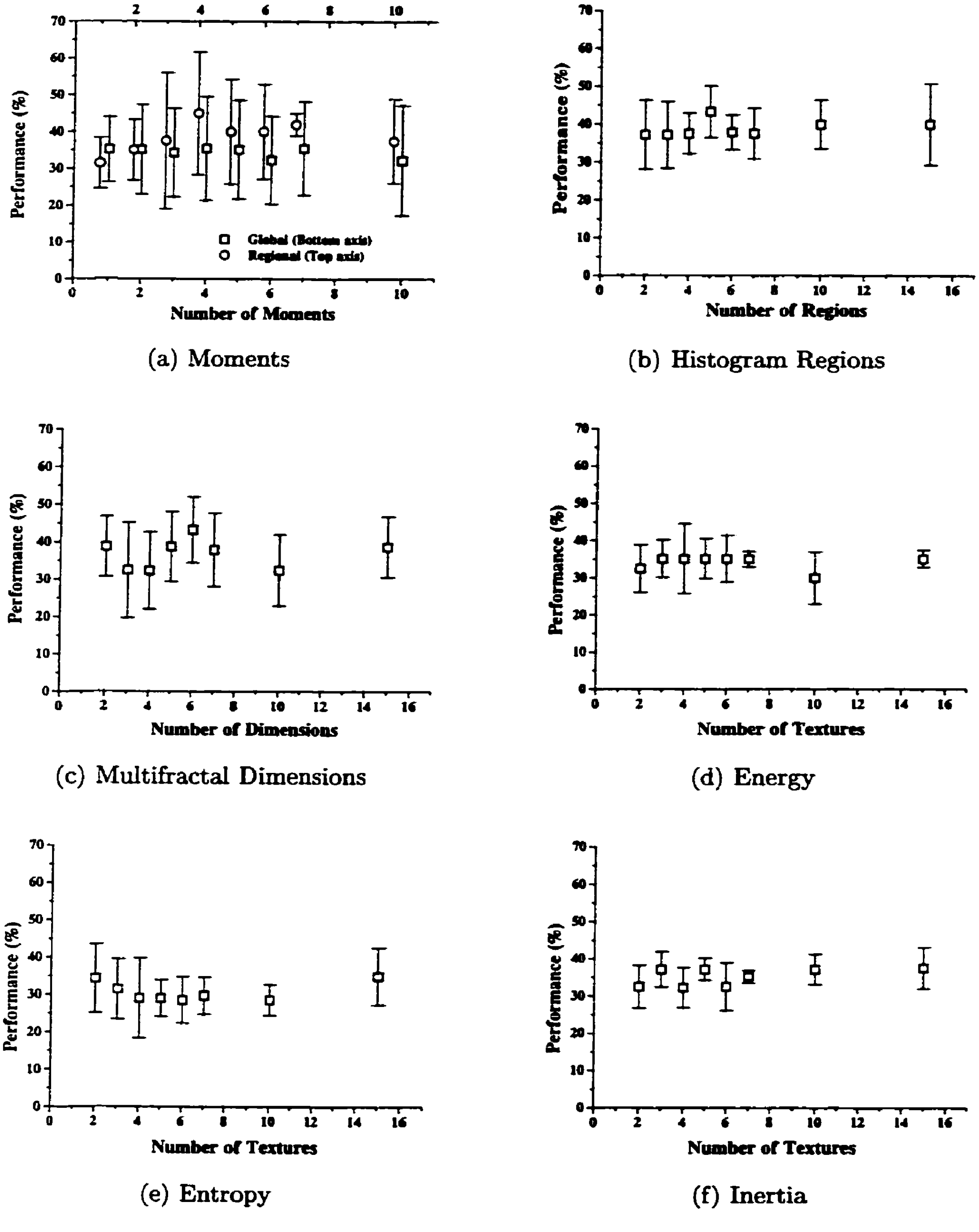
Figure 4.1: Typical examples of the histograms for a mammogram from each density grade. The images were normalised prior to extraction of the histograms.

In addition, in order to avoid subtle biases, the same side (left) was used for all cases. The patient ages span the full range under consideration, 40–69, and the cases were randomly divided into a training and test set with the training cases containing roughly twice the number in the test set. To reduce and quantify effects due to the choice of which images were in the training and test sets, four additional random divisions of the same pool of images into training and test sets were selected. This allowed an exploration of the effects of the distribution of cases on the classification accuracy.

The overall performance, that is the percentage of cases classified correctly, for each feature when considered individually is shown in Figure 4.2. The plotted points represent the median of the classification performance while the error bars showed the standard deviation of the five random training and test set selections. Since the standard deviation was calculated for a small group of values, the error bars in the following figures may not be representative of the variability of the classification performance of the feature set. In some cases the error bars seemed unusually large or small.

A few general observations were immediately apparent. Most notably, the classification accuracy was approximately 40%[1] regardless of the property under consideration and regardless of the number of features used in the discriminant function. In addition, the range of values for most of the classification results were sufficiently large that all the properties can be taken as having comparable classification ability. The size of the standard deviation was noticeably smaller for the properties where the optimum set of parameters were selected by the genetic algorithm (histogram sub-

[1] A collection of mammograms assigned at random into 4 density grades would have been expected to have a ~ 25% classification accuracy.



(a) Moments

(b) Histogram Regions

(c) Multifractal Dimensions

(d) Energy

(e) Entropy

(f) Inertia

Figure 4.2: Overall classification performance (percentage of correctly classified cases) for the **Den** class.

regions, texture energy, texture entropy and texture inertia) rather than through an exhaustive search (global moments, regional moments and multifractal dimensions). This may indicate that the genetic algorithm is more robust to overfitting so while the exhaustive search may find a combination of properties with a better overall performance on the training set, the genetic algorithm gives a result which was more representative of the typical performance.

The relatively poor performance of these properties was equally likely to be due to the small sample size and unbalanced distribution of cases in the various density grades rather than from any deficiency in the approach. The organisation of the database resulted in a very uneven distribution of cases in the various density grades which can only be overcome by using a sufficient number of cases that would allow the selection of a statistically significant number of cases for each grade. The number of cases that constitute a sufficient number can be significantly different when normal/abnormal groups were considered and when density grades were considered. Since the classification of normal/abnormal groups was the primary goal, the sample size and distribution of cases was insufficient for density grade classification.

It is also important to note that the regional skewness was found to be a significant risk factor by Byng *et al.* ([Byng et al., 1996a], [Byng et al., 1999] and [Byng et al., 1996b]) and the classification accuracy on our sample using this property, $(34 \pm 15)\%$, was comparable to any of the other features that were considered here (Tables 4.1–4.8). Each table gives the number of features used in the discriminant function (leftmost column) as well as the "confusion matrix" or the distribution of the predicted classes (rows) as a function of the known classes (columns) for each trial and the overall performance, ie. the fraction of correctly classified cases, for each

No. of Regions	Trial A & Acc (%)					Trial B & Acc (%)					Trial C & Acc (%)					Trial D & Acc (%)					Trial E & Acc (%)				
	0	0	0	0		0	3	0	0		0	0	2	1		0	1	0	2		0	1	2	0	
2	0	5	6	2	32.43	0	5	0	4	35.14	0	0	10	3	25.00	0	3	1	0	22.58	2	7	0	0	40.00
	0	6	3	3		0	5	1	5		0	1	4	5		0	6	3	4		0	2	6	0	
	0	5	3	4		0	5	2	7		0	2	6	6		0	10	0	1		0	9	5	1	
	0	0	0	0		0	2	1	0		0	1	1	1		0	1	1	1		0	1	2	0	
3	0	8	3	2	35.14	0	5	2	2	37.84	0	1	9	3	30.00	0	3	1	0	16.13	2	4	3	0	28.57
	1	6	2	3		0	7	1	3		1	0	4	5		0	8	1	4		0	2	4	2	
	0	6	3	3		0	5	1	8		1	0	6	7		0	7	3	1		0	8	5	2	
	0	0	0	0		0	1	1	1		0	1	1	1		0	1	0	2		0	1	2	0	
4	0	6	6	1	29.73	0	4	3	2	32.43	1	1	7	4	25.00	0	2	2	0	16.13	2	4	2	1	22.86
	1	7	2	2		0	5	3	3		0	0	4	6		0	7	1	5		0	4	2	2	
	0	5	4	3		0	3	6	5		1	1	7	5		0	5	4	2		0	3	10	2	
	0	0	0	0		0	2	1	0		0	2	0	1		0	1	0	2		0	1	1	1	
5	2	7	2	2	32.43	0	4	3	2	35.14	2	5	3	3	40.00	0	4	0	0	25.81	0	5	2	2	37.14
	3	5	2	2		0	5	2	4		0	2	4	4		0	6	1	6		0	4	2	2	
	1	4	4	3		0	6	1	7		1	4	2	7		1	7	0	3		1	4	4	6	
	0	0	0	0		0	0	2	1		0	2	0	1		0	1	0	2		0	1	0	2	
6	1	5	5	2	32.43	0	6	2	1	35.14	1	5	4	3	40.00	0	2	1	1	22.58	2	4	3	0	31.43
	0	7	4	1		0	6	2	3		0	3	4	3		0	7	3	3		0	4	4	0	
	0	4	5	3		1	5	3	5		1	3	3	7		1	7	1	2		1	5	6	3	
	0	0	0	0		0	2	1	0		0	1	1	1		0	2	0	1		0	1	1	1	
7	3	4	4	2	29.73	1	4	3	1	37.84	2	5	3	3	37.50	0	3	1	0	25.81	1	5	3	0	40.00
	2	4	4	2		0	3	3	5		0	1	5	4		0	6	2	5		0	4	2	2	
	1	5	3	3		0	3	4	7		1	2	6	5		2	4	2	3		0	2	6	7	
	0	0	0	0		0	3	0	0		0	0	2	1		0	1	0	2		0	1	2	0	
10	1	6	5	1	32.43	0	4	2	3	29.73	0	4	7	2	30.00	0	2	1	1	25.81	2	4	3	0	34.29
	0	9	2	1		0	4	4	3		0	3	4	3		0	8	2	3		0	2	4	2	
	0	4	4	4		0	8	3	3		1	2	7	4		1	3	3	4		0	7	4	4	
	0	0	0	0		0	3	0	0		0	0	3	0		0	1	0	2		0	1	1	1	
15	2	4	5	2	29.73	1	4	2	2	29.73	1	4	6	2	37.50	1	2	1	0	22.58	1	4	4	0	28.57
	3	5	4	0		0	4	5	2		1	3	4	2		0	9	2	2		0	3	2	3	
	0	5	4	3		1	9	2	2		3	2	2	7		1	4	3	3		1	5	5	4	

Table 4.1: Outcome for **Den** cohort using sub-regions of the histogram. For each trial the overall classification performance and confusion matrix is given.

trial. Each trial (A–E) represent the five repeated divisions of the entire sample into training and test sets.

The single best regional moment was found to be the second regional moment (regional variance). Its overall classification performance was comparable to the regional skewness, with a smaller uncertainty, but the lower variance may be a small sample size effect. For a direct comparison between all properties that were under consideration see Tables 4.1–4.8.

From Table 4.1 (only histogram sub-regions) the best classification[2], 37.5%, occurs

No. of Mom.	Trial A and Acc (%)	Trial B and Acc (%)	Trial C and Acc (%)	Trial D and Acc (%)	Trial E and Acc (%)
	0 0 0 0	0 1 2 0	0 1 2 0	0 2 0 1	0 1 2 0
1	0 11 1 1 51.35	0 6 3 0 35.14	0 10 3 0 40.00	0 1 0 3 32.26	0 7 2 0 28.57
	0 4 1 7	0 4 7 0	0 4 6 0	0 4 0 9	0 5 3 0
	0 3 2 7	0 4 10 0	0 1 13 0	0 2 0 9	0 4 11 0
	0 0 0 0	0 1 2 0	0 1 2 0	0 1 0 2	0 1 2 0
2	0 11 0 2 56.76	0 6 3 0 35.14	0 10 1 2 45.00	0 1 2 1 32.26	0 6 3 0 25.71
	0 4 3 5	0 4 7 0	0 4 4 2	0 4 2 7	0 5 3 0
	0 3 2 7	0 4 10 0	0 1 9 4	0 2 2 7	0 3 12 0
	0 0 0 0	0 1 2 0	0 1 2 0	0 1 0 2	0 1 1 1
3	0 11 0 2 51.35	0 5 1 3 32.43	0 10 1 2 52.50	0 1 3 0 25.81	2 5 2 0 34.29
	0 5 1 6	0 3 2 6	0 3 3 4	0 5 2 6	0 4 4 0
	0 4 1 7	0 4 5 5	1 1 4 8	0 3 3 5	0 3 9 3
	0 0 0 0	0 1 2 0	0 2 1 0	0 1 0 2	0 1 1 1
4	0 10 1 2 59.46	1 5 0 3 27.03	0 10 1 2 52.50	1 2 0 1 35.48	1 5 3 0 31.43
	0 3 6 3	1 3 1 6	0 3 3 4	0 3 4 6	0 5 3 0
	0 4 2 6	0 4 6 4	0 3 3 8	1 2 3 5	0 3 9 3
	0 0 0 0	0 1 1 1	0 2 0 1	0 1 0 2	0 1 2 0
5	0 11 1 1 56.76	0 5 0 4 35.14	0 10 2 1 52.50	0 2 1 1 29.03	1 5 3 0 28.57
	0 4 3 5	0 3 2 6	0 3 4 3	0 6 2 5	0 5 1 2
	1 3 1 7	0 4 4 6	0 1 6 7	1 3 2 5	1 3 7 4
	0 0 0 0	0 1 1 1	0 2 1 0	1 0 0 2	0 1 2 0
6	2 9 1 1 56.76	0 4 1 4 29.73	0 9 2 2 42.50	2 1 0 1 32.26	2 5 2 0 28.57
	0 2 7 3	0 3 2 6	0 3 2 5	0 3 3 7	0 5 1 2
	1 3 3 5	0 4 5 5	1 1 6 6	1 2 3 5	1 3 7 4
	0 0 0 0	0 1 2 0	0 2 1 0	1 1 0 1	0 2 1 0
7	1 9 1 2 59.46	0 5 1 3 32.43	0 10 2 1 47.50	1 1 1 1 35.48	1 5 2 1 28.57
	1 1 8 2	0 3 2 6	0 2 4 4	0 3 4 6	0 5 1 2
	1 3 3 5	0 3 6 5	1 2 6 5	1 3 2 5	0 2 9 4
	0 0 0 0	0 2 1 0	0 2 1 0	1 1 0 1	0 1 2 0
10	1 10 1 1 59.46	0 5 0 4 29.73	0 9 3 1 47.50	1 1 0 2 32.26	2 4 2 1 22.86
	0 2 8 2	0 3 1 7	0 1 3 6	0 4 3 6	0 4 2 2
	0 2 6 4	0 3 6 5	1 1 5 7	1 3 2 5	1 5 7 2

Table 4.2: Outcome for **Den** cohort using global moments the histogram. For each trial the overall classification performance and confusion matrix is given.

when 7 subregions of the histogram, {77–79, 100–104, 107–108, 121–123, 124–126, 254–256, 257–262}[3], were used in the discriminant function. For the global moments (Table 4.2), the performance was slightly smaller at 35.5% for 4 and 7 global moments. Since we desire the simplest discriminant function possible, the best discriminant function occurs when using 4 global moments, $\{1, 4, 5, 11\}$ where the 11^{th} moment was

[2] The classification for a given number of variables used in the discriminant function was taken as the median of trials A–E in the tables. The best was then selected from the list of median values for the various numbers of features.

[3] Recall that every 16 successive grey levels in the full histogram were grouped together so that the range of possible values were 1–256, Section 3.3.2. The remaining 257–267 bins were 10 repeated copies of the patients' age.

the patients' age. Similarly, the best classification using regional moments (45% in Table 4.3) occurs when using the 4 regional moments {3, 4, 6, 8} and the multifractal dimensions (Table 4.4) gave a discriminant function with 43% (Table 4.5) of the cases classified correctly for 6 "dimensions" ($q = \{-3.2, -1.7, -1.4, -1.1, -0.8\}$ and the patient's age.). When the texture energy was used, comparable classification accuracy (35%) appeared when using 3, 4, 5, 6 and 15 different sets of textures. The sample size for the density classification was likely insufficient to allow one discriminant function to clearly have better performance than the others. However, for the function with the fewest number of textures, 3, the energy textures used were {129–180, 1132–1161, 1383–1385}. For the texture entropy (Table 4.6), the best classification accuracy (35%) occurred for 15 textures. The use of 15 textures is excessive and comparable performance (34%) occurred for 2 textures, {765–819, 1378–1402}. Similarly, the highest accuracy when using the texture inertia (37.5%, Table 4.7) appeared for 15 textures but the use of 3 textures gave comparable results (37%) with the textures {900–1084, 1612–1620, 2721–2808}.

In order to be able to make some connection to previous studies in this area, Pearson's correlation coefficient [Bevington, 1969], r, was calculated for the results given in all the previous tables. The correlation coefficient is given by

$$r = \frac{N\sum_{i=1}^{N} x_i y_i - \left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right)}{\left\{\left[N\sum_{i}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2\right]\left[N\sum_{i}^{N} y_i^2 - \left(\sum_{i=1}^{N} y_i\right)^2\right]\right\}^{\frac{1}{2}}} \tag{4.1}$$

where each case, i, in the sample of N cases had a classification predicted by the LDA of x_i and a known classification of y_i. Then, when the correlation coefficient was used to quantify the performance, the best overall classification occurred when 7

No. of Mom.	Trial A and Acc (%)					Trial B and Acc (%)					Trial C and Acc (%)					Trial D and Acc (%)					Trial E and Acc (%)				
	0	0	0	0		0	3	0	0		1	1	1	0		0	1	0	2		0	3	0	0	
1	1	10	0	2	43.24	0	9	0	0	24.32	0	6	7	0	30.00	0	2	0	2	32.26	0	6	3	0	31.43
	0	5	0	7		0	11	0	0		0	5	5	0		0	8	0	5		0	3	5	0	
	0	6	0	6		0	13	1	0		0	4	10	0		0	3	0	8		0	9	6	0	
	0	0	0	0		0	3	0	0		1	1	0	1		0	2	0	1		0	2	1	0	
2	1	9	0	3	45.95	0	5	4	0	24.32	0	7	5	1	35.00	0	1	0	3	29.03	0	6	2	1	37.14
	0	4	2	6		0	6	3	2		0	4	5	1		0	3	0	10		0	4	4	0	
	0	3	3	6		0	6	7	1		0	4	9	1		0	3	0	8		0	5	7	3	
	0	0	0	0		0	3	0	0		0	1	1	1		0	3	0	0		0	2	1	0	
3	3	9	0	1	54.05	0	5	4	0	27.03	0	7	3	3	37.50	1	1	0	2	6.45	0	6	0	3	45.71
	1	2	2	7		0	6	3	2		0	2	4	4		0	6	0	7		0	4	4	0	
	0	3	0	9		0	6	6	2		0	3	7	4		0	9	1	1		0	5	4	6	
	0	0	0	0		0	3	0	0		0	2	0	1		0	3	0	0		0	2	1	0	
4	3	9	0	1	51.35	0	5	3	1	32.43	1	9	2	1	45.00	1	1	1	1	9.68	0	6	0	3	45.71
	1	2	2	7		1	5	4	1		0	1	7	2		1	5	1	6		0	4	3	1	
	0	4	0	8		0	6	5	3		0	4	8	2		0	9	1	1		0	5	3	7	
	0	0	0	0		0	3	0	0		0	2	0	1		0	3	0	0		1	2	0	0	
5	3	10	0	0	51.35	2	5	2	0	40.54	2	9	2	0	40.00	1	1	2	0	12.90	1	5	3	0	34.29
	1	2	2	7		0	5	5	1		0	3	4	3		1	5	2	5		1	3	4	0	
	0	5	0	7		0	2	7	5		0	1	10	3		0	8	2	1		0	5	8	2	
	0	0	0	0		0	3	0	0		0	2	0	1		0	2	0	1		0	3	0	0	
6	3	9	1	0	54.05	2	5	2	0	40.54	2	9	1	1	37.50	1	2	1	0	19.35	0	5	1	3	45.71
	0	2	3	7		0	5	5	1		0	3	3	4		1	5	3	4		1	3	3	1	
	0	3	1	8		0	2	7	5		0	1	10	3		0	9	1	1		0	3	4	8	
	0	0	0	0		0	3	0	0		0	3	0	0		0	3	0	0		0	2	1	0	
7	3	10	0	0	45.95	2	4	3	0	37.84	2	9	1	1	42.50	0	1	2	1	41.94	0	6	2	1	40.00
	1	2	1	8		0	5	4	2		0	3	5	2		0	4	4	5		1	2	5	0	
	0	3	3	6		0	2	6	6		0	1	10	3		0	2	1	8		0	2	10	3	
	0	0	0	0		0	3	0	0		0	2	0	1		1	0	0	2		0	2	0	1	
10	5	6	2	0	45.95	2	5	2	0	40.54	2	8	2	1	37.50	2	0	2	0	16.13	0	7	1	1	37.14
	1	2	6	3		0	6	3	2		0	4	3	3		4	3	1	5		1	4	2	1	
	0	2	5	5		0	5	2	7		0	5	5	4		4	3	1	3		0	4	7	4	

Table 4.3: Outcome for **Den** cohort using regional moments of the image. For each trial the overall classification performance and confusion matrix is given.

No. of Dim.	Trial A and Acc (%)		Trial B and Acc (%)		Trial C and Acc (%)		Trial D and Acc (%)		Trial E and Acc (%)	
	0 0 0 0		0 1 2 0		0 1 2 0		0 1 2 0		0 0 3 0	
2	0 11 1 1	54.05	0 5 2 2	35.14	0 9 3 1	42.50	0 2 0 2	38.71	0 7 1 1	34.29
	1 2 4 5		0 3 3 5		0 2 7 1		0 4 3 6		0 3 4 1	
	0 3 4 5		0 4 5 5		0 1 12 1		0 3 1 7		0 3 11 1	
	0 0 0 0		0 1 2 0		0 1 1 1		0 2 1 0		0 1 2 0	
3	1 9 2 1	45.95	0 5 2 2	32.43	0 10 3 0	52.50	1 1 0 2	32.26	2 4 2 1	20.00
	1 3 3 5		0 5 3 3		0 2 3 5		1 4 3 5		0 5 2 1	
	0 2 5 5		0 3 7 4		0 1 5 8		0 4 1 6		0 3 11 1	
	0 0 0 0		0 1 2 0		0 1 1 1		0 1 0 2		0 1 2 0	
4	0 10 1 2	43.24	0 5 2 2	29.73	0 10 3 0	52.50	0 1 3 0	32.26	1 5 3 0	28.57
	0 3 4 5		0 4 2 5		0 2 3 5		1 3 3 6		0 5 3 0	
	0 2 8 2		0 3 7 4		0 1 5 8		0 3 2 6		0 3 10 2	
	0 0 0 0		0 1 2 0		0 1 2 0		0 1 2 0		0 2 1 0	
5	1 10 0 2	45.95	0 5 3 1	35.14	0 10 3 0	52.50	1 1 1 1	38.71	2 5 1 1	28.57
	1 3 4 4		0 4 2 5		0 2 4 4		1 3 3 6		0 5 3 0	
	0 2 7 3		0 3 5 6		0 2 5 7		0 3 0 8		0 3 10 2	
	0 0 0 0		1 1 1 0		0 0 3 0		0 1 0 2		0 1 2 0	
6	1 10 0 2	45.95	0 5 3 1	43.24	0 9 4 0	45.00	0 1 0 3	29.03	2 5 1 1	28.57
	1 3 4 4		0 4 3 4		0 2 6 2		0 3 3 7		0 5 3 0	
	0 2 7 3		0 4 3 7		0 3 8 3		0 4 2 5		0 2 11 2	
	0 0 0 0		1 1 1 0		0 0 3 0		0 1 0 2		0 2 1 0	
7	1 10 0 2	48.65	0 5 3 1	37.84	0 10 3 0	47.50	0 1 0 3	32.26	2 5 1 1	25.71
	1 2 5 4		0 4 2 5		0 2 6 2		0 3 4 6		0 5 3 0	
	0 2 7 3		0 3 5 6		0 1 10 3		0 4 2 5		0 4 10 1	
	0 0 0 0		0 1 2 0		0 0 3 0		0 1 0 2		0 2 1 0	
10	3 7 3 0	40.54	0 4 2 3	32.43	0 9 2 2	45.00	0 1 0 3	32.26	3 4 2 0	20.00
	1 2 7 2		0 5 3 3		0 3 6 1		0 3 5 5		0 5 3 0	
	2 2 7 1		0 4 5 5		0 1 10 3		0 3 4 4		3 2 10 0	
	0 0 0 0		0 1 2 0		0 1 2 0		0 1 0 2		1 1 0 1	
15	3 9 0 1	43.24	0 4 3 2	32.43	0 8 5 0	40.00	0 1 2 1	38.71	2 4 3 0	22.86
	1 2 6 3		0 3 4 4		0 2 4 4		0 4 5 4		0 5 2 1	
	2 2 7 1		1 5 4 4		0 1 9 4		0 3 2 6		2 3 9 1	

Table 4.4: Outcome for **Den** cohort using multifractal dimensions. For each trial the overall classification performance and confusion matrix is given.

No. of Textures	Trial A and Acc (%)					Trial B and Acc (%)					Trial C and Acc (%)					Trial D and Acc (%)					Trial E and Acc (%)				
	0	0	0	0		0	1	0	2		0	2	1	0		0	2	0	1		0	2	1	0	
2	2	4	0	7	35.14	0	4	3	2	37.84	0	6	4	3	32.50	0	4	0	0	22.58	0	7	2	0	25.71
	0	6	1	5		0	4	6	1		0	2	5	3		0	9	0	4		0	6	2	0	
	1	2	1	8		0	3	7	4		0	2	10	2		0	7	1	3		1	8	6	0	
	0	0	0	0		0	1	1	1		0	3	0	0		0	2	0	1		0	1	2	0	
3	2	8	2	1	35.14	0	5	2	2	37.84	0	6	4	3	37.50	0	3	1	0	25.81	0	5	3	1	31.43
	1	7	0	4		0	4	5	2		0	2	5	3		0	7	2	4		0	3	4	1	
	0	3	4	5		0	7	3	4		0	2	8	4		0	6	2	3		1	3	9	2	
	0	0	0	0		0	1	0	2		0	3	0	0		0	3	0	0		0	1	2	0	
4	2	9	1	1	37.84	0	6	1	2	35.14	0	8	4	1	42.50	0	2	2	0	19.35	0	6	2	1	25.71
	2	6	0	4		1	5	4	1		0	4	2	4		1	5	1	6		0	6	1	1	
	1	4	2	5		0	7	4	3		0	2	5	7		0	6	2	3		1	5	7	2	
	0	0	0	0		0	1	0	2		0	3	0	0		0	2	1	0		0	1	2	0	
5	3	8	1	1	37.84	0	7	2	0	35.14	1	7	4	1	40.00	0	3	1	0	25.81	0	6	3	0	34.29
	1	6	1	4		1	5	2	3		0	1	5	4		0	7	1	5		0	5	2	1	
	1	5	1	5		0	5	5	4		0	2	8	4		0	4	3	4		1	6	4	4	
	0	0	0	0		1	1	0	1		0	2	0	1		1	1	0	1		0	1	2	0	
6	2	10	1	0	43.24	1	5	2	1	35.14	0	6	5	2	37.50	0	2	1	1	25.81	0	7	2	0	34.29
	2	5	2	3		1	4	3	3		0	4	4	2		0	8	1	4		0	6	1	1	
	1	5	2	4		0	5	5	4		1	2	6	5		0	4	3	4		1	5	5	4	
	0	0	0	0		0	1	1	1		1	2	0	0		0	1	0	2		0	2	1	0	
7	4	7	1	1	35.14	1	4	1	3	32.43	0	8	4	1	35.00	0	3	0	1	32.26	0	6	3	0	37.14
	3	3	2	4		1	5	2	3		0	5	2	3		0	6	3	4		0	4	3	1	
	2	4	2	4		0	4	4	6		2	2	7	3		0	6	1	4		1	4	6	4	
	0	0	0	0		1	1	1	0		1	2	0	0		0	1	1	1		0	2	1	0	
10	2	9	1	1	43.24	1	6	1	1	40.54	0	8	4	1	30.00	0	2	1	1	29.03	0	7	2	0	28.57
	0	3	2	7		1	3	4	3		0	2	2	6		0	5	3	5		0	4	2	2	
	1	2	4	5		0	4	6	4		1	4	8	1		0	4	3	4		2	6	6	1	
	0	0	0	0		0	1	1	1		0	2	0	1		0	1	2	0		0	1	0	2	
15	2	7	1	3	37.84	0	5	3	1	35.14	1	8	4	0	37.50	0	2	1	1	32.26	0	6	2	1	34.29
	2	4	3	3		0	4	5	2		1	4	3	2		0	6	4	3		1	3	2	2	
	2	3	3	4		0	5	6	3		2	1	7	4		1	4	2	4		2	4	5	4	

Table 4.5: Outcome for **Den** cohort using texture energy on the wavelet transform of the image. For each trial the overall classification performance and confusion matrix is given.

No. of Textures	Trial A and Acc (%)	Trial B and Acc (%)	Trial C and Acc (%)	Trial D and Acc (%)	Trial E and Acc (%)
	0 0 0 0	0 0 2 1	1 1 1 0	0 1 0 2	0 1 2 0
2	0 11 0 2 48.65	0 4 3 2 27.03	0 7 4 2 37.50	0 2 0 2 25.81	0 8 0 1 34.29
	0 5 1 6	0 4 1 6	0 3 6 1	0 8 0 5	0 5 2 1
	0 5 1 6	0 7 2 5	1 2 10 1	0 5 0 6	0 8 5 2
	0 0 0 0	0 0 3 0	1 1 0 1	0 1 0 2	0 1 2 0
3	0 8 3 2 37.84	0 3 1 5 27.03	0 8 3 2 45.00	0 2 0 2 25.81	0 7 1 1 31.43
	0 7 1 4	0 4 1 6	0 5 4 1	1 7 0 5	0 7 0 1
	0 3 4 5	0 4 4 6	1 4 4 5	0 4 1 6	1 7 3 4
	0 0 0 0	0 0 2 1	0 1 0 2	0 1 0 2	0 0 1 2
4	1 9 1 2 45.95	0 3 1 5 18.92	0 6 5 2 27.50	0 2 0 2 29.03	1 5 1 2 40.00
	1 6 2 3	0 6 1 4	0 5 4 1	1 5 2 5	0 6 1 1
	0 4 2 6	0 7 4 3	1 4 8 1	0 5 1 5	0 4 3 8
	0 0 0 0	0 1 0 2	1 1 0 1	0 1 0 2	0 1 1 1
5	1 7 4 1 37.84	0 4 1 4 27.03	0 7 2 4 32.50	1 2 0 1 29.03	1 4 3 1 25.71
	0 7 4 1	2 5 4 0	1 5 2 2	1 3 3 6	0 5 3 0
	0 6 3 3	1 6 5 2	1 3 7 3	1 6 0 4	0 3 10 2
	0 0 0 0	1 0 0 2	0 2 0 1	0 1 0 2	0 1 1 1
6	1 8 3 1 40.54	0 5 1 3 24.32	0 7 5 1 27.50	0 3 0 1 29.03	0 6 1 2 28.57
	0 6 4 2	1 7 2 1	0 4 3 3	1 4 1 7	1 4 3 0
	1 4 4 3	0 8 5 1	0 4 9 1	0 4 2 5	0 6 8 1
	0 0 0 0	0 1 0 2	0 1 0 2	0 1 0 2	0 1 1 1
7	1 9 1 2 37.84	0 6 0 3 29.73	0 7 4 2 30.00	0 3 0 1 25.81	1 4 3 1 25.71
	0 5 3 4	2 4 2 3	0 5 2 3	1 6 1 5	0 4 3 1
	0 8 2 2	0 3 8 3	1 4 6 3	1 5 1 4	0 5 8 2
	0 0 0 0	0 1 1 1	0 1 0 2	0 1 0 2	0 0 2 1
10	2 7 1 3 32.43	0 5 2 2 27.03	1 9 0 3 32.50	0 3 1 0 22.58	1 4 3 1 28.57
	2 3 3 4	1 4 4 2	1 5 3 1	1 5 1 6	1 4 3 0
	2 5 3 2	1 6 6 1	2 3 8 1	1 4 3 3	0 4 8 3
	0 0 0 0	1 0 2 0	0 1 1 1	0 2 0 1	0 1 2 0
15	2 8 1 2 43.24	0 5 1 3 37.84	0 8 2 3 35.00	0 3 0 1 25.81	1 4 2 2 25.71
	1 4 4 3	2 2 5 2	1 4 3 2	1 6 3 3	0 4 3 1
	1 6 1 4	1 6 4 3	3 4 4 3	1 5 3 2	0 3 10 2

Table 4.6: Outcome for **Den** cohort using the texture entropy on the wavelet transform of the image. For each trial the overall classification performance and confusion matrix is given.

No. of Textures	Trial A and Acc (%)	Trial B and Acc (%)	Trial C and Acc (%)	Trial D and Acc (%)	Trial E and Acc (%)
	0 0 0 0	0 2 1 0	1 2 0 0	0 2 0 1	0 2 1 0
2	2 7 2 2 32.43	0 6 1 2 37.84	1 6 2 4 45.00	0 2 0 2 32.26	0 8 0 1 31.43
	1 6 0 5	0 4 4 3	0 3 5 2	0 7 1 5	0 6 1 1
	0 7 0 5	0 4 6 4	0 2 6 6	0 4 0 7	0 7 6 2
	0 0 0 0	1 1 1 0	1 1 0 1	0 1 1 1	0 2 1 0
3	2 9 0 2 43.24	0 5 0 4 35.14	1 6 2 4 42.50	0 3 0 1 32.26	0 7 1 1 37.14
	1 4 3 4	2 3 1 5	0 4 3 3	0 5 1 7	0 3 4 1
	0 4 4 4	1 3 4 6	0 3 4 7	0 5 0 6	0 8 5 2
	0 0 0 0	0 1 0 2	1 2 0 0	1 1 0 1	0 1 2 0
4	2 10 0 1 40.54	0 3 1 5 27.03	1 6 3 3 35.00	0 2 1 1 32.26	1 5 3 0 28.57
	0 7 0 5	0 5 4 2	0 4 3 3	0 5 2 6	0 4 2 2
	0 6 1 5	0 5 6 3	2 3 5 4	0 5 1 5	0 6 6 3
	0 0 0 0	1 1 1 0	1 2 0 0	1 1 0 1	0 1 2 0
5	0 9 1 3 35.14	0 4 3 2 37.84	1 8 2 2 40.00	0 3 1 0 32.26	0 7 2 0 37.14
	2 5 1 4	1 5 2 3	0 7 2 1	0 7 2 4	0 4 3 1
	1 4 4 3	2 5 0 7	0 4 5 5	1 4 2 4	0 3 9 3
	0 0 0 0	0 2 0 1	1 1 1 0	1 0 0 2	0 0 3 0
6	2 7 3 1 37.84	1 4 3 1 32.43	1 8 2 2 32.50	0 3 1 0 45.16	0 6 2 1 28.57
	1 3 2 6	0 6 3 2	0 5 2 3	0 5 3 5	0 4 4 0
	1 3 3 5	2 3 4 5	0 4 8 2	1 2 1 7	0 4 11 0
	0 0 0 0	1 1 0 1	1 1 1 0	0 1 1 1	0 0 3 0
7	1 8 2 2 35.14	0 4 0 5 35.14	1 8 2 2 35.00	0 3 1 0 35.48	1 6 2 0 31.43
	2 6 1 3	1 4 1 5	0 5 2 3	0 4 3 6	0 4 4 0
	2 3 3 4	0 5 2 7	1 3 7 3	0 4 2 5	0 4 10 1
	0 0 0 0	1 1 1 0	1 1 0 1	0 1 0 2	0 0 2 1
10	2 8 2 1 35.14	0 5 3 1 29.73	2 8 1 2 40.00	0 2 1 1 38.71	1 5 2 1 37.14
	1 5 1 5	2 5 1 3	0 3 3 4	0 4 3 6	0 4 4 0
	1 3 4 4	0 5 5 4	1 4 5 4	0 2 2 7	0 5 6 4
	0 0 0 0	0 2 0 1	1 1 1 0	0 1 0 2	0 2 1 0
15	2 9 1 1 43.24	2 2 4 1 29.73	0 9 2 2 37.50	0 3 1 0 41.94	0 8 0 1 34.29
	2 2 4 4	0 4 2 5	0 4 3 3	0 3 5 5	0 5 2 1
	1 5 3 3	1 3 3 7	0 3 9 2	1 2 3 5	0 2 11 2

Table 4.7: Outcome for **Den** cohort using the texture inertia on the wavelet transform of the image. For each trial the overall classification performance and confusion matrix is given.

Trial A and Acc(%)	Trial B and Acc (%)	Trial C and Acc (%)	Trial D and Acc (%)	Trial E and Acc (%)
0 0 0 0	0 3 0 0	0 2 1 0	0 2 0 1	0 2 1 0
0 12 1 0 59.46	0 5 4 0 27.03	0 10 3 0 37.5	0 4 0 0 19.35	0 8 1 0 34.29
0 3 3 6	0 6 5 0	0 5 5 0	0 8 0 5	0 4 4 0
0 5 0 7	0 7 7 0	0 6 8 0	0 9 0 2	0 10 5 0

Table 4.8: Outcome for **Den** cohort using regional skewness of the image. For each trial the overall classification performance and confusion matrix is given.

regional moments were used in the discriminant function with a correlation coefficient $r = 0.57 \pm 0.08$. Since the use of 7 moments is somewhat excessive we also gave the results for 5 regional moments $r = 0.5 \pm 0.2$ and the single best regional moment $r = 0.18 \pm 0.09$. The correlation coefficient for the best classifier using the percentage of correctly classified cases, 4 regional moments, gave a correlation coefficient[4] $r = 0.3 \pm 0.2$. However, prior to presenting the detailed results from other studies it should be reiterated that the only feature common to this thesis and previous studies is the regional skewness. The method used to create many of the features in this thesis was similar to the approach used in the other studies but a small change in the method can result in considerably different characteristics of the new feature. This is particularly true for the textures which we have used.

Karssemeijer [Karssemeijer, 1998] was able to obtain 65% agreement with the density classification provided by radiologists on 615 mammograms and 80% agreement if the study was constrained to use the more recent mammograms with a more consistent quality (125 cases, 1991–1994). Additionally, the majority of misclassifications were incorrect by a single density grade. (For the entire set of images and the feature set used for the 65% result, the minor error rate was 0.33 and the major error rate[5] 0.023.) For the data set in this study it was found that the minor and major error rates were higher than that reported by Karssemeijer. All the properties considered had roughly similar error rates. The minor error rate at ~ 0.45 was on the order of the value found by Karssemeijer while the major error rate (~ 0.2) was substantially

[4] The difference in the selected set of features which gave the best classification performance may be a small sample size effect or due to the properties of the correlation coefficient. The correlation coefficient considers cases which are "nearly" classified correctly whereas the fraction of correctly classified cases does not.

[5] Karssemeijer defined the error rate as the number misclassified by one grade (minor) or more than one grade (major) normalised by the total number sample size, including the correctly classified cases.

Classifiers	% Exact Agreement	% Minor Disagreement	% Major Disagreement
R_0 vs. computer	66.0	22.0	12.0
R_1 vs. computer	64.8	20.9	14.3
R_2 vs. computer	68.8	16.8	14.4
R_3 vs. computer	65.8	16.5	12.4
R_4 vs. computer	68.5	16.5	15.0

Table 4.9: Classification performance for the approach used by Tahoces. (Results taken from [Tahoces et al., 1995]).

greater. For these considerations only (error rates) the multifractal dimensions performed slightly better than the regional moments followed by the remaining properties with very similar but slightly greater error rates.

Tahoces [Tahoces et al., 1995] on the other hand, used Wolfe grades to classify the mammograms into risk groups and in comparing the computer classification against 5 radiologists (R_0–R_4) Tahoces found the results summarized in Table 4.9.

In another study, Byng [Byng et al., 1996b] used mammograms classified by radiologists into a six class density grade system (SCC) on 100 cases. The particular results which are relevant for comparison to this work is the correlation between

- SCC and their semiautomatic system of calculating the percent density (Pearson correlation coefficient, $r = 0.811$)
- SCC vs. regional skewness: $r = -0.761$
- SCC vs. their fractal dimension: $r = -0.649$.

(See [Byng et al., 1996b]. The negative correlation coefficients indicates that the higher skewness values and fractal coefficients are associated with the lower density grades and vice versa.) Finally, Boone *et al.* [Boone et al., 1998] used a continuous scale (BDI) which was specifically constructed to correlate with the radiologists' ranking of images. On 160 patients, they found a correlation of $r = 0.907$ between the BDI based on the results calculated using the radiologists' ranking and the computerised ranking.

A cursory examination of the results shown in the tables, above, indicate a poorer correlation between the properties selected for this study and the density grade classification than previous studies by other groups. While it is possible that the properties themselves were less suitable for this classification task, there were significant differences between this work and the studies described above. First, the number of density grades differed for most of the studies and although Karssemeijer also used four density grades, Byng used six density grades, Boone a continuous scale up to 100 and Tahoces used Wolfe grades. Second, the dataset itself was different. This is distinct from the difference in sample sizes, discussed below. For example, Karssemeijer also reported that he was unable to reproduce the results of Byng using the fractal dimension on a set of images from Nijmegen, although the failure in that case may be due to the variation in film quality. The majority of the studies, discussed above, used locally obtained mammograms and to the best of my knowledge a comparison of the performance of property sets on the datasets from other groups has not been performed.

The most significant difference between this study and the work cited above was that the sample size which was used for classification of the density grades was considerably smaller ($\sim$ 80 in total and $\sim$ 30 for the test set) than in any of the other studies and much fewer than the complete 240 cases. Most likely, the small sample size and uneven distribution of density grades severely inhibited the ability of the sys-

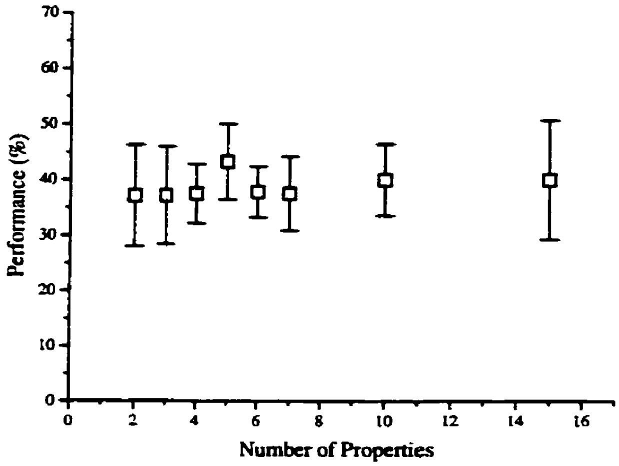


Figure 4.3: Overall classification performance for **Den** cohort allowing for features to be selected from all calculated properties.

tems from identifying the most useful, general features. For completeness, the results were given when all properties were combined and the genetic algorithm allowed to select combinations of features from all calculated properties (Figure 4.3 and Table 4.10). The small sample size resulted in many unrelated feature combinations with comparable performance so that a detailed analysis of the actual features selected yielded little that was generalisable.

4.1.2 Diagnosis Classification

For this part of the analysis, the mammograms that contained a diagnosed malignancy and the mammograms from the patients who had both breasts diagnosed as normal (i.e. **Diag** cohort), were used to select the features which were subsequently used in the analysis for the remaining studies. The results for this cohort shared a few of the same general characteristics which had been observed for the analysis of the density grade classification (**Den** cohort). For example the classification performance was, for

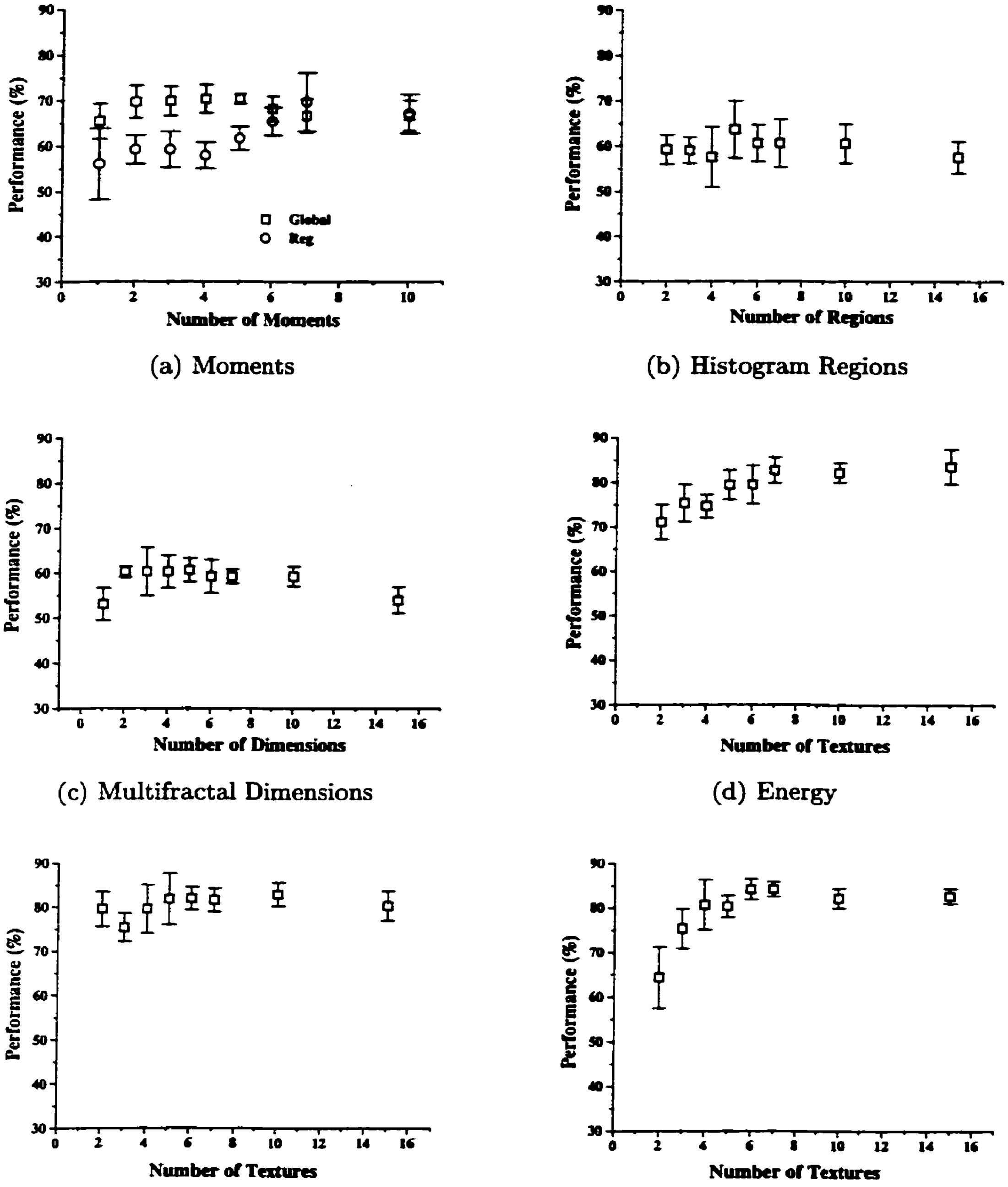
No. of Prop.	Trial A and Acc (%)					Trial B and Acc (%)					Trial C and Acc (%)					Trial D and Acc (%)					Trial E and Acc (%)				
	0	0	0	0		0	1	0	2		1	2	0	0		1	2	0	0		0	1	2	0	
2	3	7	1	2	40.54	0	6	1	2	35.14	1	8	2	2	47.50	0	2	1	1	22.58	0	7	1	1	37.14
	0	4	1	7		0	6	1	4		0	3	6	1		0	9	1	3		0	4	3	1	
	0	3	2	7		0	7	1	6		1	3	6	4		0	7	1	3		0	4	8	3	
	0	0	0	0		0	0	3	0		1	1	0	1		1	0	1	1		0	1	2	0	
3	2	9	0	2	45.95	1	3	3	2	32.43	1	6	5	1	40.00	0	2	1	1	22.58	0	5	3	1	37.14
	1	3	4	4		0	3	6	2		0	2	4	4		0	9	1	3		0	3	4	1	
	0	4	4	4		1	2	8	3		1	2	6	5		0	5	3	3		0	3	8	4	
	0	0	0	0		0	1	1	1		1	2	0	0		0	1	0	2		0	1	2	0	
4	2	8	1	2	43.24	0	4	3	2	32.43	1	7	4	1	37.50	0	4	0	0	32.26	0	6	3	0	42.86
	0	3	3	6		1	4	2	4		0	2	2	6		0	5	5	3		0	2	4	2	
	1	1	5	5		1	5	2	6		1	1	7	5		0	8	2	1		0	5	5	5	
	0	0	0	0		0	1	2	0		1	0	1	1		0	0	1	2		0	0	3	0	
5	1	8	2	2	43.24	1	5	2	1	32.43	1	6	4	2	45.00	0	3	1	0	32.26	0	7	2	0	45.71
	0	4	2	6		1	5	2	3		0	2	4	4		0	5	5	3		0	2	4	2	
	0	3	3	6		1	5	3	5		0	1	6	7		0	6	3	2		0	6	4	5	
	0	0	0	0		0	1	2	0		1	1	1	0		1	0	1	1		0	2	0	1	
6	1	11	1	0	37.84	1	4	2	2	32.43	1	7	5	0	42.50	1	2	0	1	32.26	1	6	2	0	40.00
	0	5	2	5		0	5	3	3		0	2	4	4		0	5	5	3		0	4	3	1	
	0	1	10	1		1	5	3	5		0	2	7	5		1	7	1	2		0	6	4	5	
	0	0	0	0		0	1	2	0		1	1	0	1		0	1	1	1		0	1	2	0	
7	1	9	1	2	43.24	0	7	1	1	40.54	1	8	3	1	37.50	0	2	0	2	25.81	1	7	1	0	37.14
	1	4	1	6		0	5	3	3		0	3	3	4		0	5	5	3		0	4	2	2	
	0	2	4	6		0	6	3	5		1	2	8	3		1	5	4	1		0	6	5	4	
	0	0	0	0		2	0	1	0		1	1	1	0		1	1	1	0		0	1	1	1	
10	2	5	3	3	43.24	2	4	0	3	32.43	2	6	3	2	40.00	0	4	0	0	29.03	1	7	0	1	42.86
	0	3	5	4		2	7	1	1		0	2	3	5		0	7	3	3		0	4	2	2	
	0	2	4	6		2	3	4	5		0	1	7	6		1	5	4	1		0	5	4	6	
	0	0	0	0		1	1	1	0		1	1	0	1		0	1	2	0		0	1	2	0	
15	2	8	2	1	48.65	4	4	1	0	24.32	1	10	1	1	42.50	0	3	1	0	25.81	0	7	1	1	40.00
	1	2	3	6		0	4	4	3		0	3	4	3		1	6	3	3		0	3	3	2	
	0	1	4	7		1	6	7	0		0	3	9	2		0	2	7	2		0	6	5	4	

Table 4.10: Outcome for **Den** cohort allowing for features to be selected from all calculated properties. For each trial the overall classification performance and confusion matrix is given.

a large part, independent of the number of properties used to form the discriminant function. (See Figure 4.4 and the detailed performance results in Tables 4.11–4.18.)

However, the differences compared to the results for the **Den** cohort were equally prominent. Most obvious was that the overall performance was somewhat higher for all properties. The poorest performers were the regional moments, histogram sub-regions and the multifractal dimensions. The behaviour of the histogram sub-regions seemed to be the simplest to characterise in that the accuracy was essentially unchanged at $\sim$ 60% when the number of sub-regions used to form the classifier varied from 2–15. While the classification accuracy when using the multifractal dimensions was $\sim$ 53% for a single multifractal dimension, it rose to $\sim$ 60% as more fractal dimensions were employed. The performance also reached a plateau at $\sim$ 60% for 2–10 dimensions and dropped back to $\sim$ 50% for 15, Figure 4.4(c). The global moments appear to have the best performance of the features described so far at $\sim$ 70% for two global moments, Figure 4.4(a). The behaviour of the regional moments was similar to the multifractal dimensions except that for more then 5 regional moments the accuracy increases to $\sim$ 70%, Figure 4.4(a). The most significant features for classification appear to be the textures. All three had an overall peak accuracy at 80–85% and the texture energy and inertia had similar overall characteristics. Both had relatively low accuracy for a single texture and attain peak performance at 6 or 7 textures which was maintained at nearly that level for up to 15 textures. On the other hand, the performance of the texture entropy was essentially uniform at $\sim$ 80% when the number of textures used in the classifier is varied from 2–15.

From examination of the data used for the summary figures, Tables 4.11–4.18, it is clear that the summary figures do not make apparent some important information



(a) Moments

(b) Histogram Regions

(c) Multifractal Dimensions

(d) Energy

(e) Entropy

(f) Inertia

Figure 4.4: Overall classification performance for **Diag** class

No. of	Trial A			Trial B			Trial C			Trial D			Trial E		
Regions	and Acc (%)			and Acc (%)			and Acc (%)			and Acc (%)			and Acc (%)		
2	25	13	60.49	26	14	55.42	21	10	59.09	28	17	59.21	23	12	64.38
	19	24		23	20		17	18		14	17		14	24	
3	22	16	56.79	28	12	59.04	23	8	59.09	30	15	64.47	19	16	58.90
	19	24		22	21		19	16		12	19		14	24	
4	23	15	50.62	26	14	54.22	19	12	57.58	32	13	63.16	25	10	67.12
	25	18		24	19		16	19		15	16		14	24	
5	25	13	55.56	26	14	54.22	20	11	63.64	36	9	68.42	24	11	65.75
	23	20		24	19		13	22		15	16		14	24	
6	23	15	55.56	30	10	57.83	20	11	60.61	32	13	65.79	22	13	63.01
	21	22		25	18		15	20		13	18		14	24	
7	25	13	53.09	26	14	56.63	18	13	60.61	35	10	65.79	23	12	64.38
	25	18		22	21		13	22		16	15		14	24	
10	24	14	54.32	26	14	55.42	18	13	60.61	29	16	64.47	22	13	61.64
	23	20		23	20		13	22		11	20		15	23	
15	23	15	56.79	24	16	55.42	18	13	57.58	35	10	64.47	22	13	58.90
	20	23		21	22		15	20		17	14		17	21	

Table 4.11: Outcome for **Diag** cohort using sub-regions of the histogram. For each trial the number of features used in the discriminant function, confusion matrix and overall classification performance are shown.

No. of	Trial A			Trial B			Trial C			Trial D			Trial E		
Mom.	and Acc (%)			and Acc (%)			and Acc (%)			and Acc (%)			and Acc (%)		
1	29	9	65.43	27	13	66.27	22	9	60.61	28	17	71.05	24	11	63.01
	19	24		15	28		17	18		5	26		16	22	
2	23	15	74.07	24	16	69.88	23	8	69.70	23	22	64.47	21	14	67.12
	6	37		9	34		12	23		5	26		10	28	
3	22	16	67.90	25	15	69.88	23	8	66.67	30	15	75.00	22	13	71.23
	10	33		10	33		14	21		4	27		8	30	
4	26	12	70.37	25	15	69.88	24	7	65.15	30	15	73.68	23	12	71.23
	12	31		10	33		16	19		5	26		9	29	
5	26	12	70.37	26	14	68.67	23	8	69.70	28	17	71.05	24	11	71.23
	12	31		12	31		12	23		5	26		10	28	
6	25	13	66.67	25	15	66.27	23	8	68.18	28	17	71.05	26	9	72.60
	14	29		13	30		13	22		5	26		11	27	
7	25	13	64.20	25	15	66.27	23	8	66.67	30	15	73.68	26	9	69.86
	16	27		13	30		14	21		5	26		13	25	
10	22	16	65.43	25	15	66.27	23	8	66.67	31	14	73.68	26	9	67.12
	12	31		13	30		14	21		6	25		15	23	

Table 4.12: Outcome for **Diag** cohort using global moments of the histogram. For each trial the number of features used in the discriminant function, confusion matrix and overall classification performance are shown.

in the data. For example, there were several cases where the overall accuracy was relatively high ($\gtrsim$ 60%) but the true positive or true negative fraction was very poor ($<$ 50%). Fortunately, these were not present for any of the cases using texture properties. However, there were a few situations where the true positive or true negative fraction was $\sim$ 60%. These typically only appear for a small number of textures where the median performance was also relatively poor. It should also be noted that the performance of the regional skewness was essentially random at 53±4%.

Additional useful observations can be drawn from the selection of features for each of the extracted features (Tables[6]4.19–4.23.) but two points that should be

[6] The tables contain the number of features used in the discriminant function for each property, leftmost column, and the selected features for each of the 5 redistributions of the sample cases, Trials A–E.

No. of Mom.	Trial A and Acc (%)			Trial B and Acc (%)			Trial C and Acc (%)			Trial D and Acc (%)			Trial E and Acc (%)		
1	29	9	56.79	33	7	54.22	25	6	56.06	28	17	72.37	29	6	53.42
	26	17		31	12		23	12		4	27		28	10	
2	26	12	64.20	27	13	55.42	21	10	59.09	24	21	59.21	21	14	60.27
	17	26		24	19		17	18		10	21		15	23	
3	27	11	65.43	31	9	56.63	23	8	57.58	25	20	59.21	25	10	63.01
	17	26		27	16		20	15		11	20		17	21	
4	28	10	56.79	32	8	63.86	22	9	57.58	24	21	57.89	27	8	60.27
	25	18		22	21		19	16		11	20		21	17	
5	30	8	66.67	29	11	60.24	22	9	60.61	25	20	61.84	28	7	61.64
	19	24		22	21		17	18		9	22		21	17	
6	27	11	65.43	26	14	60.24	23	8	68.18	24	21	64.47	28	7	67.12
	17	26		19	24		13	22		6	25		17	21	
7	29	9	74.07	24	16	57.83	23	8	69.70	26	19	64.47	29	6	71.23
	12	31		19	24		12	23		8	23		15	23	
10	28	10	72.84	28	12	61.45	23	8	69.70	26	19	65.79	24	11	67.12
	12	31		20	23		12	23		7	24		13	25	

Table 4.13: Outcome for **Diag** cohort using regional moments of the image. For each trial the number of features used in the discriminant function, confusion matrix and overall classification performance are shown.

noted prior to any further discussion was:

1. The exact number of selected features or regions may vary, particularly when the genetic algorithm was used. Due to the specific implementation of the genetic algorithm in **ga_ors**, if a set of features was identified which classifies the images well, the set was retained even though the number of features may be less than what was desired. Naturally, the program cannot search through all possible combinations so that **ga_ors** does respect an upper limit to the number of properties to use in the discriminant function, for the most part.

2. The specific features that were selected varied somewhat for the different training/test groups and became more apparent as the number of selected features increased. This is partly due to overfitting — a greater number of features

No. of	Trial A			Trial B			Trial C			Trial D			Trial E		
Dim.	and Acc (%)			and Acc (%)			and Acc (%)			and Acc (%)			and Acc (%)		
1	29	9	56.79	23	17	57.83	17	14	53.03	13	32	51.32	20	15	49.32
	26	17		18	25		17	18		5	26		22	16	
2	24	14	60.49	20	20	59.04	17	14	60.61	17	28	57.89	21	14	60.27
	18	25		14	29		12	23		4	27		15	23	
3	24	14	54.32	18	22	53.01	22	9	62.12	23	22	65.79	20	15	60.27
	23	20		17	26		16	19		4	27		14	24	
4	23	15	55.56	19	21	55.42	22	9	62.12	21	24	63.16	21	14	60.27
	21	22		16	27		16	19		4	27		15	23	
5	28	10	60.49	19	21	59.04	20	11	60.61	20	25	63.16	22	13	65.75
	22	21		13	30		15	20		3	28		12	26	
6	25	13	59.26	19	21	56.63	19	12	53.03	19	26	59.21	21	14	63.01
	20	23		15	28		19	16		5	26		13	25	
7	28	10	60.49	19	21	56.63	20	11	59.09	21	24	59.21	18	17	60.27
	22	21		15	28		16	19		7	24		12	26	
10	26	12	60.49	21	19	60.24	20	11	56.06	20	25	59.21	17	18	56.16
	20	23		14	29		18	17		6	25		14	24	
15	26	12	59.26	20	20	55.42	17	14	51.52	18	27	53.95	15	20	53.42
	21	22		17	26		18	17		8	23		14	24	

Table 4.14: Outcome for **Diag** cohort using multifractal dimensions. For each trial the number of features used in the discriminant function, confusion matrix and overall classification performance are shown.

No. of	Trial A			Trial B			Trial C			Trial D			Trial E		
Textures	amd Acc (%)			amd Acc (%)			amd Acc (%)			amd Acc (%)			amd Acc (%)		
2	30	8	74.07	26	14	71.08	20	11	75.76	26	19	65.79	24	11	69.86
	13	30		10	33		5	30		7	24		11	27	
3	28	10	77.78	28	12	72.29	24	7	80.30	29	16	69.74	26	9	75.34
	8	35		11	32		6	29		7	24		9	29	
4	31	7	77.78	28	12	74.70	23	8	71.21	30	15	72.37	26	9	75.34
	11	32		9	34		11	24		6	25		9	29	
5	32	6	82.72	33	7	79.52	25	6	77.27	31	14	73.68	29	6	79.45
	8	35		10	33		9	26		6	25		9	29	
6	33	5	85.19	32	8	79.52	26	5	75.76	33	12	75.00	29	6	82.19
	7	36		9	34		11	24		7	24		7	31	
7	31	7	82.72	30	10	78.31	27	4	83.33	36	9	77.63	32	3	83.56
	7	36		8	35		7	28		8	23		9	29	
10	33	5	83.95	34	6	84.34	26	5	78.79	38	7	81.58	31	4	82.19
	8	35		7	36		9	26		7	24		9	29	
15	32	6	85.19	31	9	78.31	25	6	77.27	39	6	85.53	32	3	83.56
	6	37		9	34		9	26		5	26		9	29	

Table 4.15: Outcome for **Diag** cohort using texture energy on the wavelet transform of the image. For each trial the number of features used in the discriminant function, confusion matrix and overall classification performance are shown.

No. of	Trial A			Trial B			Trial C			Trial D			Trial E		
Textures	amd Acc (%)			amd Acc (%)			amd Acc (%)			amd Acc (%)			amd Acc (%)		
2	30	8	80.25	32	8	79.52	19	12	80.30	27	18	71.05	24	11	76.71
	8	35		9	34		1	34		4	27		6	32	
3	28	10	77.78	31	9	77.11	18	13	71.21	32	13	71.05	24	11	75.34
	8	35		10	33		6	29		9	22		7	31	
4	29	9	81.48	31	9	79.52	24	7	80.30	29	16	69.74	21	14	71.23
	6	37		8	35		6	29		7	24		7	31	
5	31	7	83.95	29	11	79.52	25	6	81.82	28	17	69.74	26	9	83.56
	6	37		6	37		6	29		6	25		3	35	
6	32	6	85.19	31	9	81.93	22	9	81.82	30	15	78.95	28	7	84.93
	6	37		6	37		3	32		1	30		4	34	
7	32	6	82.72	34	6	79.52	23	8	80.30	33	12	81.58	29	6	86.30
	8	35		11	32		5	30		2	29		4	34	
10	32	6	82.72	34	6	81.93	24	7	78.79	35	10	82.89	30	5	86.30
	8	35		9	34		7	28		3	28		5	33	
15	33	5	80.25	30	10	77.11	23	8	80.30	35	10	81.58	29	6	86.30
	11	32		9	34		5	30		4	27		4	34	

Table 4.16: Outcome for **Diag** cohort using the texture entropy on the wavelet transform of the image. For each trial the number of features used in the discriminant function, confusion matrix and overall classification performance are shown.

No. of	Trial A			Trial B			Trial C			Trial D			Trial E		
Textures	amd Acc (%)			amd Acc (%)			amd Acc (%)			amd Acc (%)			amd Acc (%)		
2	28	10	64.20	32	8	79.52	25	6	63.64	27	18	64.47	23	12	64.38
	19	24		9	34		18	17		9	22		14	24	
3	27	11	76.54	35	5	80.72	21	10	69.70	32	13	71.05	23	12	75.34
	8	35		11	32		10	25		9	22		6	32	
4	29	9	79.01	33	7	80.72	27	4	86.36	33	12	71.05	28	7	82.19
	8	35		9	34		5	30		10	21		6	32	
5	31	7	77.78	32	8	81.93	25	6	80.30	36	9	84.21	27	8	79.45
	11	32		7	36		7	28		3	28		7	31	
6	32	6	83.95	37	3	85.54	25	6	80.30	36	9	84.21	29	6	86.30
	7	36		9	34		7	28		3	28		4	34	
7	34	4	85.19	32	8	81.93	24	7	81.82	35	10	84.21	28	7	84.93
	8	35		7	36		5	30		2	29		4	34	
10	33	5	83.95	34	6	84.34	25	6	78.79	35	10	81.58	28	7	82.19
	8	35		7	36		8	27		4	27		6	32	
15	32	6	82.72	33	7	83.13	23	8	80.30	35	10	82.89	27	8	79.45
	8	35		7	36		5	30		3	28		7	31	

Table 4.17: Outcome for **Diag** cohort using the texture inertia on the wavelet transform of the image. For each trial the number of features used in the discriminant function, confusion matrix and overall classification performance are shown.

Trial A			Trial B			Trial C			Trial D			Trial E		
and Acc (%)			and Acc (%)			and Acc (%)			and Acc (%)			and Acc (%)		
26	12	55.56	24	16	46.99	23	8	53.03	18	27	52.63	23	12	45.21
24	19		28	15		23	12		9	22		28	10	

Table 4.18: Outcome for **Diag** cohort using regional skewness of the image. For each trial the confusion matrix and overall classification performance are shown.

requires a greater sample for a general discriminant function[7]. This is another reason to favour a small number of features but it would be premature to discount the importance of the properties under consideration without more significant evidence.

If we begin with the sub-regions selected from the histogram, Table 4.19, we see that when a small number of regions was selected, the genetic algorithm chose regions from the lower third of the histogram as well as one region from the middle or upper third. For two regions, predominantly the lower and middle third was used while for three, one region from each tended to be selected. As the number of regions increases, regions from the lower third are chosen more frequently than either the middle or upper third but for 10 regions or more, regions from the middle third were also highly represented. It is also important to note that the number of regions for the "best" classification began to be fewer than the requested number of regions as the number exceeds 7.

A cursory examination of the global and regional moments (Table 4.20) which had the greatest discriminatory power, seemed to be somewhat disappointing, in that frequently rather high moments were selected. The difficulty with the high moments was that they are extremely sensitive to small variations in the distribution and attempting to use them for a stable classification method would be ill-advised. Recall from Figure 4.4(a) that the classification accuracy of the moments was essentially unchanged for 2–6 global moments so that a different number of moments with more favourable characteristics can be used with little impact on the classification performance. From 5–7 regional moments the accuracy improved somewhat but the

[7] In optimisation terminology, multiple local minima have been found but it is not possible to identify a global minimum without a sample more representative of the full population.

No. of Regions	Trial A	Trial B	Trial C	Trial D	Trial E
2	35-38 112-113	44-52 112-114	41-42 112-113	38-39 214-216	45-49 112-113
3	38-39 97-100 101-105	48-50 109-110 214-215	41-42 115-117 220-221	1-2 38-39 213-216	112-113 225-227 238-239
4	22-24 55-59 61-62 112-119	1-2 44-50 109-110 213-215	41-42 112-113 164-166 190-191	1-2 4-7 38-39 220-221	45-47 112-113 225-227 238-239
5	22-24 57-58 61-62 112-119 253-254	1-2 45-50 97-102 103-106 213-215	27-28 112-117 140-149 150-153 220-221	1-2 4-14 38-39 183-186 212-216	18-20 47-49 127-131 225-227 238-239
6	22-24 57-58 61-62 97-100 101-108 213-214	2-3 15-17 45-49 99-101 108-110 213-215	26-28 78-79 113-117 122-125 150-153 220-221	1-2 4-14 15-17 18-20 38-39 220-224	1-3 5-6 18-20 127-131 223-227 238-239
7	22-24 25-31 57-58 61-62 112-119 220-222 253-254	1-2 13-19 33-34 45-50 97-102 103-110 213-215	26-29 78-79 82-84 113-117 122-125 151-153 220-221	1-2 4-6 18-20 25-26 35-39 185-186 220-221	1-4 5-8 47-49 112-113 225-228 233-238
10	22-24 55-58 61-62 84-86 97-100 101-104 158-161 167-172 253-256	1-2 15-19 33-34 45-50 97-102 103-110 213-215	26-29 32-34 78-81 82-84 113-117 122-125 160-166 170-177 179-184 189-194	1-2 3-8 37-38 102-104 111-113 126-127 129-130 145-149 201-203 213-214	24-29 32-34 38-39 131-139 140-142 149-156 220-230 238-239
15	1-2 19-20 42-43 53-59 60-62 97-100 103-104 117-119 198-203 207-208 213-216 254-256	45-52 55-58 97-102 103-104 109-110 132-134 145-149 201-203 204-206	26-29 32-34 41-42 79-81 82-84 102-104 113-117 122-125 157-169 170-177 182-186 194-196	1-2 4-7 36-38 102-108 111-113 128-130 164-166 213-217	1-3 5-7 12-16 38-42 127-131 135-139 140-142 188-190 225-228 236-239

Table 4.19: Actual subregions of the histogram selected as the "best" properties for each trial in the **Diag** cohort corresponding to Table 4.11. The selected regions represent the rebinned grey levels (Section 3.3.2) from 1–256. Bins 257–266 were assigned the patients' age.

No. of Moments	Trial A	Trial B	Trial C	Trial D	Trial E
1	2	2	2	2	2
2	3	4	4	3	6
	4	5	5	4	7
3	1	3	3	3	3
	3	6	4	6	6
	4	10	8	10	9
4	1	2	3	2	3
	3	3	4	3	4
	4	6	5	6	8
	7	8	6	8	9
5	1	2	2	2	1
	3	3	3	3	3
	4	4	4	6	4
	9	5	5	8	8
	10	9	6	9	9
6	3	1	1	1	1
	5	2	2	2	3
	7	3	3	3	4
	8	4	4	6	5
	9	5	5	8	7
	10	7	6	9	8
7	1	1	2	2	1
	3	2	3	3	3
	5	3	5	4	4
	7	5	6	6	5
	8	6	7	8	7
	9	7	8	9	8
	10	10	9	10	11
10	1	1	1	1	1
	2	2	2	2	2
	3	3	3	4	3
	4	4	5	5	4
	5	5	6	6	6
	6	7	7	7	7
	7	8	8	8	8
	9	9	9	9	9
	10	10	10	10	10
	11	11	11	11	11

(a) Global Moments

No. of Moments	Trial A	Trial B	Trial C	Trial D	Trial E
1	4	4	4	4	4
2	1	1	1	1	1
	4	4	4	4	4
3	1	1	1	1	1
	2	4	4	6	4
	4	9	7	9	9
4	4	4	1	4	1
	5	6	2	6	3
	6	8	4	8	4
	8	10	7	10	5
5	3	4	1	3	1
	4	6	2	4	2
	5	7	3	6	3
	6	8	4	8	4
	8	9	5	10	5
6	3	3	3	3	3
	4	4	4	4	4
	5	6	6	6	6
	6	7	7	7	7
	7	8	8	8	8
	8	9	9	9	9
7	2	3	2	2	2
	3	4	3	3	4
	4	5	4	4	5
	5	6	5	6	6
	6	7	6	7	7
	7	8	8	8	9
	8	9	9	9	10
10	1	1	1	1	1
	2	2	2	2	2
	3	3	3	3	3
	4	4	4	4	4
	5	5	5	5	5
	6	6	6	6	6
	7	7	7	7	7
	8	8	8	8	9
	9	9	9	9	10
	11	11	11	11	11

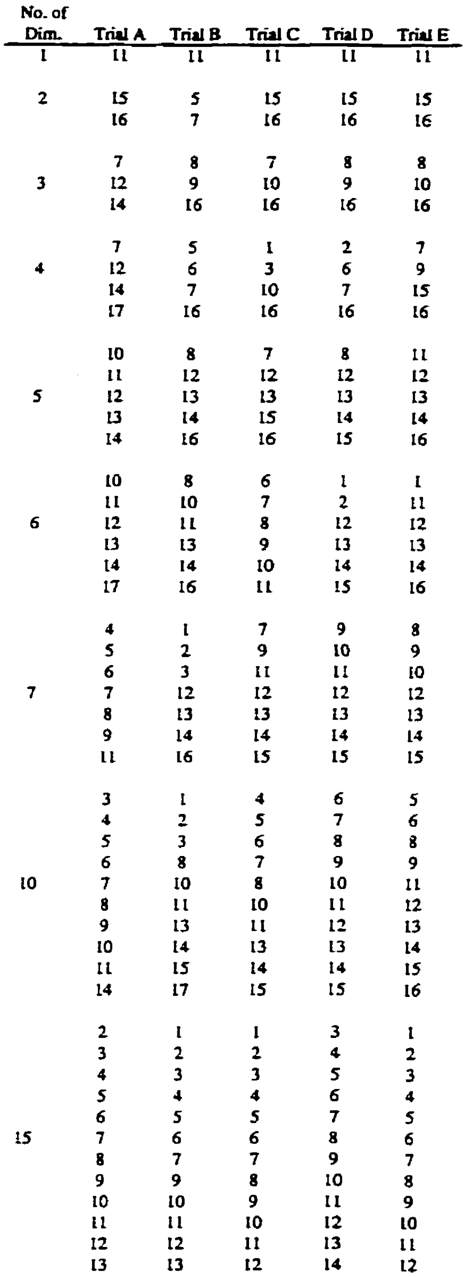
(b) Regional Moments

Table 4.20: Best moments selected for the classification of the **Diag** cohort for a varying number of moments in the discriminant function. The entries are associated with the corresponding entries in Tables 4.12 and 4.13. The moment numbered 11 is the patients' age and not the 11th moment.

improvement does not justify the cost of using substantially more moments. Further, if only one moment was selected, the preferred moment was either the global variance or regional kurtosis. If two moments were used, the preferred pair was the regional mean and kurtosis or a pair of global moments from $\{3, 4, 5\}$. Any of these is low enough that a stable classifier was much more likely than if the sixth or greater moments was used.

For the multifractal dimensions, D_q was found for 20 values of q from -5 up to 0.7. However, the classification performance varied considerably ($\sim$ 25%) for the five redistribution trials of the cases between the training and test sets. Since the selected features also favoured the higher values of q, this may indicate that that the range selected for q still contains a region where the multifractal model was not valid (See Chapter 3.3.3.) or may simply be due to overfitting. When the range of q was constrained to the first 16 values ($q = -5$ to $q = -0.5$) the stability of the results improved significantly, Figure 4.4(c) and Table 4.21. The preferred dimensions, selected by the search algorithm lie close to the upper part of the range for q. However, the single best multifractal dimension corresponded to D_{-2} (bin 11). Additionally, it appeared that dimensions far from the upper range for q were selected only when "forced". For example if 15 dimensions were requested then, with no repetitions, dimensions far from the upper q range *must* be selected.

The parameters, d and θ, in the calculation of the textures resulted in a particularly large number of features. For these properties, the sheer number of variables required that a random search method be used, i.e. **ga_ors**. The features selected by ga_ors are given in Tables 4.22 and 4.23. If the genetic algorithm is to be applied for these properties the features must be arranged to form an artificial spectrum as



No. of Dim.	Trial A	Trial B	Trial C	Trial D	Trial E
1	11	11	11	11	11
2	15	5	15	15	15
	16	7	16	16	16
3	7	8	7	8	8
	12	9	10	9	10
	14	16	16	16	16
4	7	5	1	2	7
	12	6	3	6	9
	14	7	10	7	15
	17	16	16	16	16
5	10	8	7	8	11
	11	12	12	12	12
	12	13	13	13	13
	13	14	15	14	14
	14	16	16	15	16
6	10	8	6	1	1
	11	10	7	2	11
	12	11	8	12	12
	13	13	9	13	13
	14	14	10	14	14
	17	16	11	15	16
7	4	1	7	9	8
	5	2	9	10	9
	6	3	11	11	10
	7	12	12	12	12
	8	13	13	13	13
	9	14	14	14	14
	11	16	15	15	15
10	3	1	4	6	5
	4	2	5	7	6
	5	3	6	8	8
	6	8	7	9	9
	7	10	8	10	11
	8	11	10	11	12
	9	13	11	12	13
	10	14	13	13	14
	11	15	14	14	15
	14	17	15	15	16
15	2	1	1	3	1
	3	2	2	4	2
	4	3	3	5	3
	5	4	4	6	4
	6	5	5	7	5
	7	6	6	8	6
	8	7	7	9	7
	9	9	8	10	8
	10	10	9	11	9
	11	11	10	12	10
	12	12	11	13	11
	13	13	12	14	12

Table 4.21: Best multifractal dimensions selected for the classification of **Diag** cohort for a varying number of dimensions in the discriminant function. The entries were associated with the corresponding entries in Table 4.14 and the dimensions were numbered sequentially from 1 for the generalised dimension, $D_q|_{q=-5}$ (from Equation 2.6) to 16 for $D_q|_{q=-0.5}$ in equal increments of $q = 0.3$. The patients' age was in the bin numbered 17.

No. of Textures	Trial A	Trial B	Trial C	Trial D	Trial E
2	548-609 2373-2403	558-608 2374-2408	554-608 2371-2403	530-612 2395-2402	555-608 2385-2402
3	556-608 2867-2908 2952-2956	558-608 2331-2399 2804-2805	557-608 2838-2894 2957-2971	549-609 2814-2865 2937-2973	557-608 2894-2906 2919-2970
4	558-608 2888-2907 2959-2972	375-425 571-696 737-770 2359-2407	553-608 1167-1217 2891-2898 2923-2976	504-617 1810-1856 2201-2204 2328-2337	537-735 739-789 2835-2882 2953-2971
5	411-417 676-687 736-770 2890-2907 2955-2964	465-495 715-719 768-774 2849-2868 2926-2959	308-334 577-712 714-762 2352-2378 2817-2868	552-609 1889-1890 1914-1922 2060-2171 2903-2953	325-359 588-648 732-771 2205-2255 2314-2363
6	668-715 734-779 1076-1093 1151-1166 1238-1270 2338-2402	390-417 589-648 742-781 1698-1717 2879-2885 2933-2961	425-480 512-613 661-726 735-782 1603-1651 2951-3000 3001-3007	560-611 1002-1069 1142-1210 1697-1745 2831-2883 2906-3000 3001-3010	284-332 575-674 720-777 1641-1680 2889-2914 2950-2973
7	427-430 557-593 665-711 714-764 1224-1273 2259-2307 2339-2389	244-283 313-316 595-619 761-772 1177-1223 2813-2849 2930-2979	508-518 536-604 680-725 743-777 1250-1298 2824-2876 2920-2997	561-611 1087-1136 1143-1207 1683-1735 2368-2416 2834-2907 2918-3000 3001-3002	449-466 562-593 640-657 745-774 1240-1286 2205-2248 2297-2309
10	411-424 580-604 632-649 733-780 1042-1078 1136-1174 1307-1370 2234-2270 2323-2372	211-255 377-425 462-497 551-602 648-692 733-775 1354-1410 1754-1787 2851-2900 2901-2994	195-206 461-474 499-532 580-590 661-692 729-777 1396-1485 1693-1705 2827-2850 2854-2901	9-57 304-305 424-463 681-701 732-778 1041-1042 1151-1168 1193-1241 2831-2857 2932-2958	9-57 96-144 314-362 486-518 567-568 686-690 763-770 820-871 2820-2868 2939-2984
15	425-428 432-460 543-591 663-711 716-764 1063-1127 1158-1164 1270-1288 1706-1754 2163-2211 2259-2307 2597-2631 2635-2656 2953-2999	240-288 293-303 410-455 583-631 649-695 732-769 1004-1024 1072-1096 1147-1167 1357-1405 1764-1812 1910-1959 2202-2246 2299-2382 2895-2943	316-364 439-454 660-681 740-762 949-954 962-977 1009-1051 1159-1163 1166-1208 1386-1403 1548-1596 1982-2012 2023-2031 2880-2908 2954-2972	246-270 509-535 632-691 714-762 825-914 1082-1105 1138-1170 1226-1277 1853-1866 1888-1889 1911-1914 1987-2021 2041-2067 2107-2155 2837-2848	281-329 449-488 515-547 641-690 714-763 764-814 951-967 1122-1145 1148-1170 1616-1640 2421-2468 2838-2877 2927-2980

(a) Texture Energy

No. of Textures	Trial A	Trial B	Trial C	Trial D	Trial E
2	195-244 379-391	191-293 359-397	300-343 361-400	163-210 812-815	156-209 784-825
3	183-316 350-399 748-846	193-298 361-398 1229-1319	121-224 767-860 2594-2635	160-209 1262-1286 1305-1374	159-211 788-837 2417-2418
4	195-310 358-398 547-597 2096-2157	284-285 362-363 812-865 2785-2794	312-321 389-395 477-523 821-845	145-222 933-944 961-986 1797-1827	165-211 831-879 2248-2252 2401-2429
5	236-284 361-400 593-641 881-885 2494-2567	301-350 351-381 805-853 1892-1938 2260-2281	262-315 343-397 453-509 812-860 2919-2967	162-214 852-880 954-961 1795-1842 2240-2334	197-236 352-397 563-610 1972-1997 2452-2482
6	212-258 380-387 574-636 1403-1435 1483-1513 2597-2690	13-61 283-319 326-369 1205-1216 1311-1314 2405-2453	193-282 335-384 496-547 1380-1428 1461-1529 2660-2769	273-304 353-384 419-429 862-906 954-972 1793-1844	199-297 356-396 487-532 0804-849 1973-1988 2594-2631
7	192-288 338-388 599-645 712-740 1235-1242 1308-1329 2067-2170	251-275 361-396 802-852 1820-1868 1924-1939 2282-2350 2560-2569	196-246 381-399 554-594 880-928 1445-1473 1486-1530 2720-2765	277-293 363-379 490-496 851-898 951-974 1382-1430 2011-2033	199-307 348-398 477-541 1228-1239 1341-1344 1993-2041 2371-2419
10	284-298 346-388 458-506 587-590 1210-1252 1312-1317 1410-1430 1477-1510 2000-2043 2607-2626	196-265 276-322 361-387 464-472 1210-1242 1363-1415 1505-1508 1537-1538 2390-2433 2543-2560	350-359 380-392 575-579 628-674 712-760 871-917 1328-1355 1455-1464 1510-1560 2022-2054	169-215 282-311 355-379 516-546 594-610 909-946 949-973 1251-1299 1809-1845 2074-2122	10-41 211-255 359-396 562-580 984-995 1236-1240 1297-1333 2406-2426 2568-2585 2634-2658
15	14-62 90-138 227-259 332-370 579-593 698-735 814-826 1123-1171 1227-1251 1318-1321 1383-1437 2129-2177 2532-2546 2682-2778	55-93 279-299 356-373 515-544 1100-1120 1207-1245 1249-1255 1377-1401 1947-1994 2694-2731 2751-2799	287-303 358-388 485-557 561-564 677-680 745-747 845-894 960-1007 1035-1078 1437-1456 1505-1529 1965-1978 2128-2190	84-94 231-261 356-386 523-546 585-633 650-684 755-819 847-895 944-976 1433-1473 2028-2058 2170-2218 2434-2474	170-171 279-306 354-382 578-623 643-652 710-758 900-946 954-995 1372-1421 1674-1696 1747-1757 1928-1998 2306-2354 2474-2478 2526-2540

(b) Texture Entropy

Table 4.22: Range of texture properties selected for the classification of the **Diag** cohort for a varying number of texture energies and texture entropies in the discriminant function. The entries are associated with the corresponding entries in Tables 4.15 and 4.16 and the values in each range of features are numbered as described in Section 3.3.4. The patients' age was placed in bins 3001–3010.

No. of Textures	Trial A	Trial B	Trial C	Trial D	Trial E
2	1062-1113	204-216	990-1011	999-1010	986-1028
	1116-1132	364-385	2730-2758	2942-2965	2951-2964
3	264-294	290-349	217-270	1063-1089	216-267
	351-382	352-393	824-875	1109-1123	801-885
	849-866	1757-1804	2151-2154	2302-2364	2954-2964
4	282-295	285-317	176-222	434-481	121-169
	360-364	365-390	272-301	931-943	216-243
	811-865	877-881	361-396	959-1001	332-371
	2721-2769	1827-1835	2127-2174	2910-2961	2361-2410
5	205-235	311-319	121-143	143-179	102-173
	338-351	381-386	286-309	296-346	243-303
	1047-1098	901-951	364-393	367-398	328-377
	1101-1149	1794-1839	874-886	1021-1024	1204-1252
	2164-2169	2734-2742	2106-2155	1097-1146	1880-1965
6	163-205	146-193	197-215	150-177	162-170
	211-250	227-275	296-344	320-356	257-291
	362-388	363-395	368-386	361-387	345-392
	1021-1024	1222-1229	899-934	1063-1066	829-840
	1145-1192	1762-1812	2803-2856	1112-1138	964-1001
	2154-2167	2342-2343	2875-2884	2922-2948	1882-1919
7	177-195	282-310	10-56	179-222	154-200
	229-273	356-387	107-131	245-292	203-251
	384-390	836-884	227-250	374-389	366-382
	1043-1099	938-968	362-386	838-895	1224-1240
	1104-1148	1797-1801	816-850	929-1003	1924-1968
	2122-2168	1873-1927	949-974	2031-2079	2814-2840
	2721-2767		2369-2403	2634-2689	2901-2903
10	19-23	118-166	194-197	68-95	194-196
	129-139	202-229	280-313	103-137	315-336
	203-251	365-389	362-389	257-327	360-410
	344-368	438-486	498-603	347-392	473-500
	1045-1093	979-1028	1078-1101	474-571	882-930
	1106-1134	1066-1205	1102-1146	814-862	941-971
	1750-1800	1784-1823	1473-1521	953-987	1465-1513
	2002-2050	1887-1931	2654-2692	1556-1590	2095-2112
	2480-2533	2807-2839	2823-2854	2209-2255	2380-2404
	2637-2685	2888-2911	2911-2927		2711-2721
15	60-74	113-133	171-185	65-113	64-82
	124-156	201-233	266-301	117-150	98-131
	215-224	362-389	345-387	313-327	267-291
	379-382	908-916	607-611	348-374	368-415
	390-437	964-977	696-733	885-897	449-497
	648-696	1153-1261	860-905	949-996	890-919
	816-864	1727-1728	953-963	1870-1920	941-1033
	1088-1110	1781-1829	1194-1238	2111-2139	1081-1134
	1123-1138	1870-1929	1764-1812	2173-2177	1443-1487
	1969-1989	2183-2214	1940-1988	2239-2297	1542-1716
	2341-2368	2367-2419	1991-2062	2683-2731	1717-1765
	2806-2830	2828-2840	2118-2151	2781-2812	1911-1972
	2891-2919	2888-2902	2360-2408	2885-2905	2706-2754
	2941-2958		2603-2609		2836-2889
			2882-2905		2898-2933

Table 4.23: Range of inertia texture properties selected for the classification of the Diag cohort for a varying number of texture inertia in the discriminant function. The entries are associated with the corresponding entries in Table 4.17 and the values in each range of features are numbered as described in Section 3.3.4. The patients' age was placed in bins 3001–3010.

described in Section 3.3.4. The genetic algorithm was used in the selection of the discriminatory regions in the histogram as well but the OD of a film would not change abruptly without some intermediate OD values. Therefore, the grey values were not entirely independent and we were justified in taking the mean for a sub-region of the histogram as a variable in the discriminant function. This was not the case, in principle, for the texture measures since they were arranged into a spectrum in an arbitrary order and there were "natural" boundaries between each texture. In practice, the selected group of textures did not always respect the divisions between different textures. For example, practically all the selected regions span several different textures formed simply by changing the parameters d and θ in the SGLD matrix[8]. This implies that the features the program selects were isotropic and independent of distance, up to the range of pixel separations that were used in this work (max $d = 16\sqrt{2}$ and $\theta \in \{45°, 135°, 225°, 315°\}$). The next most frequent "boundary crossing" was between the three quadrants which were distinguished by the order and type of 1D wavelet filter that was applied. In particular, **ga_ors** would try to combine texture values calculated from different quadrants of Figure 2.3 for the same level of the wavelet decomposition. That is, between quadrants where the label for the quadrant only differs in the last two letters of Figure 2.3. While the difference in the order of application of the 1D wavelet filters had an impact on the output of the transform, the resulting coefficients between the three quadrants were similar and it would not be surprising that the program occasionally took means of the values across these quadrant boundaries.

Finally, the last boundary occurred between the different resolutions for each iteration in the wavelet transform (quadrants in Figure 2.3 with different numbers

[8] These quantities were described in Section 2.2.2.

of letters in their labels). Sets of features which cross this boundary occurred more frequently for the texture energy and texture entropy and then predominantly for the iterations that resulted in the lowest resolution part of the transform. The intervals with this behaviour only appeared when a small number of textures was desired and it may be significant that the texture inertia which had the fewest of this type of selected regions also performed the best in classifying the images. The frequency of occurrence also varied significantly. The first type of boundary crossing, ignoring d and θ, were extremely common while regions crossing quadrants were considerably less frequent and the resolution (or scaling) level crossing regions were quite rare. The range of the specific features that were selected for the best classification performance[9] using 7 texture energy features, 6 texture entropy features and 6 texture inertia features are listed in Table 4.24. The table gives the endpoints for each range of features when the textures are arranged as described in Section 3.3.4. Each texture was described by the channel number in the artificial spectrum as well as the wavelet level (L1, L2, L3, L4 and L5), wavelet transformed quadrant (LH, HH and HL) and their (d, θ) combination as expressed as a Cartesian vector $((1,0),(0,1),\cdots$ See Section 3.3.4.).

Overall, the feature sets which were selected appear to be derived predominantly from the iterations of the transform which resulted in the lowest and highest resolution. Generally, more features were selected from the lower resolution levels than the higher ones.

The remaining study for this section allowed the genetic algorithm to select any combination of features from the moments, histogram regions, multifractal dimensions and texture measures[10]. Combining the properties allowed the creation of a

[9] The number of features were selected for the best classification performance with the fewest features. However, the trial (A–E) was selected as the trial with the median classification performance.

[10] Previously the selection was constrained to only use one property.

Property	"Spectrum"	Feature Endpoints	
	Range	Start	End
	427-430	L5, HL, (1,-1)	L2, HL, (1,-1)
	557-593	L5, HL, (8,8)	L5, HL, (16,16)
	665-711	L4, LH, (2,-2)	L4, LH, (4,4)
Energy	714-764	L4, LH, (4,-4)	L4, LH, (0,16)
	1224-1273	L3, LH, (1,-1)	L3, LH, (2,2)
	2259-2307	L2, HL, (2,0)	L2, HL, (4,-4)
	2339-2389	L2, HL, (8,0)	L2, HL, (16,-16)
	262-315	L5, HH, (2,-2)	L5, HH, (4,4)
	343-397	L5, HH, (8,-8)	L5, HH, (16,16)
Entropy	452-509	L5, HL, (2,0)	L5, HL, (4,-4)
	812-860	L4, HH, (1,0)	L4, HH, (2,0)
	2919-2967	L1, HL, (4,4)	L1, HL, (0,16)
	150-177	L5, LH, (8,-8)	L5, LH, (16,0)
	320-356	L5, HH, (4,4)	L5, HH, (8,8)
Inertia	361-387	L5, HH, (0,16)	L5, HH, (16,-16)
	1063-1066	L4, HL, (2,-2)	L4, HL, (2,-2)
	1112-1138	L4, HL, (4,-4)	L4, HL, (8,0)
	2922-2948	L1, HL, (0,8)	L1, HL, (8,-8)

Table 4.24: Texture features selected by ga_ors for 7 energy textures (Trial A), 5 entropy textures (Trial C) and 6 inertia textures (Trial D). L1, L2, L3 or L4 refers to the level of the wavelet decomposition; LH, HH or HL refers to the quadrant for the decomposed image and $(\cdot, \cdot)$ refers to the Cartesian vector that corresponds to the d and θ combination. (See Section 3.3.4.)

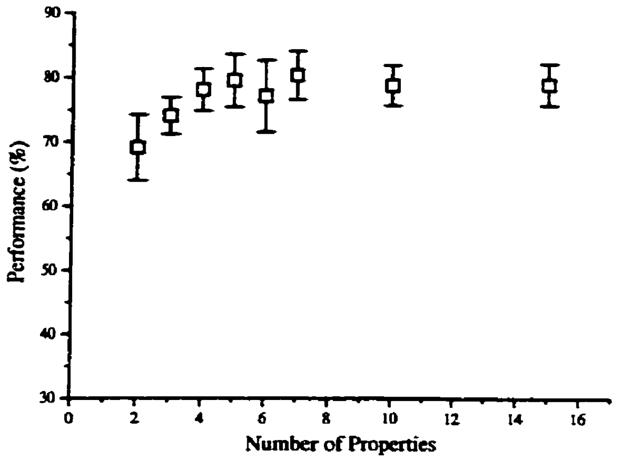


Figure 4.5: Overall classification performance for the **Diag** cohort allowing for features to be selected from all calculated properties.

discriminant function utilising properties that characterise unrelated aspects of the mammogram and, ideally, giving an improved classification accuracy. When the process was performed using all properties, the overall accuracy, the details of its performance and the selected regions are shown in Figure 4.5, Table 4.25 and Table 4.26 respectively. In order to use the genetic algorithm on the combined features, a spectrum was formed by placing the rebinned histogram data (rank ordered and binned to 256 values) in the channels numbered 1–256 (as for the selection of the histogram sub-regions). The global moments, regional moments and multifractal dimensions were entered in bins 257–356, 357–456 and 457–656 respectively. The features for these three properties were repeated for 10 bins so that, for example, the 10 moments occupy 100 bins. The patients' age was entered in bins 657–666 and the age was also repeated for the 10 bins. The texture energy, texture entropy and texture inertia was placed in the bins 667–3666, 3667–6666, 6667–9666 respectively. Within each range the textures were ordered as described in Section 3.3.4.

No. of	Trial A			Trial B			Trial C			Trial D			Trial E		
Properties	and Acc (%)			and Acc (%)			and Acc (%)			and Acc (%)			and Acc (%)		
2	22	16	69.14	29	11	73.49	20	11	69.70	25	20	63.16	25	10	76.71
	9	34		11	32		9	26		8	23		7	31	
3	24	14	74.07	27	13	72.29	22	9	78.79	28	17	72.37	24	11	76.71
	7	36		10	33		5	30		4	27		6	32	
4	26	12	74.07	30	10	78.31	23	8	80.30	27	18	72.37	25	10	78.08
	9	34		8	35		5	30		3	28		6	32	
5	27	11	76.54	32	8	79.52	25	6	81.82	28	17	72.37	29	6	82.19
	8	35		9	34		6	29		4	27		7	31	
6	26	12	76.54	29	11	77.11	24	7	84.85	28	17	72.37	30	5	84.93
	7	36		8	35		3	32		4	27		6	32	
7	30	8	81.48	32	8	78.31	25	6	80.30	28	17	73.68	29	6	83.56
	7	36		10	33		7	28		3	28		6	32	
10	29	9	79.01	29	11	74.70	24	7	78.79	30	15	75.00	28	7	82.19
	8	35		10	33		7	28		4	27		6	32	
15	30	8	82.72	30	10	77.11	22	9	78.79	31	14	78.95	30	5	84.93
	6	37		9	34		5	30		2	29		6	32	

Table 4.25: Outcome for the **Diag** cohort allowing for features to be selected from all calculated properties.

The appearance of Figure 4.5 highly resembled the curves in Figure 4.4(d) or Figure 4.4(f). The reason is immediately clear from Table 4.26. The features which were chosen were almost exclusively from the texture properties but there does not appear to be any single texture type which dominates. There was near uniform representation among the texture energy (667–3666), texture entropy (3667–6666) and texture inertia (6667–9666) but almost no features selected from any of the other extracted properties (collectively 1–666). This implies that the discriminatory power of the texture properties was much stronger than any other property that we have considered but that there was little difference in the discriminatory power among the three different textures.

There was also a significant difference in the peak classification accuracy in Figure 4.5 and when using the textures individually, Figures 4.4(d)–4.4(f). This was likely

No. of Properties	Trial A	Trial B	Trial C	Trial D	Trial E
2	1056-1270 6872-6875	1176-1310 6890-6941	1181-1334 6897-7206	1175-1339 9023-9028	1183-1343 6982-7145
3	1172-1225 5647-5799 6947-6957	1166-1319 2980-3132 6894-7018	1132-1283 3659-3666 3667-3911 6891-6918	1166-1327 2979-3131 6938-7133	1156-1311 3625-3666 3667-3673 6924-7080
4	174-266 1116-1266 5413-5564 7002-7057	1051-1360 6867-7019 8385-8538 9389-9429	632-666 667-784 1169-1322 6861-7018 8732-8856	1157-1314 2468-2575 6858-7115 8962-8989	1254-1269 2636-2792 3660-3666 3667-3811 6910-6958
5	1117-1288 3534-3601 3640-3666 3667-3925 5273-5784 6879-7098	1123-1367 6950-7029 7046-7066 8420-8572 8678-8831	1072-1117 6565-6582 6729-6881 6932-6966 6994-7030	1177-1311 3095-3246 3662-3666 3667-3791 6885-7012 7622-7747	1190-1340 3116-3269 4413-4566 6828-7055 9557-9623
6	479-536 1113-1243 2387-2538 5654-5710 6957-7004 7619-7708	68-159 1185-1317 4353-4507 7258-7487 8564-8716 9185-9485	484-535 1176-1194 3041-3111 3119-3156 3496-3648 6943-6990	1167-1311 2877-2965 6965-7091 7627-7706 9482-9633	1199-1305 1891-2043 3469-3621 3627-3666 3667-3743 6939-7022 7610-7759
7	148-299 343-536 1118-1246 5887-6039 6884-6983 7585-7725 9342-9494	457-609 1186-1332 3489-3518 3613-3645 4529-4573 5599-5664 6963-7036	1164-1192 3050-3055 3468-3620 5875-6027 6714-6866 6957-6999 7014-7054	485-539 1164-1315 5033-5185 7005-7084 7622-7715 7734-7886 9470-9566	1152-1304 2695-2844 3435-3580 3645-3666 3667-3788 6896-6993 7586-7747 8235-8389
10	486-535 2021-2173 2413-2465 3857-3867 5299-5303 5320-5352 6896-6914 7629-7779 8744-8836	84-120 1214-1285 2533-2682 3427-3666 3667-3679 3732-3774 4283-4315 4355-4404 6421-6569 8414-8569 9577-9642	629-655 966-1038 1190-1209 1768-1920 5534-5549 6999-7018 7040-7058 9320-9346 9469-9490 9500-9562	257-297 486-522 1188-1289 3351-3484 5161-5191 6925-7030 7614-7716 8439-8568 9457-9610	491-512 1204-1322 2283-2497 3654-3666 3667-3791 6888-6914 7222-7272 7645-7703 8462-8581 9143-9276
15	303-457 507-543 1137-1186 1546-1698 2467-2588 3670-3683 3935-3972 4005-4022 5397-5487 6312-6511 6831-6931 7703-7722 7750-7802 9241-9386	460-490 534-648 1149-1212 1240-1295 1639-1755 3017-3048 3485-3637 3934-3968 3984-4057 4496-4648 5400-5552 6983-7135 7519-7572 7950-8016 8894-8944	29-92 138-288 456-483 1000-1045 1361-1389 1397-1432 3120-3124 3921-3967 5165-5209 5995-5997 7159-7224 7650-7802 8175-8290 8857-9028	489-503 1022-1087 1253-1341 1382-1451 1711-1800 3017-3169 3280-3430 3876-4028 4599-4677 5717-5719 6928-6990 7695-7746 7775-7845 8897-8913	346-421 487-543 556-601 1234-1282 2474-2476 3575-3666 3667-3727 4745-4767 5798-5831 5849-6001 6398-6525 6859-7012 7017-7050 8619-8646 9173-9185

Table 4.26: Properties selected for the **Diag** cohort when allowing for features to be selected from all calculated properties.

due to the stochastic examination of the potential feature sets inherent in a genetic algorithm. For this part of the study, the total number of features was more then three times larger than when the texture energy, entropy or inertia was used individually. Therefore, the genetic algorithm was able to examine a greater proportion of the solution space, and find a better solution, when each property was examined individually as compared to the combination of all properties.

For all instances, the patients' age was included in the "spectrum" as an extra property, however, it was rarely chosen as a factor in any of the studies. The instances where it was selected seemed to occur only when a large number of features (10 or 15) were requested to be used in the discriminant function. This suggested that the discriminatory power of the selected features are nearly age independent. However, the evidence is circumstantial and a more explicit analysis for the presence of an age dependence was considered in Section 4.3.

4.2 Contralateral Mammograms

When the mammograms containing a malignancy and the set of normal cases (**Diag** class) was selected, a large subset of the collected images was excluded. Recall that the selection of the training and test groups was made at random from a pool of images with twice as many normal mammograms as abnormal ones[11]. Additionally, the number of normal and abnormal cases was kept approximately equal[12]. Therefore, there were many normal images which were not used in any of the prior training or test groups. As well, all the mammograms of the contralateral breasts from the women

[11] The mammograms for both breasts for each patient was included in the total pool of images and a random sample was selected.

[12] The equality between the groups varied slightly due to the random selection of cases.

with a unilateral abnormality were not included in the pool of images for the **Diag** class. It would be instructive to examine the results of the classification system on this set of images.

For the mammograms of the breasts contralateral to the side where an abnormality was found, the classification should be "normal" if the radiologists' diagnosis is taken. However, it was desirable to distinguish this set of images from the mammograms from subjects with both mammograms diagnosed as normal. Therefore, for this section the abnormal classification was altered to describe the *patient* diagnosis rather than the diagnosis of the breast itself. Using this altered definition, a patient with an abnormality in either breast would carry the abnormal classification and only if both breasts are normal were the mammograms given the normal classification. It should be noted that since the clinically normal images from the two different categories were kept in two isolated groups, it would be a simple matter to generate the classification performance for the original definition of normal and abnormal mammograms by regrouping the elements in the confusion matrices.

In order to evaluate the system on this set of images, linear discriminant analysis was used, which requires a discriminant function to be formed

$$Z = \sum_{i=1}^{m} c_i x_i \tag{4.2}$$

for m features, x_i, and m constants, c_i. There were two aspects that must be addressed. First, the most significant set of variables, x_i, needed to be found and can be performed using a genetic algorithm or an exhaustive search of all possible combinations of variables. Second, the set of coefficients, c_i, which produced the specific function that classifies the images the "best" must be determined. This can be performed using LDA.

For this part of the analysis, the genetic algorithm was not re-applied to the data set to select the best set of x_i. Rather, the properties which were selected in Section 4.1.2 were considered. However, due to several random factors such as the selection of five training and test groups, there was a certain amount of variation in the set of best properties that were found for each trial.

Since $\{x_i\}$ or $\{c_i\}$ can be varied to form different discriminant functions, two approaches to the evaluation were used.

1. Only the $\{c_i\}$ was calculated for each property. The $\{x_i\}$ was selected from the feature sets using the **Diag** cohort (Section 4.1.2) and taken from Tables 4.19–4.23. For each property, the specific feature set was taken as that which resulted in the median classification performance for the 5 trials, A–E, but with the best classification accuracy for the various numbers of features in the discriminant function (1–15). For example, for the sub-regions of the histogram the feature set for Trial C using 5 sub-regions was selected. The various feature sets that were selected for each property is described in detail below.

   Then, 5 different $\{c_i\}$ were calculated, using LDA, by using the test datasets for the 5 trials corresponding to the selected feature set. i.e. for the sub-regions of the histogram, 5 different $\{c_i\}$ were formed using the test sets for Trials A–E, for 5 sub-regions. The performance of the 5 discriminant functions were then tested using the contralateral images. The median and standard deviation was taken as the classification accuracy and uncertainty for these features (**Uniform**).

2. Both $\{x_i\}$ and $\{c_i\}$ were varied simultaneously (**Individual**). For this case the classification performance of the discriminant functions created for the 5 trials,

A–E in the **Diag** cohort, was tested on the contralateral images. The five functions were investigated only for the number of features which gave the best classification when the number of features was varied from 1–15. Therefore, for the sub-regions of the histogram, the discriminant functions which were used to give the results in Table 4.11 when using 5 sub-regions were tested for their classification ability on the contralateral mammograms.

Recall that for each property that has been considered, histogram subregions, global and regional moments, multifractal dimension, etc., the classification performance was examined for a varying number of features. For example, when considering the global moments, the single best moment was found, the combination of the best pair of moments, best triplet and so on. For this part of the analysis, the feature that gives the highest classification performance for each property was the only feature set considered, for the most part. However, there were several cases where the best feature set required an excessively large number of features or features that tend to be unstable. For these cases an alternative feature set was selected that has comparable performance to the feature set with the highest accuracy. The actual feature sets that were used for the analysis of the **Uniform** features will be discussed in turn.

Global Moments For this property the best performance was attained when using 4 features but for many of the trials, some of the high moments are needed, which tend to be very sensitive to small changes in the histogram distribution. Therefore two moments were used (median outcome: Trial C) which has nearly the same overall performance and only used the lower moments. See Figure 4.4 and Table 4.20(a).

Regional Moments A similar situation exists for this property as for the Global

Moments. The best performance occurred when four moments were used but two has nearly the same outcome without requiring the high moments (median outcome: Trial D.). See Figure 4.4 and Table 4.20(b).

Histogram Subregions This case is straightforward. The best performance occurred for 5 regions (median outcome: Trial C). See Figure 4.4 and Table 4.19.

Multifractal Dimensions The calculations using 2 to 5 dimensions had comparable performance to each other. Therefore, 2 dimensions were used as this resulted in the simplest discriminant function (median outcome: Trial E.). See Figure 4.4 and Table 4.21.

Energy In order to balance the best performance with the smallest number of features, 7 energy textures was selected for this texture property. (median outcome: Trial A). See Figure 4.4 and Table 4.22(a).

Entropy Again, there was nearly equal classification performance when using $\gtrsim 5$ entropy textures. Therefore five features were used (median outcome: Trial C). See Figure 4.4 and Table 4.22(b).

Inertia This also seemed to have the best performance/smallest number of features at 6 features (median outcome: Trial D). See Figure 4.4 and Table 4.23.

4.2.1 Results

An overview of the classification performance was shown in Figure 4.6. The figure shows the median result when using the **Uniform** and **Individual** feature sets for each property that was considered for this work. Clearly, there was little difference

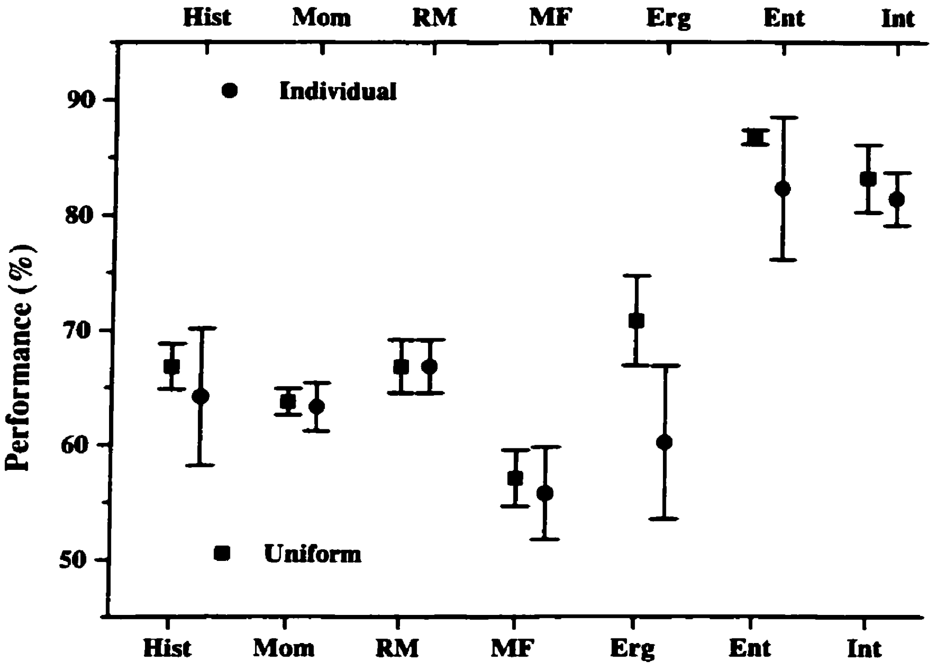


Figure 4.6: Overall classification performance for the **Individual** and **Uniform** feature sets when applied to the cohort of remaining images. Note that the upper or lower horizontal axis is associated with the data series in the legend closest to each respective axis. The properties are listed in the order: Histogram sub-regions (Hist), Global moments (Mom), Regional moments (RM), Multifractal dimensions (MF), Texture energy (Erg), Texture entropy (Ent) and Texture inertia (Int).

Property	Trial A			Trial B			Trial C			Trial D			Trial E		
Hist	71	61	56.19	94	38	69.03	85	47	64.16	98	34	71.68	88	44	63.27
	38	56		32	62		34	60		30	64		39	55	
Mom	74	58	65.49	69	63	63.72	71	61	63.27	56	76	59.73	70	62	62.83
	20	74		19	75		22	72		15	79		22	72	
RM	94	38	68.58	89	43	65.49	88	44	66.81	65	67	62.83	86	46	68.14
	33	61		35	59		31	63		17	77		26	68	
MF	63	69	55.75	55	77	47.79	66	66	57.08	101	31	57.52	56	76	53.10
	31	63		41	53		31	63		65	29		30	64	
Energy	120	12	72.57	107	25	60.18	121	11	68.58	101	31	57.08	115	17	59.29
	50	44		65	29		60	34		66	28		75	19	
Entropy	114	18	82.30	106	26	85.40	117	15	86.73	96	36	71.68	116	16	77.43
	22	72		7	87		15	79		28	66		35	59	
Inertia	113	19	79.65	109	23	77.88	115	17	81.42	114	18	83.19	110	22	83.19
	27	67		27	67		25	69		20	74		16	78	

Table 4.27: Classification details for the **Individual** feature set and cohort of remaining images for each property under consideration. Both the confusion matrix and overall performance (%) are included.

in the results between the two feature sets. The consistency was also evident from an examination of the classification details, Tables 4.27–4.28. The results are also consistent with those given for the **Diag** cohort with the exceptions of the Global Moments and texture Energy. The Global Moments had a lower performance as compared to the results for the corresponding properties on the **Diag** cohort. The texture Energy was of particular importance since it had comparable performance to the other textures on the **Diag** cohort ($\sim$ 80% from Figure 4.4) but considerably lower performance for the **Uniform** feature set on the set of contralateral images ($\sim$ 70%) and lower still for the **Individual** feature set ($\sim$ 60%). The texture entropy and inertia exhibited high classification rates ($\gtrsim$ 80%) for most of the cohorts and feature sets shown so far. Therefore the texture entropy or inertia may be a better choice as a classification property due to their similar performance over more diverse

Property	Trial A			Trial B			Trial C			Trial D			Trial E		
Hist	93	39	66.81	86	46	65.93	85	47	64.16	89	43	69.03	93	39	68.58
	36	58		31	63		34	60		27	67		32	62	
Mom	69	63	63.72	69	63	63.72	71	61	63.27	61	71	61.06	68	64	63.72
	19	75		19	75		22	72		17	77		18	76	
RM	94	38	68.58	89	43	65.49	88	44	66.81	65	67	62.83	86	46	68.14
	33	61		35	59		31	63		17	77		26	68	
MF	63	69	55.75	77	55	59.73	66	66	57.08	101	31	57.52	56	76	53.10
	31	63		36	58		31	63		65	29		30	64	
Energy	120	12	72.57	118	14	70.80	121	11	70.80	109	23	63.27	118	14	65.93
	50	44		52	42		55	39		60	34		63	31	
Entropy	117	15	87.61	114	18	86.73	117	15	86.73	110	22	86.73	117	15	88.05
	13	81		12	82		15	79		8	86		12	82	
Inertia	119	13	78.32	112	20	83.19	118	14	83.19	114	18	83.19	115	17	77.43
	36	58		18	76		24	70		20	74		34	60	

Table 4.28: Classification details for the **Uniform** feature set and cohort of remaining images for each property under consideration. Both the confusion matrix and overall performance (%) are included.

conditions.

Both properties produced comparable results on the **Diag** cohort for the selected number of textures but the variation of the classification performance when the entropy textures were used on the five trials for the **Individual** and **Uniform** feature sets were greater than those for the inertia. On the other hand, the false positive rate tended to be higher for the inertia than the entropy.

4.2.2 Conclusions

There were several additional observations which can be drawn from these results. For example, with the exception of the properties mentioned above (texture energy), the performance of all the remaining properties were comparable when tested on the **Individual** and **Uniform** feature sets. The results were also generally comparable to the classification results obtained when the features were applied to the **Diag** cohort, from where they were originally selected.

Recall that the **Uniform** features refer to using the same variables in the discriminant function but allowing the coefficients to vary by changing the images used in determining the coefficients. On the other hand the **Individual** features refer to the results when allowing both the variables and coefficients to vary in the discriminant function. The similarity in the results between the two approaches suggest that the selection of features was robust against variation due to the distribution of images between the training and test groups, variation due to the random nature of the genetic algorithm and (possibly) variation inherent between patients.

In addition, the classification performance of the various classifiers on these contralateral images were on the same order of magnitude (percent difference of ~ 6) as as for the mammograms diagnosed with an abnormality (**Diag** group). That is, clinically normal mammograms appeared to be classified as abnormal if the contralateral mammogram contains an abnormality. The exception is when the texture energy was considered where the classification accuracy was reduced by a percent difference exceeding 10 for the **Contra** images compared to the **Diag** cohort.

One possibility for the similarity in the classification accuracy between the **Diag** and **Contra** cohorts was that the left and right breasts for the majority of patients were symmetric enough that the calculated properties were similar. Therefore, if a mammogram with an abnormality was classified correctly then the contralateral mammogram is likely to be classified correctly as well. This explanation is compelling from an examination of the actual texture properties which were selected. The textures for the lower resolution components of the wavelet transform were favoured over the more detailed ones. The lower resolution components can only distinguish broad overall characteristics such as the mammographic density. Boyd *et al.* in [Boyd et al., 1995] suggested, from some of their unpublished data, that there is a high degree of left/right symmetry with respect to the density.

Additionally, it would be expected that the same patients would be misclassified regardless of the mammogram examined. In particular, for the histogram sub-regions, global moments, multifractal dimensions and texture energy, the fraction of misclassified mammograms which came from the same patient was greater than or equal to 64%[13] and for the global moments it was as high as 79%. The remaining properties, regional moments, texture entropy and inertia, still had a considerable fraction of misclassifications for the same patient at 55%, 50% and 48% respectively although they

[13] The observations for the histogram and textures only involved the actual trial selected for the **Uniform** feature set while for the remaining properties the results represent the average over trials A to E.

were somewhat lower than for the previously mentioned properties. It is also important to note that since the textures performed much better than the other properties, there was a considerably smaller total number of misclassified mammograms for the texture entropy and inertia compared to the spectral properties or the multifractal dimensions.

An alternative explanation to left/right symmetry is that the program may be detecting true characteristics of disease. For example, the presence of the malignancy may be producing some agent which caused a global and detectable influence on both breasts simultaneously. Both possible explanations for the observed results can be tested with the appropriate data. An analysis of mammograms from the same patient over a long period of time should reveal a change in the classification as the malignancy develops, for the cancer cases. On the other hand, testing whether the effects are due to an inherent symmetry, which would be more important as a risk factor than as a diagnostic tool, is possible through the analysis of a set of images from patients who do not have symmetric mammograms, with approximately half of the cases falling into both the normal and abnormal classifications. Creating such a set of images can be difficult since the number of patients who satisfy this criterion represent a relatively small proportion of the population and a study of this nature is beyond the scope of this thesis.

Regardless of whether the program was detecting features characteristic of disease or an inherent left/right symmetry, the results (above) along with the fact that the feature sets had been selected to distinguish normal/abnormal classes was suggestive that the selected texture properties was of interest in the classification of mammograms. Further work is needed to identify the exact nature underlying the observed behaviour. The outcome of such further studies will determine how the system can be used in practice.

4.3 Age Dependence

Since age is a significant risk factor for breast cancer it would be prudent to perform a more detailed examination of whether there is any age dependence in the results. It is also possible that the classification performance can be improved by optimizing the choice of feature sets for many smaller age ranges.

The results thus far seem to imply that there was little significant age dependence in the selected features. The analysis of the **Diag** class included the patients' age as an additional feature. Although there were a few exceptions, for the vast majority of the cases, the age was not selected for the set of "best" features. This is, however, rather indirect evidence. For the analysis presented in this section, a new cohort was created from the images based on the patient's age. A search for direct evidence of an age dependence was then made.

4.3.1 Method

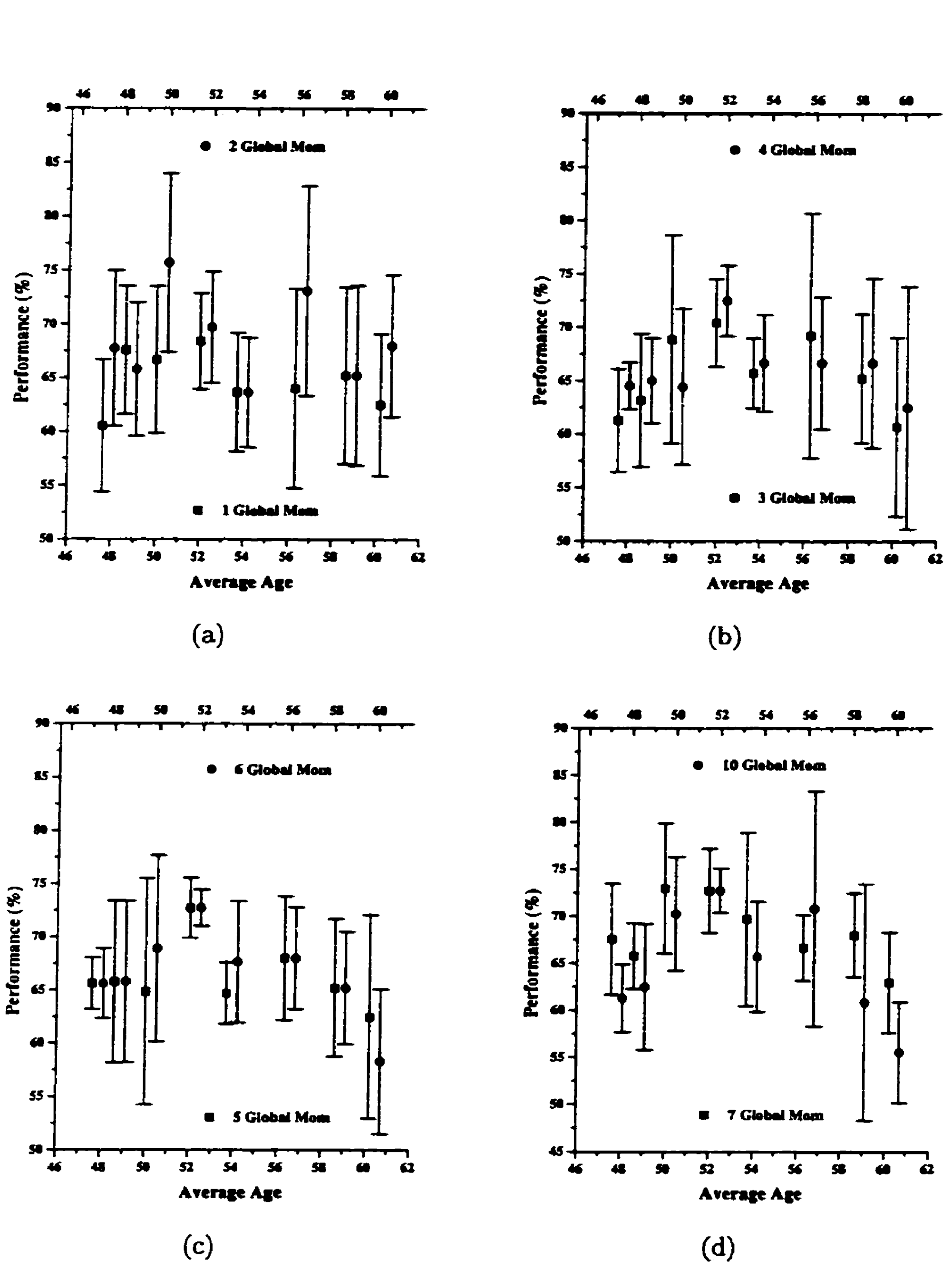
The most straightforward method of dividing the **Diag** pool of images into age groups was to choose a separate group for each decade. The difficulty with this approach was that the number of images in the 50–59 age range would be considerably larger than for the 40–49 or 60–69 age groups. Another consideration was that this approach only gives 3 different age groups which makes it difficult to clearly identify a pattern in the results. As an alternative the images were divided into eight age groups, 40–54, 42–56, 44–58, · · ·, 54–68. This gave a sufficient number of data points to make trends in the data clearer and allowed easier identification of anomalous results (noise). The drawback with this approach was that the groups were no longer disjoint[14]. From this point the analysis proceeded as for the **Diag** cohort in Section 4.1.2 for each age group. Briefly,

- Five random divisions into training and test sets were made with each age group.
- The best set of 1, 2, · · ·, 7, 10 and 15 features from each property under consideration (moments, multifractal dimensions, etc.) was found using an exhaustive search or a genetic algorithm as appropriate.
- The overall performance of each set of features was evaluated using LDA for each property.

4.3.2 Results

It would not be surprising to observe differences in the functional relationship of the classification performance with the patients' age when using different properties in the classifier, unless the properties are dependent. The overall classification performance on the various age groups when global moments were used are shown in Figure 4.7. The classification was performed using linear discriminant analysis and allowing 1–7 and 10 moments to be used in the discriminant function. Similarly, the results for the Regional Moments are shown in Figure 4.8, the sub-regions of the histogram for 2–7, 10 and 15 regions in Figure 4.9, Figure 4.10 for the multifractal dimensions (1–7 and 10 dimensions), while the energy, entropy and inertia results (for 2–7, 10 and 15

[14] The size of each group may still be too small but the statistics of any study could always stand some improvement.



(a)

(b)

(c)

(d)

Figure 4.7: Age dependence of Global Moments of the **Diag** cohort when considering 1–7 and 10 moments in the discriminant function. In each case the upper or lower horizontal axis is associated with the data series in the legend closest to each respective axis.

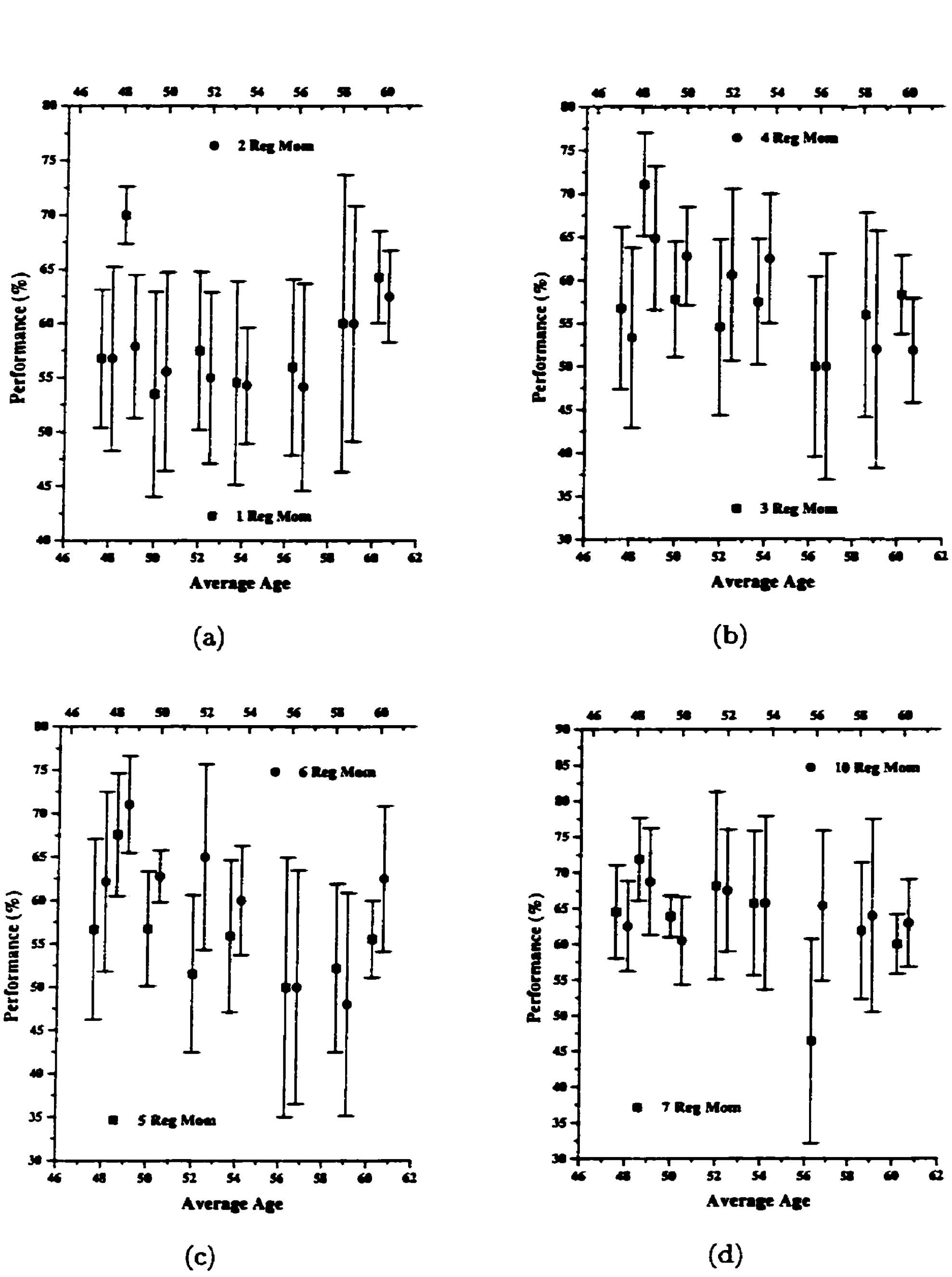
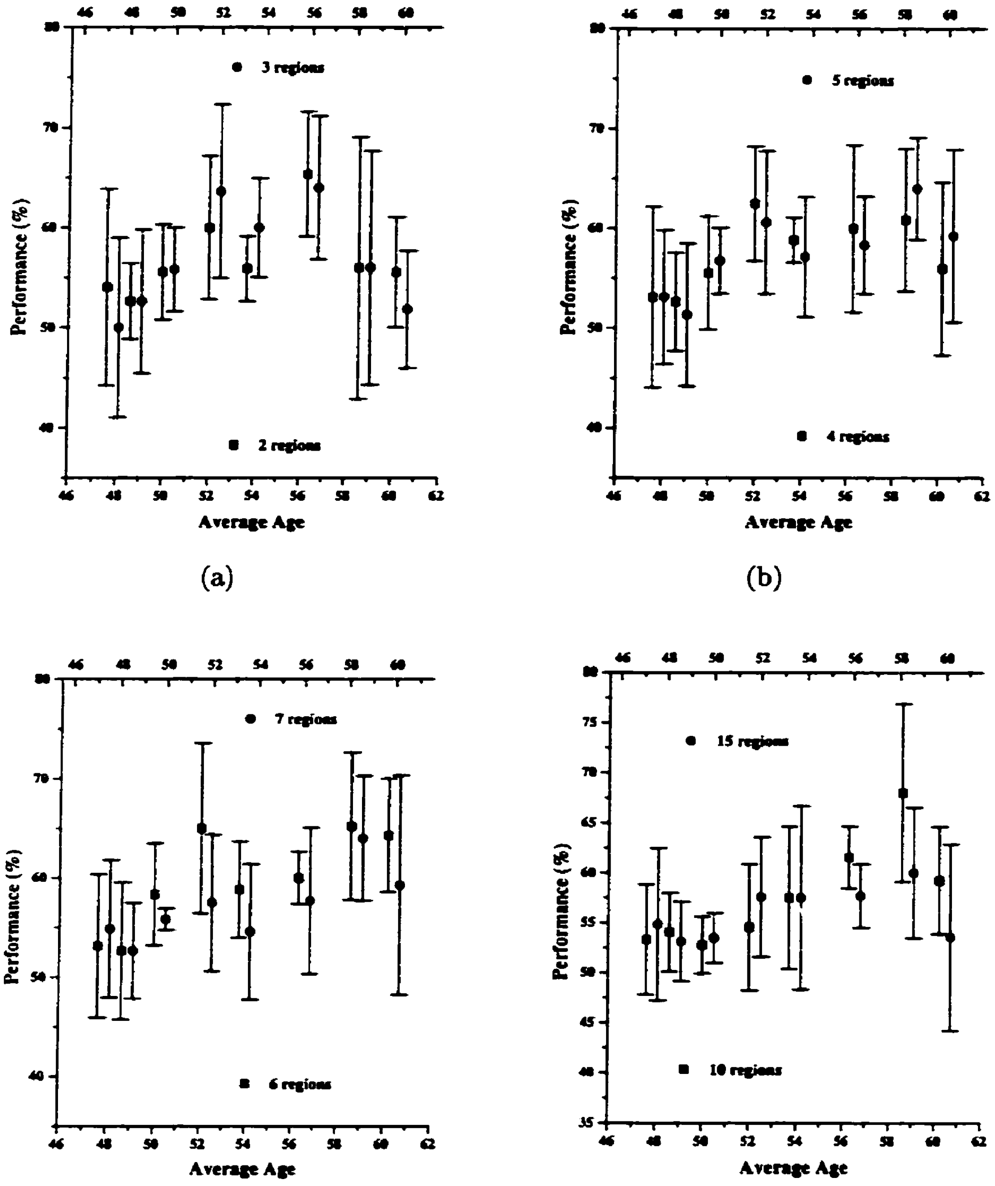


Figure 4.8: Age dependence of Regional Moments of the **Diag** cohort when considering 1–7 and 10 moments in the discriminant function. In each case the upper or lower horizontal axis is associated with the data series in the legend closest to each respective axis.



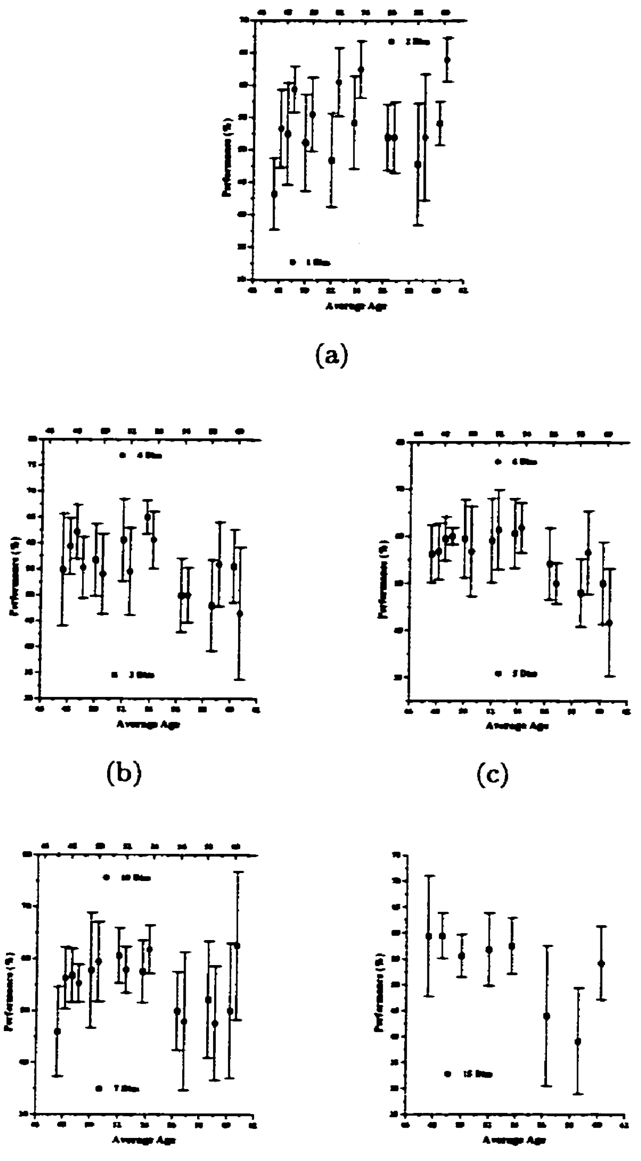
(a)

(b)

(c)

(d)

Figure 4.9: Age dependence of Histogram sub-regions of the **Diag** cohort when considering 1–7 and 10 regions in the discriminant function. In each case the upper or lower horizontal axis is associated with the data series in the legend closest to each respective axis.



(a)

(b)

(c)

(d)

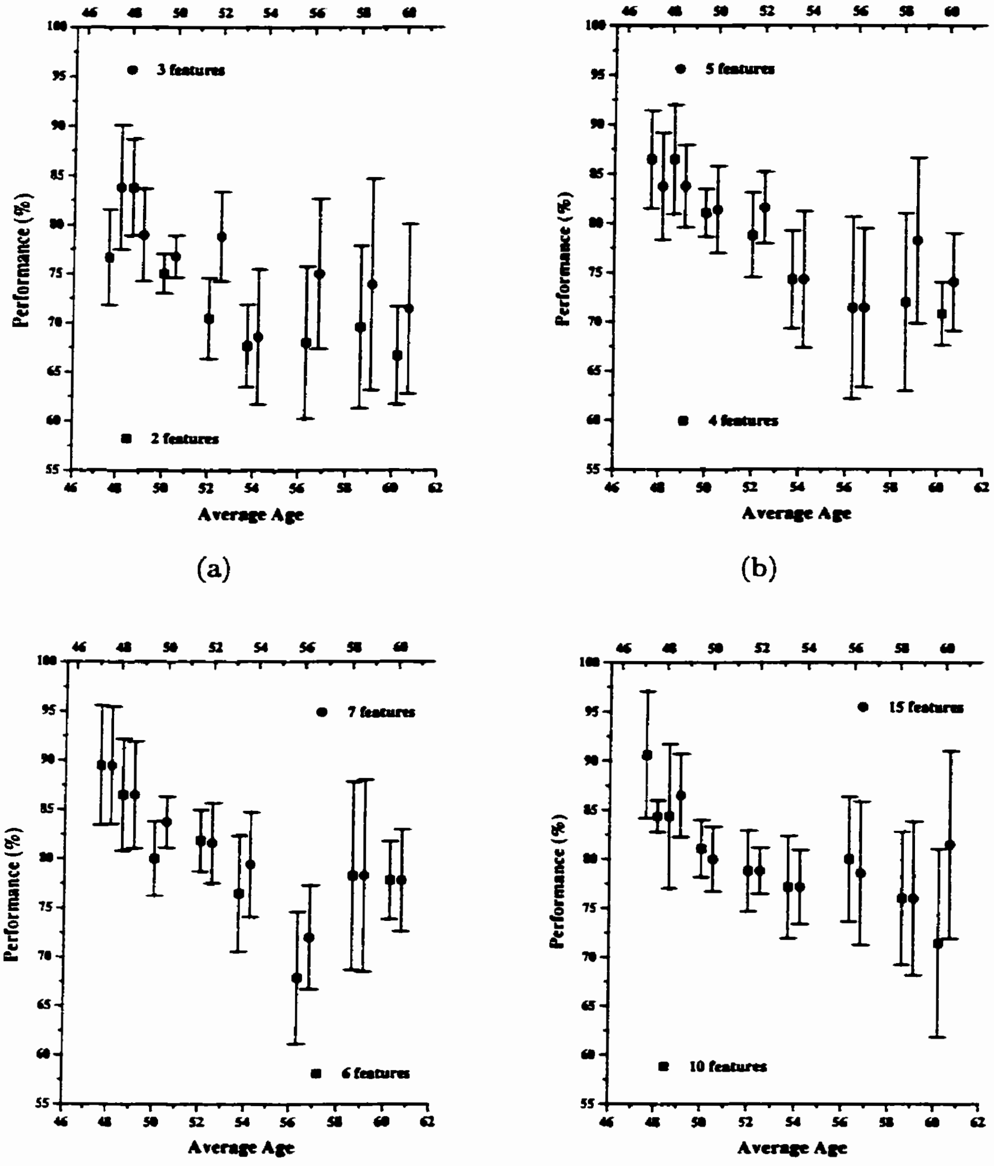
(e)

Figure 4.10: Age dependence of Multifractal dimensions of the **Diag** cohort when considering 1–7 and 10 dimensions in the discriminant function. In each case the upper or lower horizontal axis is associated with the data series in the legend closest to each respective axis.

textures) appear in Figures 4.11–4.13 respectively. In many of the instances, there appeared to be a clear age dependence and in some cases the function was rather non-linear, as in Figure 4.7(d) or Figure 4.12(a). However, the majority of these feature sets either classify the images poorly or utilise more features than would be desirable. Therefore our attention was focussed on the particular feature sets selected in Section 4.2.

A least squares fit of polynomials in the age up to the third degree was made for each data set corresponding to the feature sets selected in Section 4.2. The fits themselves were performed using a commercial product, Jandel Scientific's Table Curve. In addition, a weighting was applied to each datum proportional to its uncertainty. However, the uncertainty in the results given previously was taken as the standard deviation for a small number of points and may not be representative of the uncertainty. Therefore, for this part of the analysis, the mean of the standard deviation values were used for the uncertainty for the points in each dataset. That is, a different uncertainty was found only for different properties and different numbers of features used in the discriminant function. i.e. for each curve in Figures 4.7–4.13.

Table Curve ranked the three polynomials by the quality of the fit based on the root mean square error. When strictly considering the fits' ranking, a non-linear fit would often appear to be appropriate but the final choice of the best fit polynomial was based on a combination of the ranking of the fits, using the fit standard error (FSE or root mean square error) and a partial F-test for statistical significance [Bevington, 1969, Flury and Riedwyl, 1988]. An F-test evaluates the likelihood (probability or p) of two samples being drawn from the same distribution. In this case the partial F-test was performed using the residuals of two candidate polynomials



(a)

(b)

(c)

(d)

Figure 4.11: Age dependence of texture energy of the **Diag** cohort when considering 2–7, 10 and 15 textures in the discriminant function. In each case the upper or lower horizontal axis is associated with the data series in the legend closest to each respective axis.

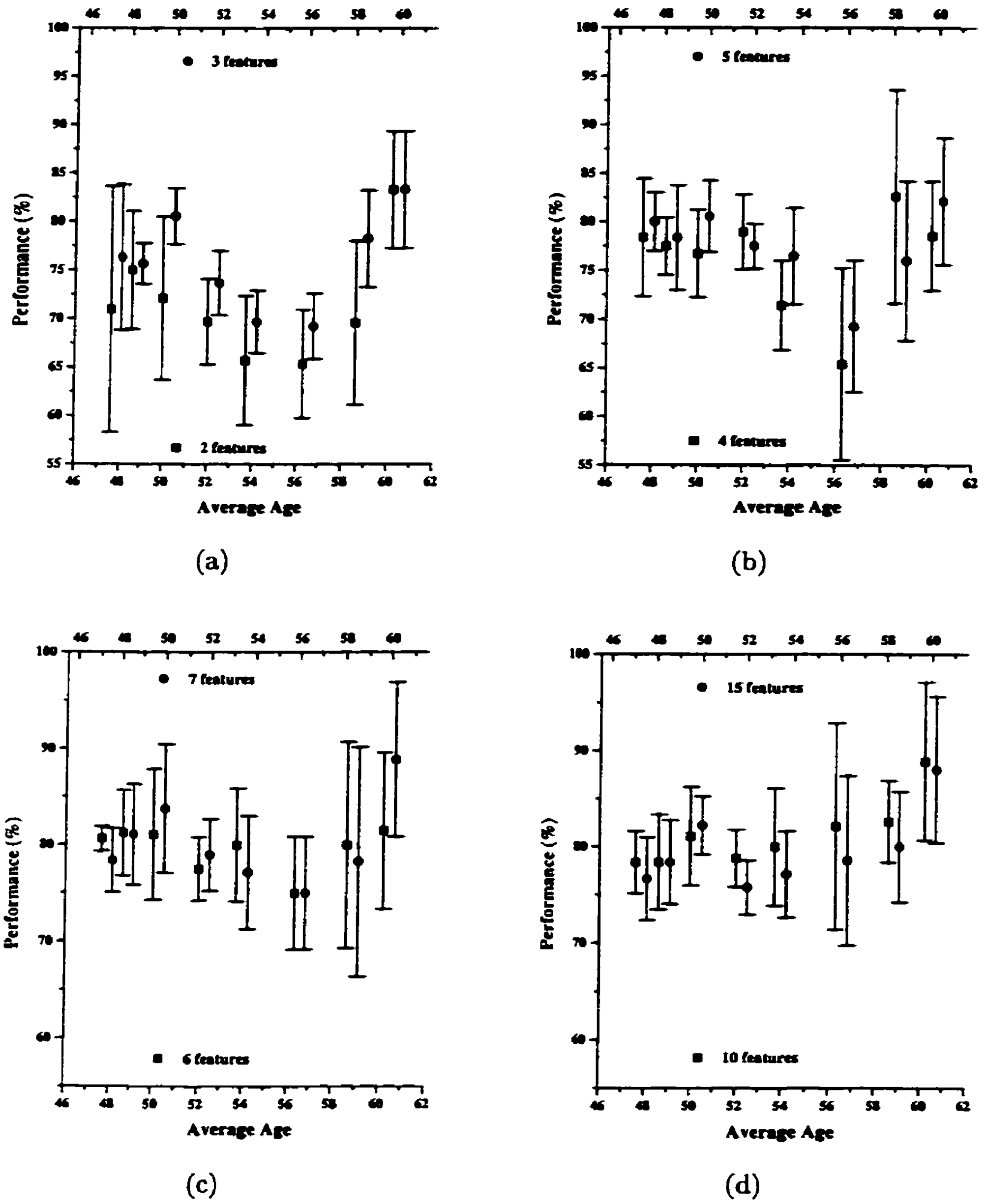


Figure 4.12: Age dependence of texture entropy of the **Diag** cohort when considering 2–7, 10 and 15 textures in the discriminant function. In each case the upper or lower horizontal axis is associated with the data series in the legend closest to each respective axis.

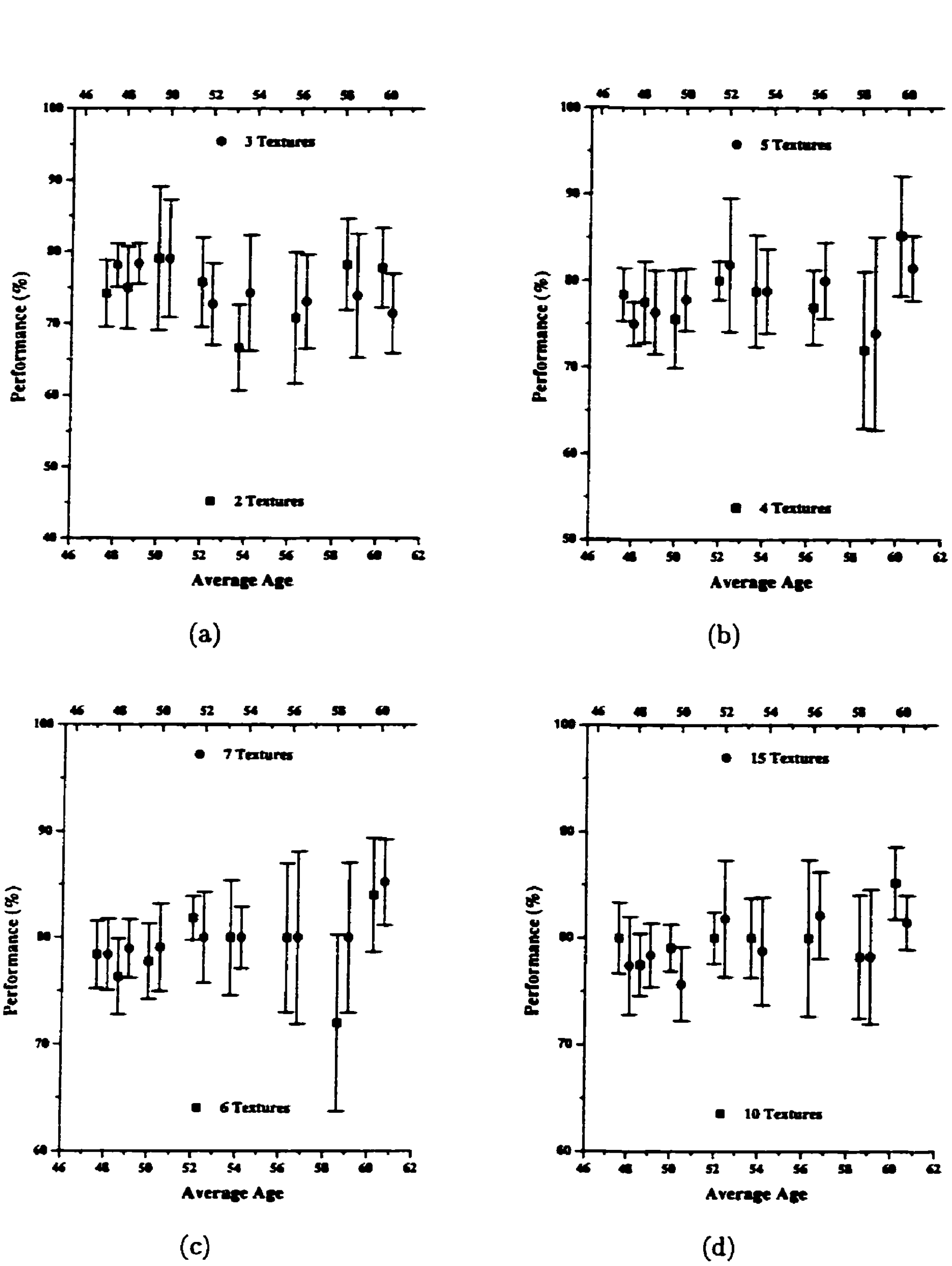


Figure 4.13: Age dependence of texture inertia of the **Diag** cohort when considering 2–7, 10 and 15 textures in the discriminant function. In each case the upper or lower horizontal axis is associated with the data series in the legend closest to each respective axis.

Property	Degree of Polynomial	Partial F comp. to const (p)	Partial F comp. to linear (p)
Global Moments	1	0.103 (> 0.75)	N/A
Regional Moments	3	15.7 (0.01)	17.0 (0.01)
Histogram Regions	2	5.27 (0.06)	1.31 (0.33)
Multifractal Dimensions	3	1.66 (0.34)	2.44 (0.22)
Texture Energy	2	17.7 (0.006)	7.68 (0.04)
Texture Entropy	3	7.66 (0.04)	10.8 (0.02)
Texture Inertia	1	0.13 (0.74)	N/A

Table 4.29: Partial F-test results for best polynomial fit (according to the fit standard error up to degree 3) for the age dependence data.

and if the sample residuals were statistically similar (p greater than a critical value) then the addition of the extra parameter in the fitted function did not improve the fit significantly. The lower order polynomial was then selected as the best fit. On the other hand, a statistically significant difference (p less than a critical value, say 0.05) in the partial F-test indicated the converse and the best fit function was selected on the basis of its (FSE) ranking. Additionally, since the uncertainty of the points was quite large for many of the cases, the best fit polynomial was always compared to a linear fit as well as to the weighted average of the points (i.e. a horizontal line). For both situations the fit was tested for a statistically significant improvement in the description of the data over a linear fit and a horizontal line. The partial F-test values for the best fit (FSE) polynomial compared to a constant and a straight line along with their p values are shown in Table 4.29.

If a relatively high value of p was selected as the cutoff for statistical significance, say 0.05, then the use of the regional moments and texture entropy appear to produce results with a cubic dependence on age while the texture energy gave results with a quadratic age dependence and classification results for the remaining properties (global moments, histogram sub-regions, multifractal dimensions and texture inertia) were independent of age. However, if p is reduced to 0.01 only the results for the regional moments still exhibit an age dependence and for $p = 0.001$, which is not unreasonable, none of the properties show any age dependence in the classification results.

In summary, if the critical value for p was taken as 0.05 the age dependence of the various feature combinations vary for different properties. The combination of 2 regional moments and the combination of 5 texture entropies that were selected show a cubic age dependence in the results. The classification ability for 7 texture energies show a quadratic dependence on age and the remaining feature sets (2 global moments, 5 histogram subregions, 2 multifractal dimensions and 6 texture inertial features) were *independent* of age. The characteristics of the dataset would suggest that these fits should only be accepted if the dependencies were very clear. Therefore for $p = 0.001$ the classification ability of all the properties are independent of age. This conclusion is supported by the results of the selection of the best feature sets where the age was not selected in combination with any of the properties[15].

4.4 Scanner Dependence

One characteristic of the data set that caused some concern was that the x-ray scanner used to digitise the film had some correlation with the classification of the images. Specifically, the majority of the Normal cases were digitised using the DBA scanner while the majority of the Abnormal cases used the LS scanner. Therefore, features

[15] This was for the feature sets selected for the highest classification accuracy with the fewest number of features. See Section 4.2.

Property	Trial A			Trial B			Trial C			Trial D			Trial E		
Histogram	20	14	58.82	18	16	52.94	20	14	58.82	15	19	44.12	13	21	38.24
	0	0		0	0		0	0		0	0		0	0	
GM	20	14	58.82	17	17	50.00	20	14	58.82	17	17	50.00	20	14	58.82
	0	0		0	0		0	0		0	0		0	0	
RM	13	21	38.24	16	18	47.06	17	17	50.00	11	23	32.35	15	19	44.12
	0	0		0	0		0	0		0	0		0	0	
MF	17	17	50.00	15	19	44.12	17	17	50.00	10	24	29.41	16	18	47.06
	0	0		0	0		0	0		0	0		0	0	
Energy	22	12	64.71	19	15	55.88	23	11	67.65	19	15	55.88	21	13	61.76
	0	0		0	0		0	0		0	0		0	0	
Entropy	14	20	41.18	9	25	26.47	13	21	38.24	10	24	29.41	10	24	29.41
	0	0		0	0		0	0		0	0		0	0	
Inertia	17	17	50.00	15	19	44.12	15	19	44.12	13	21	38.24	15	19	44.12
	0	0		0	0		0	0		0	0		0	0	

Table 4.30: Classification details for the Abnormal mammograms digitised using the DBA x-ray scanner. The table shows the overall classification performance (%) and confusion matrix for each property using the feature set of Section 4.2

which can be used to characterise the specific scanner would also appear to be able to correctly classify the mammograms scanned with the respective digitisers. For the results presented earlier, it was assumed that the inherent patient to patient variation masked any dependencies due to the scanner after the normalizations removed the obvious scanner characteristics. There were insufficient data to fully test this assumption but a small number of cases *were* present that did not have the scanner/patient outcome correlation. Namely, 34 patients with a biopsy confirmed malignancy had their mammograms digitised with the DBA scanner. The classification ability of "best" set of features for each property, as described in Section 4.2, was tested just using this group of images. The results are shown in Table 4.30 which indicates that the classification performance is lower for all properties and many are close to 50%.

Property	Trial A	Trial B	Trial C	Trial D	Trial E
Histogram	0.34	0.13	0.34	0.02	0.00
GM	0.12	0.01	0.12	0.01	0.12
RM	0.01	0.10	0.18	0.00	0.05
MF	0.15	0.04	0.15	0.00	0.08
Energy	0.01	0.00	0.02	0.00	0.00
Entropy	0.00	0.00	0.00	0.00	0.00
Inertia	0.00	0.00	0.00	0.00	0.00

Table 4.31: Total probability of obtaining classification results as given in Table 4.30 or poorer, assuming a binomial distribution.

The likelihood that the change in performance observed in Table 4.30 was due to the small sample size can be calculated, with a few judicious assumptions. If a set of images, with a known classification, were assigned at random into one of two categories such that the assigned category was correct with probability p_c, then the distribution of correct and incorrect cases follow a binomial distribution. A sample of N cases with at most N_c correctly classified and $N - N_c$ incorrectly classified cases can be calculated as the total probability of obtaining $N - N_c, N - N_c + 1, \cdots, N$ incorrectly classified cases or

$$P(N - N_c; N, p_c) = \sum_{i=N-N_c}^{N} \binom{N}{i} p_c^{N-i}(1 - p_c)^i \tag{4.3}$$

If the probability for a correct classification, p_c, is taken as observed in Section 4.1.2, for the **Diag** cohort, then the total probabilities for each trial and each property are given in Table 4.31. For example, the table (4.31) indicates that the total probability, due to random chance alone, is 0.12 for obtaining at most 20 correctly classified cases[16] using the global moments and the distribution of cases as in Trial C. Since the

[16] From Table 4.30.

value is small it would seem to indicate that the poor classification accuracy is not simply a statistical effect. However, there may be small sample size effects being exhibited in the calculations. In particular, the probabilities for correctly classifying the mammograms, p_c, were taken from a different and significantly larger sample of cases and may not be accurately reflected in this smaller sample. In addition, a binomial distribution only is an approximation to the act of randomly classifying the mammograms. For Equation (4.3), the probability for a correct classification, p_c, is assumed to be constant. p_c actually changes somewhat as the number of mammograms to be assigned to the classes is depleted. Clearly, this approximation is more valid as the total number of cases increases and deviations can be expected for a small number of cases.

It should also be noted that the corrections that were applied to the images encompass the full range of normalizations that can be applied in practice and should be sufficient to compensate for differences in the scanner characteristics. However, the significant variability in the appearance of the mammograms from patient to patient combined with the non-uniformity in the appearance of the disease made it very difficult to assess the quality of the correction procedure with a random sample of mammograms from the population. Therefore, the use of data specifically intended for testing the correction procedure is imperative for this type of study.

4.5 Conclusions

Superficially, the performance of the system for the density grade classification, was disappointing. However the regional skewness, that Byng *et al.* [Byng et al., 1996b, Byng et al., 1999] found to be a significant risk factor also had a similar performance for our sample of images. The difficulty that was encountered was that the sample itself was quite small for this type of investigation[17] and the distribution of cases among the density grades was very non-uniform. For the small sample size, there is little to discuss; atypical results are very common if the sample size is not sufficiently large and the more subtle an effect, the larger the necessary sample size. The calculated properties were being used to quantify a characteristic of mammograms which is extremely variable (the mammographic density). Without a sufficiently large sample of images that is representative of the amount of variation present in the population it is unlikely that the features that are useful to classify the density can be identified even through very powerful techniques such as a genetic algorithm.

The distribution of cases among the density grades, on the other hand, requires some elaboration. This unevenness was also partly due to the sample size but in addition to this, the cases in the DDSM were selected to fall into normal/abnormal/abnormal but benign categories without regard to their density grade classification. This resulted in the majority of the cases falling into the middle two density grades (~ 15) with few ($\lesssim 3$) in the lowest density grade, in particular[18]. The unevenness gave **ga_ors** some difficulty because

- With 5 random training and test groups, some of the samples did not expose the program to examples in all four density grades.
- The program would simply try to classify the second and third density grade correctly and allow the classification of cases in the extreme grades to be incorrect. During the training phase the genetic algorithm and LDA tried to maximise the number of correctly classified cases and even with the errors from the

[17] On the order of 50 test cases.

[18] The highest density grade had a comparable number of cases to the middle density grades.

mammograms falling in the extreme density grades, the system would classify the vast majority of the cases correctly. The genetic algorithm then "decided" that this solution was nearly optimal and mainly explored solutions with similar behaviour. Then, during the testing phase, many cases were still classified correctly but there were more errors due to the patient variability. This combined with all the errors from the extreme density grades gave a low overall performance.

Of the properties that were considered for this study, the regional moments and the multifractal dimensions classified the density grades the best and both properties had similar analogues in [Byng et al., 1996a] and [Byng et al., 1994] who found a correlation with the mammographic density. Due to these factors, it seemed likely that the poor performance was not necessarily from the choice of the extracted property or the fault of **ga_ors**, and the analysis should be repeated with a larger sample.

The remaining investigations which were performed involved the diagnosis of the images or patient cases. When considering the mammograms with a diagnosed malignancy and one mammogram for each normal case, **Diag** cohort, the properties: global and regional moments, sub-regions of the histogram and the multifractal dimensions all had similar overall classification performance at roughly 60–70%. The texture properties had an even better classification performance at ~ 80–85%. Similarly, when the images under consideration were changed to the mammograms contralateral to the breasts where a malignancy was diagnosed and those normal mammograms which were not selected by the random selection of training and test sets (i.e. the **Contra** set) the performance is largely comparable for all properties between the results in the **Contra** and **Diag** cohorts. However the performance of the texture energy, in particular, was seriously degraded in comparison to the **Diag** cohort. Finally, the results did not exhibit any dependence with age for a reasonable choice of the critical value, $p = 0.001$, given the characteristics of the dataset itself[19].

Overall, the texture inertia appeared to have the best combination of characteristics. The classification performance was consistently high for the **Diag** and the set of images in Section 4.2 (the **Contra** cohort) as well as having a relatively small variance in the classification accuracy after random redistribution of the cases into training and test groups. Additionally, there did not appear to be any age dependence in the results.

[19] The sample size was small and there was an overlap of cases between adjacent age groups.

Chapter 5

Discussion

The overall objective of the project was to identify global characteristics of mammograms that may be useful in assisting in the diagnosis of breast cancer or assessing breast cancer risk. In order to assess risk, an established mammographic risk factor was used, the mammographic density grade while the diagnostic ability of the computer system was compared to the known clinical diagnosis of the mammogram.

The images that were used for the study were obtained from a publicly available database, the digital database for screening mammography, from the University of South Florida. The database represents a first step toward a standard database of images to be used for mammographic image research. The full database is to contain a large number of images when complete and the images were obtained with high quality x-ray digitisers intended for use in mammography. This factor is quite important as the breast shadow must be segmented from the background and the task was considerably more straightforward when the images were obtained from these digitisers.

The database was not an ideal dataset for a study on breast cancer risk similar to those done by Boyd [Boyd et al., 1995] and Byng [Byng et al., 1997]. The details of each patient's reproductive history (age of menarche, age of first live birth, etc.) must be known and was not provided as part of the DDSM. As well, there were some additional deficiencies that may hinder some types of studies. For example, the database was lacking images that were obtained from the same patient over a fairly long period of time. This would be useful for examining age dependent effects or temporal changes in mammographic features.

5.1 Procedure Overview

The breast shadow was segmented from the background and corrections made for the scanner dependent effects and mammographic technique as described in Chapter 3. Next, the properties that were selected for this study were extracted. A greater number of features than what was expected to be useful for the mammographic classification was intentionally extracted since the precise combination of features that would maximise the classification accuracy was not known. From this large pool of features a subset was selected that had the best classification ability. The properties that were calculated consisted of the first 10 global moments, the first 10 regional moments, subregions of the global histogram (a maximum of 15 subregions were allowed after reduction of the histogram from 4096 grey levels to 256), multifractal dimensions (20), the texture energy, texture entropy and texture inertia (300/texture).

The energy, entropy and inertia were calculated from the wavelet transform of the image. The transform itself was executed using a biorthogonal wavelet basis as described in Sweldens [Sweldens, 1994]. The textures were found for 5 different resolution levels of the wavelet transform and for 3 of the quadrants in each level

(LLL...LH, LLL...HH, LLL...HL, see Figure 2.3). The textures quantified characteristics for pairs of pixels in the image and required two additional arbitrary parameters representative of the separation and orientation of the pair of pixels under consideration. Twenty different combinations of these arbitrary parameters were used for each texture in each quadrant and in each level for a total of 300 features per texture.

From the pool of features, the single best global moment was selected, the single best regional moment and so on for each property. Then the best pair of global moments, best pair of regional moments, etc. was found. The procedure was repeated for 3, 4, 5, 6, 7, 10 and 15 features, except for those properties where fewer than 15 features were calculated (i.e. global and regional moments). For the global moments, regional moments and the multifractal dimensions, an exhaustive search of all possible combinations of single features, pairs of features, etc. could be performed in a reasonable amount of time. However, the total number of possible combinations was too great for an exhaustive search for the remaining features (subregions of the histogram, the energy, entropy and inertia) and **ga_ors**, a program developed at the Institute for Biodiagnostics, was used as an alternative approach for finding the best set of features. **ga_ors** uses a genetic algorithm to select the optimal or nearly optimal combination of features to maximise the classification accuracy.

The procedure was used to classify the images into categories corresponding to density grades (**Den**) and according to the diagnosis of the mammograms themselves (**Diag** group). Additionally, the age dependence in the classification performance of the selected features for the **Diag** cohort was examined along with any dependence in the results to the remaining systematic scanner dependencies for a small number of mammograms. Finally, the selected features were used to classify the normal mammograms which were contralateral to those selected for the **Diag** cohort (**Contra**).

5.2 Results

The following provides a summary of the main results for each cohort and for the age dependence and scanner dependence. See Chapter 4 for details.

5.2.1 "Den" Cohort

When classifying the images into density grade categories, it was found that all the properties under consideration (global moments, regional moments, subregions of the histogram, multifractal dimensions and the texture energy, texture entropy and texture inertia applied to the wavelet transformed images) had similar classification performance. Approximately 40% were classified correctly, independent of the number of features used. This classification rate was inferior relative to other studies ($\gtrsim$ 60%).

Many of the features which were investigated were variants of the properties used in the literature, however, one property was calculated closely following the approach in [Byng et al., 1996a], the regional skewness. The results for this property on the database of images used in this study showed a similar performance to any of the other properties under consideration. Therefore, it was quite possible that the small sample size and uneven distribution of cases between density grades was a significant factor for these results. As discussed in Section 4.1.1, the influence of these factors was too great to permit a conclusion regarding the usefulness of this approach for classification according to density grades.

5.2.2 "Diag" Cohort

For the mammograms where a malignancy was found and a random selection of mammograms from the patients who were free of cancer (**Diag** cohort), the classification performance varied with the property under consideration.

- The classification performance when using the subregions of the histogram were independent of the number of regions appearing in the discriminant function. Classification of $\sim$ 60% of the cases were correct.

- The multifractal dimensions showed a broad peak in the classification results as the number of dimensions used in the discriminant function changed. The maximum performance was $\sim$ 60% correctly classified cases when using three dimensions.

- As the number of regional moments used in the classifier is increased, the behaviour in the results is quite complex. Initially the classifier had a low classification accuracy ($\sim$ 55%), reached a flat plateau ($\sim$ 60%) and rose again for a high number of moments ($\sim$ 70%). The simplest classifier with good classification lay in the plateau region and used two regional moments.

- The global moments also exhibited a plateau or a very broad peak when the classification results were examined as a function of the number of moments used. Similarly, the best performance ($\sim$ 70%) occurred for two moments.

- The texture entropy had a high classification performance of $\sim$ 80% when approximately five entropy textures were used and the performance was maintained at this level for up to at least 15 entropy textures in the discriminant function. Therefore five entropy textures were used for this property.

- Both the texture energy and texture inertia had a low classification performance for two textures and rose to a plateau as the number of textures were increased. The increase in classification performance was more dramatic for the texture inertia, changing from $\sim$ 65% (for two textures) to $\sim$ 85% for six textures while the texture energy spanned $\sim$ 70% to $\sim$ 80%, using seven textures.

An examination of the actual combination of features that were selected also revealed that the directional information contained in the textures was ignored (isotropic features were more significant than any directional information contained in the textures). Secondly, the textures calculated for the low resolution levels of the wavelet transform were predominately selected. This justified the assumption that subsampling the images to 110 μm/pixel had little impact on the classification ability of the properties. Finally, when the features used in the discriminant function were not constrained to belong to a single property the texture energy, texture entropy and texture inertia were selected to the exclusion of all the others.

5.2.3 "Contra" Cohort

In the **Diag** cohort, there were many mammograms in the dataset that were not included. The majority of these were the mammograms for the clinically normal breast contralateral to those that had a malignancy (for the abnormal cases) or, for the normal cases, the images that remained after five random samples were chosen for the **Diag** cohort. These images were collected as the **Contra** cohort. See Section 3.2. The classification performance for the feature sets selected from the **Diag** cohort were tested on these **Contra** images.

The testing involved:

1. Creating 5 discriminant functions using the best feature set for each property, described in Section 4.2, and changing only the coefficients for the 5 trials (A–E) of the **Diag** group.

2. Taking the same discriminant functions formed for the **Diag** cohort.

All these functions were then tested for their classification performance on the **Contra** cohort.

Overall the results were similar to those described above for the **Diag** cohort. The similarity in the results between the **Diag** and **Contra** cohorts may be due to an early stage of breast cancer that had yet to exhibit any clinical indicators or mirror symmetry between the left and right breasts, since the **Contra** cohort predominantly consisted of the mammograms contralateral to those in the **Diag** cohort. There is a high risk but low overall incidence of contralateral breast cancer [Chen et al., 1996], therefore it was more likely that the similarity in the classification performance between the two cohorts are due to symmetry rather than the presence of an abnormality. However, the fact that the system was trained to distinguish normal/abnormal groups and the high risk of contralateral breast cancer suggested that the selected features may be associated with a mammographic characteristic that is related to risk rather than some definite appearance of disease. While suggestive, this evidence is circumstantial and further work in this area is recommended for an unequivocal conclusion about the nature of the characteristic that is detected.

5.2.4 Age Dependence

The exact form of the dependence of the classifier results on age varied with the features used to form the classifier and with the cutoff selected for statistical significance. The images with a diagnosed malignancy and a random selection of images from patients who were cancer free, **Diag** cohort, was considered for this part of the study. The cohort was divided into 8 groups based on the patients' age; 40–54, 42–56, 44–58, $\cdots$, 54-68. The feature sets which were used to form the classifiers for this study were the discriminant functions with the best classification performance for the fewest number of features for the original **Diag** cohort, as described in Section 5.2.2. The classification results were then tested for a statistically significant linear (with a slope different from 0), quadratic or cubic dependence with age compared to a constant. Additionally, any quadratic or cubic age dependence was also tested for statistical significance when compared to a straight line.

Overall, if the limit for statistical significance was set relatively high, $p = 0.05$, the age dependence in the results was complex. For example, for the regional moments and the texture entropy there appeared to be a cubic age dependence while the texture energy resulted in a quadratic age dependence and the remaining properties gave results that were independent of age. However, these may be artifacts from the small sample size and the use of the same images in several age groups. When p was set to a lower threshold, 0.001, there was no apparent age dependence in the results.

5.2.5 Scanner Dependence

There was a small number of cases (34) with diagnosed malignancy but with mammograms digitised on the DBA scanner, which was the scanner usually used for scanning the normal cases. If the classification accuracy of the **Diag** cohort was used for the probability of a correct mammogram classification for each feature set then the probability of observing the distribution found in Section 4.4 for the 34 cases, from random chance alone, can be calculated using a binomial distribution. From Section 4.4 the resulting probabilities were low, which implies that there may be a residual scanner dependence in the data. Again, this may be a small sample size effect since the classification accuracy found for the **Diag** cohort was an average value for a sample consisting of more than 34 cases. If this was indeed the case, then the results which were observed for the 34 cases would be unlikely but not necessarily systematic.

The corrections that were applied to the images compensated for variations in the slope of the linear part of the calibration curve, resolution and non-linear effects in the calibration curve. These alterations made the resolution and contrast consistent with a single scanner. The final modification to the grey levels (the removal of the tails of the histogram followed by extending the histogram to occupy the full range of possible grey level values) provided a small correction for variations in the mammographic technique. No attempts were made to correct for noise differences aside from the inherent smoothing due to the reduction in the resolution. Further corrections for the MTF, noise and differences in the details of the mammographic technique are difficult to perform both theoretically and in practice.

If the corrections that were applied to the images are insufficient to reduce the scanner characteristics below a detectable level then the repercussions are significant. In particular, meaningful comparisons of results from different groups would not be possible unless the same database of images were used for all studies. In addition, the maintenance of a reliable system for computer aided diagnosis is made more difficult since the system may need to be retrained after any alterations to the hardware. This includes routine adjustments to a laser scanner such as a re-calibration of the look up table for conversion of the detected light signal (which is related to the optical density) to a pixel value.

5.3 Summary

In summary, none of the properties that were selected appear to be useful for the classification of density grades for this sample. However, it is advisable for an additional study be performed with a more extensive data set designed specifically for automatic density grade classification before dismissing the properties in this thesis for density grade classification.

For the normal/abnormal classification, the texture inertia using 6 features gave consistently high classification performance for the mammograms with a diagnosed malignancy and a random sample from the normal cases (**Diag**). In addition, none of the selected features appeared to exhibit any significant age dependence in the results, at the $p = 0.001$ level. The results for all properties were comparable for images of the normal breast contralateral to those with the diagnosed malignancy and the mammograms from the normal cases not previously selected (**Contra** groups). However, the classification accuracy using the texture energy was significantly lower for the **Contra** cohort relative to the **Diag** cohort results. Overall, the texture inertia appeared to exhibit the best combination of qualities for the classification of the mammograms into normal/abnormal groups.

5.4 Future Work

The work presented in this thesis signifies the beginning of a broad examination of global properties for mammographic classification. Some important aspects for the classification of mammograms with global characteristics have been revealed by this study, many possible extensions identified and details in the execution revealed. The more important issues are discussed below.

5.4.1 Segmentation

The segmentation procedure that was used for this work was a semi-automatic procedure that requires a considerable amount of user interaction. This was primarily due to the need for the procedure to be flexible and to have enough features in order to cope with any image that may be encountered. These characteristics were particularly significant in segmenting the images from the Vision Ten scanner which possessed a considerable number of severe artifacts that made the segmentation procedure difficult. With the higher quality images from the DDSM database, it would be straightforward to make the procedure fully automatic as long as the poorer quality images could be identified and set aside. Since the user interaction constrains the number of images that can be segmented in a given time, a fully automatic segmentation procedure can reduce the time needed to process a large number of images.

5.4.2 Scanner Dependence

As stated previously, it is recommended that a study be made to specifically investigate the effects of the scanner used to digitise the mammograms on the classification performance. However, an ideal investigation would examine the effects of the exposure conditions, processing of the film, etc. in addition to the scanner dependence. Therefore, the desired radiographs would consist of x-rays for an anthropomorphic phantom taken

1. At different kVp's to examine the impact of the AEC as well as the effect of the kVp setting on the images.

2. For different film/screen/processor combinations.

3. For various intervals from the most recent QA test. This examines the effectiveness of the QA procedures to maintain a consistent image quality with respect to the mammographic properties that were extracted.

4. Digitised using different scanners.

It is desirable to have many images to encompass the full range of variation that can be expected for each variable given above. A large number of images is also needed since this type of classifier is of a statistical nature rather than the more typical CAD systems that identify a suspicious area within any *individual* mammogram. The image set can be analysed for scanner dependence using the same procedure given in the previous chapters as well as examining the features individually for uniformity between images obtained from different scanners. It would also be possible to use these images to investigate the details of how the classification system distinguishes between the various categories. For example, the selected features may be related to an incidental rather than a causal effect such as the kVp used to obtain the mammogram. A dense breast has a greater risk of developing breast cancer and generally requires a higher kVp so that a classifier that was in some way sensitive to the kVp would also classify many mammograms correctly.

5.4.3 Density Grade Classification

Attempting to improve the density grade classification performance may be achieved by using a much larger data set with considerably more patient information than was available for this thesis. The database should be large enough to enable the selection of a balanced number of cases in each density grade. This factor is the primary difficulty in carrying out a more detailed analysis than what was described in previous chapters.

On the other hand a study of the risk requires a more sophisticated study altogether, such as a case control study as found in [Byng et al., 1997]. The current database does not provide the necessary patient information to be able to match the cases and controls with respect to similar risk factors that are beyond our control. In particular, the reproductive history is missing (age of first live birth, age of menarche, etc.). What is required is an entirely different database for an evaluation of the relative risk associated with the properties considered in this thesis. Once an appropriate database is assembled, the full procedure given previously should be repeated.

5.4.4 Diagnosis

It is possible that the selected feature set was characterising considerably different aspects of the mammogram; some that are indicative of cancer risk and others that are indicative of the clinical appearance of an abnormality in the current mammogram. Further, it is not possible to isolate the category to which each characteristic belongs without data selected specifically for that purpose. The simplest approach to study the diagnostic ability of any particular features would be to employ a database of images that consists of many sets of mammograms for the same patient over a long time interval. It is also useful to have roughly equal numbers in both the normal and abnormal groups and both groups should utilise mammograms obtained over approximately the same time period to compensate for improvements in the film technology. For the abnormal cases, the screening mammograms both before and after the actual diagnosis of the lesion would be useful. The images taken prior to discovery of the malignancy have to extend over a sufficiently long time that the presence of a lesion which is masked in the early mammograms is unlikely. Similarly, the normal cases require that the patient has been undergoing screening mammography for a sufficiently long time to be confident that the patient has not contracted an abnormality for some time after the date of acquisition for the last mammogram used in the study.

5.4.5 Age Dependence

The approach to investigate the age dependence given in the previous chapters is sufficient to examine this characteristic. However, a much larger database of images is required so that the results are less likely to exhibit small sample effects. Indeed, if the sample size can be made large enough, it is possible to have a statistically significant number of samples for each age group as well as disjoint groups with respect to both the cases in each group and the ages of the patients in each group.

5.4.6 Bootstrapping

Both the current study, described in the earlier chapters, and the proposed experiments given above require a training and a test set as an integral part of the analysis. The results presented for this thesis used only five training and test groups that were formed by redistributing the available cases between the two groups for each trial. This is a small number of trials and a better estimate of both the expected performance for the feature set under evaluation and the uncertainty in the performance can be obtained for a larger number of trials. While it is possible to simply increase the number of trials and to find the average and standard deviation for the results as in [Boone et al., 1998], there exists a method of combining the results for a large number of trials that gives a better estimate of the expected performance and uncertainty. This method is known as bootstrapping [McLachlan, 1992]. A prototype program has been developed at the Institute for Biodiagnostics that performs the procedure but it is in its early stages and requires the intervention of an expert statistician to perform the procedure and evaluate the results. The procedure has not been used for this current study but there is no fundamental reason that prevents its application to the data for this work once the program reaches production quality software.

5.4.7 Miscellaneous

One point that has become obvious in the literature is that there are many approaches to feature extraction in mammogram classification but few comparisons between competing techniques. One of the difficulties for such a comparison is the difference in image databases between different studies as well as differences in the reporting methods, some use the percent correctly classified (as do we), some the area beneath ROC curves, some the actual relative risk between patients with the highest and lowest values for some particular property. Others utilise various statistical parameters (several different types of correlation coefficients, etc.).

Once the necessary image database is constructed it would be a simple matter to make a direct comparison of the performance for all the features used to date for mammogram classification. This would require a careful examination of the literature since the selection of the images for the database may have many constraints to be able to reproduce the work of several different groups *exactly*. Additionally, some of the diagnosis and risk studies require very large datasets and some of the studies involving the use of Wolfe grades may be difficult to execute since it requires the assessment of the mammogram by someone experienced in classifying the mammograms into these categories. However, the popularity of Wolfe grades has declined and some of the techniques appearing in the literature can be used for a more objective measure for various mammographic classes and enable a comparison to be made without as much intervention of an experienced observer.

5.5 Final Comments

Clearly, there is a sufficient number of extensions to the work presented in this thesis that a full investigation of all aspects would extend over many years. The studies that were performed with the current set of images constitute an initial examination of a potentially ongoing investigation for mammographic classification using the selected global properties. These current results have successfully identified simple texture properties of a wavelet transformed mammogram that are useful for mammographic classification.

Appendix A

Fractal Geometry

One of the principal features that was extracted from the mammograms used in this study was based on a multifractal dimension. Since a multifractal is an extension to the basic ideas underling fractal geometry, a brief outline of some of the concepts is described below prior to discussing multifractals.

A.1 Basics

Although many of the initial steps needed for the field of fractal geometry appear at various times throughout history, one of the key people behind unifying the concepts and forming fractal geometry into its own field of study has been Benoit Mandelbrot. He recognised that the use of conventional Euclidian objects, such as lines, circles, curves, etc., was inadequate when trying to model many naturally occurring objects — the objects themselves were simply too complex. However, there did appear to be a unifying characteristic behind these complicated objects; namely, self-similarity. Symmetry has long been an important characteristic in the study of natural phenomena.

For example, Noether found that the symmetries that are a characteristic of the system itself imply the existence of some conserved quantity [Goldstein, 1980]. As well, the behaviour of systems under rescaling has been an important tool in statistical mechanics and condensed matter physics for some time [Chaikin and Lubensky, 1995]. The behaviour of an object under changes in scale is fundamental to fractal geometry as well and, in a loose sense, fractals are characterised by invariance to changes in scale [Addison, 1997].

Consider three of the "classic" mathematical or pure fractals, the Koch curve (Figure A.1), the Sierpinski Gasket (Figure A.2) and a Peano curve (Figure A.3). All of these were formed using a simple systematic iterative procedure. For example, a Koch curve was formed starting from a straight line and applying the procedure

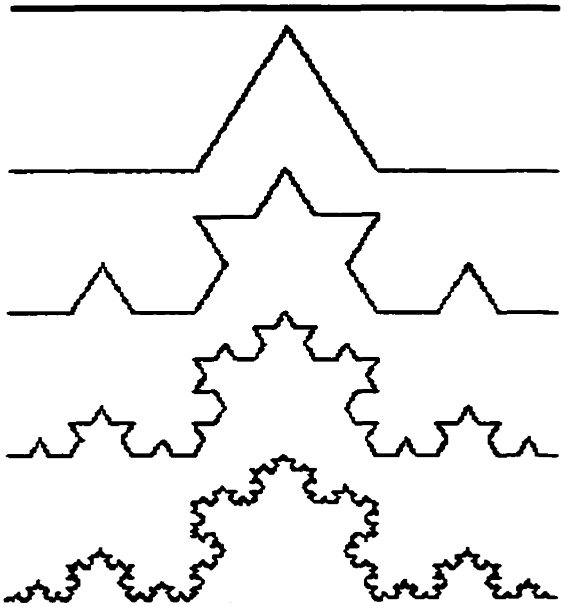


Figure A.1: Five iterations in the generation of a Koch curve starting from a straight line. Figure taken from [Peitgen et al., 1992].

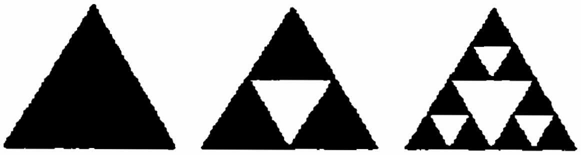


Figure A.2: Three iterations in the generation of a Sierpinski gasket. Figure taken from [Peitgen et al., 1992].

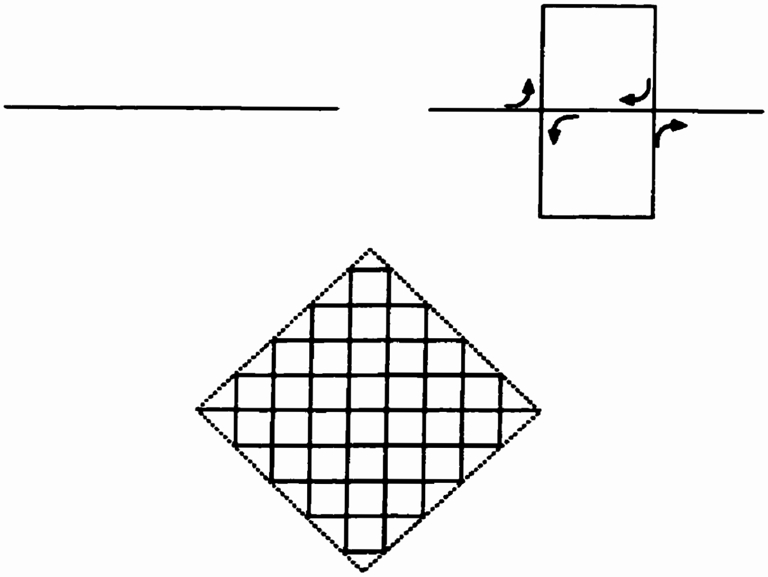


Figure A.3: Three iterations in the generation of a Peano curve. Figure taken from [Peitgen et al., 1992].

1. Remove the middle third of the line segment.

2. Place two line segments of the same length as the part that was removed in the gap, so that it forms an equilateral triangle (with one side missing).

The process was carried out for each line segment that forms a part of the curve and repeated *ad infinitum*. Similarly, a Sierpinski gasket starts as a filled triangle and $\frac{1}{4}$ of the area is removed from each remaining triangle for each stage in the iterative process. Finally, the Peano curve employed a scaling factor of $\frac{1}{3}$ at each stage and replaced each line segment by a combination of 9 line segments that formed the shape shown in the second iteration in Figure A.3.

The resulting objects have properties quite unlike other commonly encountered curves or surfaces. Both the Koch and Peano curves are continuous everywhere and differentiable nowhere and all three of these objects are strictly self-similar. If a region of the object, especially in its final state, is rescaled by the proper amount, the result exactly resembles the original, although some rotation may be necessary. In addition, the Peano curve can be shown to fill an entire region of space [Peitgen et al., 1992] and space filling curves are examples of where the conventional concept of dimension encounters difficulties. The Peano curve was formed from a single curve, a one dimensional object, but covers all points on a surface, a two dimensional object so the question arises — which dimension should be used for a Peano curve? Mandelbroit argued that these sort of objects were members of an entirely different class of objects and that the concepts which were used to define the dimensions of more conventional objects were inadequate for these. The conventional dimensions (1 because it is a line and 2 since it fills the area) still appear as the topological dimension and the Euclidean dimension, respectively. A fractal dimension, on the other hand, is formulated using some aspect of the scale invariance that is inherent in a fractal. This is usually expressed by some variation of

$$M = \kappa s^d \tag{A.1}$$

where M is some measured quantity, κ an arbitrary proportionality constant (it is not relevant for the fractal dimensions), s some distance characteristic of the resolution or scaling and d is the quantity related to the fractal dimension. Depending on the exact quantities for M and s, d may be the fractal dimension directly or may need to be offset, usually by either the topological dimension or the Euclidean dimension, before being used as a fractal dimension.

One of the more common methods of calculating a fractal dimension is known as the Hausdorff mesh or box counting dimension. The procedure to calculate this dimension is very straightforward and this simplicity combined with the ease in implementing the calculation automatically has made the box counting dimension extremely popular. Basically, the approach is as follows:

1. If the box counting dimension is desired for the object in Figure A.4, choose an initial length, ε. This length is arbitrary and for an image a convenient size is 1 pixel.

2. Superimpose a regular mesh or grid composed of cells of size ε x ε over the object.

3. Count the number of cells that contain *any* portion of the object.

4. Change the size of the mesh and repeat the procedure. Since the scaling behaviour of, in this case, the number of cells covering the object is under exami-

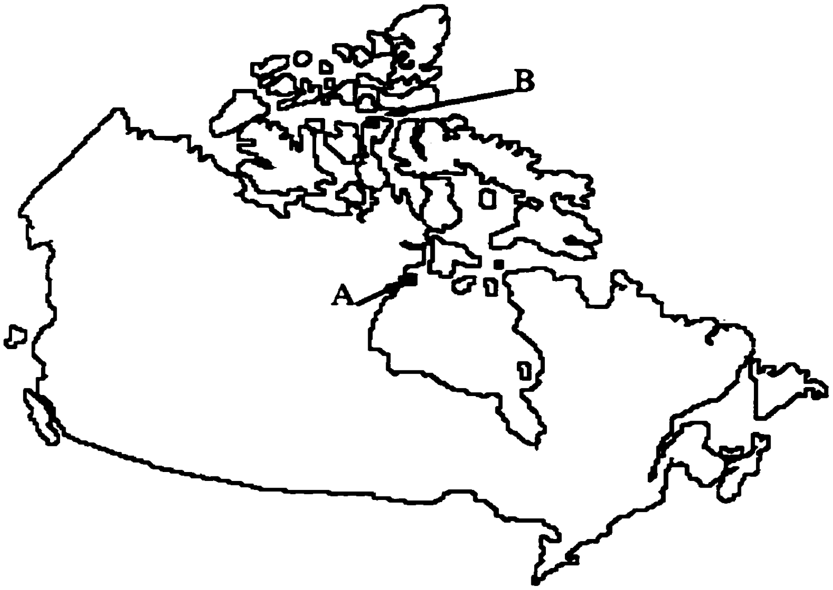


Figure A.4: Border of Canada. The labels "A" and "B" are used as part of a fractal dimension calculation (following).

nation, a wide range of resolutions is needed. Therefore, a dramatic change in the resolution is chosen for each iteration of the procedure. If the initial size is 1 pixel, and for a digitised image this is the smallest resolution possible, a common choice for the other resolutions form a dyadic sequence, i.e. 1, 2, 4, 8, 16, $\cdots$

5. Plot the logarithm of the number of cells as a function of the logarithm of the reciprocal of the cell size and look for a linear relationship[1] If one does not exist then the fractal model does not hold and the calculation of a fractal dimension is nonsensical. On the other hand, if there *is* a linear relationship the slope of the fit is the fractal dimension (for the box counting dimension). For real

[1] There is some ambiguity in the quantity that should appear on the abscissa. For the scale sizes defined as we have in this example, we use the logarithm of the *reciprocal* of the scale size.

Scale Size (Pixels)	Number of Cells to Cover
1	4175
2	2604
4	1007
8	378
16	132
32	51
64	18

Table A.1: Data for the box counting dimension applied to Figure A.4

objects there will only be a finite range of resolutions where the fractal model is applicable and usually not all the data that were calculated can be used for the regression fit.

When this procedure is applied to Figure A.4, the resulting data for the scaling properties are given in Table A.1. The data are also plotted in Figure A.5. From Figure A.5 it was clear that the linear relationship was very good except for the point that corresponds to a scale of 1 pixel. This indicates that for the finest resolution the fractal model of the border was beginning to break down but for scales from 2 to 64 pixels the model was a good representation for the border. The slope of the regression fit and the box counting fractal dimension was then 1.44 and is, as expected, greater than the topological dimension (1) and less than the Euclidean dimension (2).

There are many approaches to calculating a fractal dimension. Indeed, there are almost as many different methods as there are studies that use a fractal dimension. Only two are discussed here, the mass dimension and a fractal dimension derived from the power spectrum. For the purposes of this discussion, the mass dimension was used to demonstrate a few general properties of fractal dimensions that we would like to emphasise while the power spectrum dimension is described since it and vari-

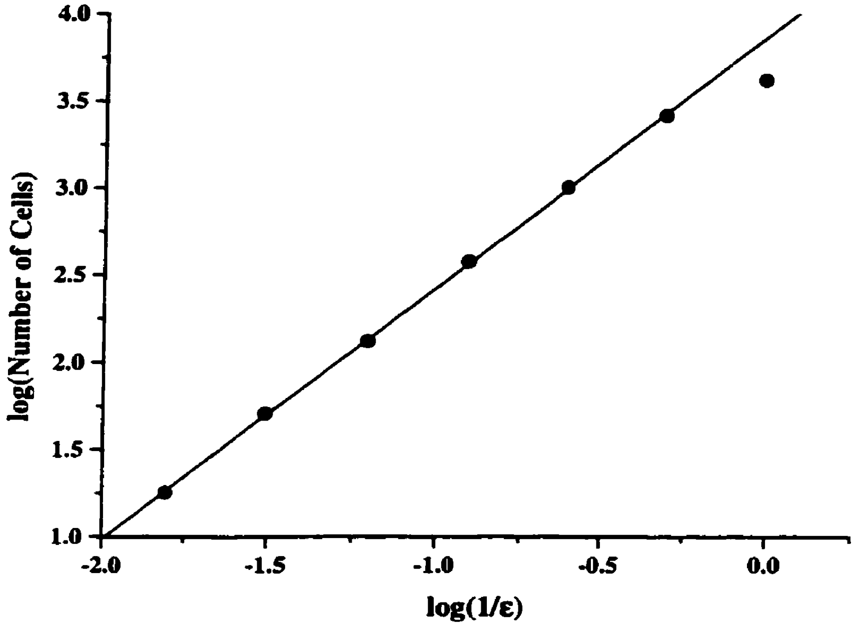
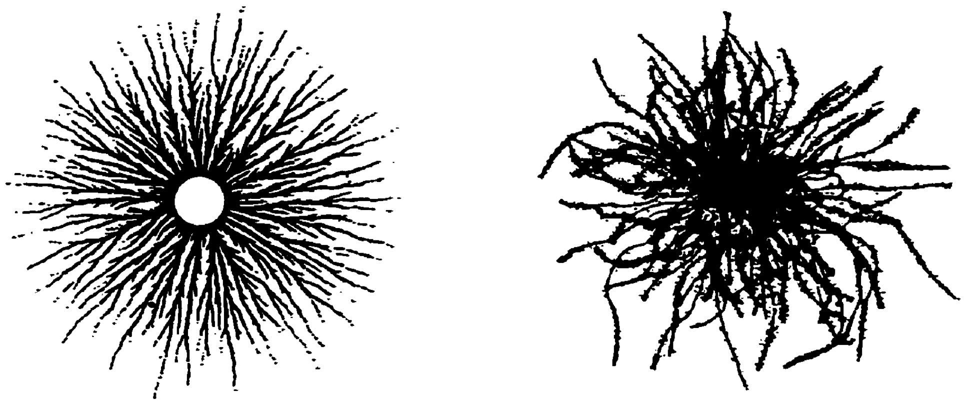


Figure A.5: Plot of data for application of the box counting dimension on Figure A.4.

ations on the technique are common in the literature for mammographic texture characterization.

The procedure necessary to find the mass dimension [Schroeder, 1991] is also straightforward. It is commonly used to calculate a fractal dimension for objects that appear to radiate from a central location such as in a Lichtenberg figure, Figure A.6(a) — formed by the electric discharge from an electrode placed at the centre of the image, or natural down, Figure A.6(b). The mass dimension is defined as



(a) Lichtenberg figure

(b) Natural down

Figure A.6: Two examples of natural fractals.

$$m \propto r^{D_m} \qquad \text{(A.2)}$$

where m is the amount or "mass" of the object contained within a radius r of some point, naturally the centre, and D_m the mass dimension. Therefore, the actual procedure is roughly similar to the procedure for the box counting dimension. For an *image* and an arbitrary selection of radii[2] find the "mass" (the number of pixels) con-

	Mass	Mass
Radius (Pixels)	(Centre : Hudson's Bay A in Figure A.4)	(Centre : Nunavut B in Figure A.4)
1	0	0
2	2	1
4	9	11
8	24	50
16	72	208
32	255	825
64	1137	1926
128	3222	2792
256	4175	4175

Table A.2: Data for the mass dimension applied to Figure A.4 with the centre for the technique placed at two different locations, Points A and B in Figure A.4.

tained within a circle of each radius. Then, if there is evidence of a linear relationship between $\log(m)$ and $\log(r)$ the slope is D_m. If this procedure were to be applied to the image of the border of Canada, Figure A.4, there is some arbitrariness as to the choice of location for the centre. In particular, consider two centres, one on the coast of Hudson's bay (point A in Figure A.4) and the other near the middle of Nunavut (point B in Figure A.4) for radii of 1, 2, 4, 8, 16, 32, 64, 128 and 256 pixels. The larger radii actually extend beyond the edge of the image but since there was no part of the Canadian border that extended beyond the image, all points outside the image were taken as white. The data for the mass dimension for these two centres are shown in Table. A.2 and Figure A.7 Again, there are some important observations from Figure A.7.

1. The range of scales over which the fractal model was applicable was finite. When the centre for the mass dimension was placed near the centre of Nunavut (point

[2] This is also usually chosen as a dyadic series.

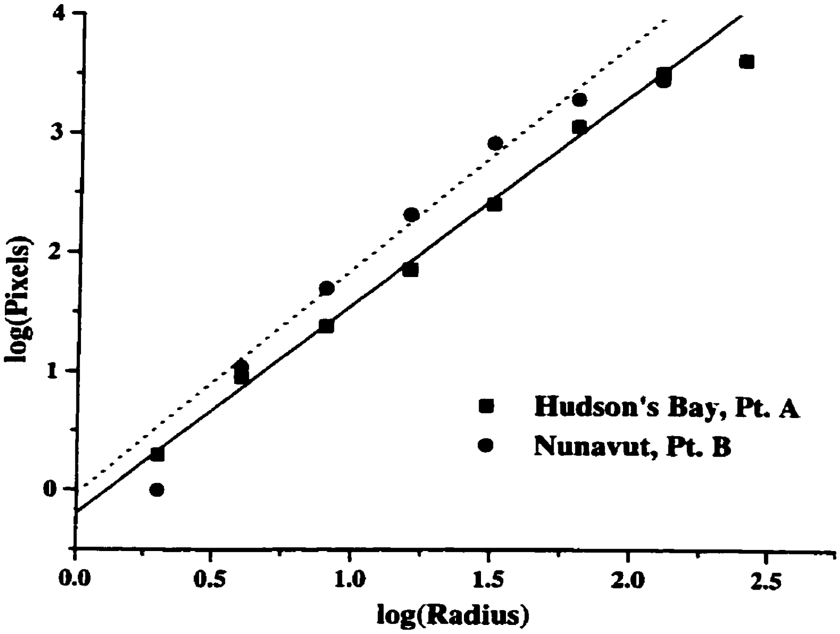


Figure A.7: Plot of data for application of the mass dimension on Figure A.4 with the centre for the technique placed at two different locations: points A and B from Figure A.4.

B) there was only a narrow range of resolutions where the curve was linear[3].

2. The slopes, and therefore the mass dimension, were *not* the same when the mass dimension was calculated at two different locations. This demonstrated that a fractal dimension may not be uniform over an entire real world object, unlike a mathematical fractal such as the examples in Figure A.1 – Figure A.3.

3. The mass dimension with a centre at Nunavut was 1.9 while the mass dimension with a centre near Hudson's bay was 1.8. Both are different from the box counting dimension (1.4) for the same object.

The repercussions of point 1 were discussed above. An arbitrary selection of resolutions was not appropriate when calculating a fractal dimension on a real object. Point 2 was also fairly obvious. A physical object may not have the same fractal dimension throughout. However, point 3 is less well known and there are several reasons behind the observed difference:

- Not all methods of calculating the fractal dimension are appropriate for all situations. The mass dimension was intended to be used on object that seem to have an obvious central point.

- There is *not* a single fractal dimension. Each method of finding a fractal dimension actually examined a different property of the object under investigation. See [Schroeder, 1991] or [Peitgen et al., 1992], for example.

Both points 2 and 3 imply that the specification of the method used in the calculation of any fractal dimension must be clear and unambiguous.

[3] The breakdown in the fractal model close to the largest scale was expected. The radius of the circle extended beyond the limits of the image and, obviously, the border cannot be fractal beyond its limits.

A.2 Random Fractals

Most naturally occurring fractals, including coastlines, are not strictly self similar. Unlike a Koch curve, a small region cannot be made to resemble the entire object exactly. Rather, statistical self similarity is the norm, since the objects are formed by processes with at least some aspect of randomness. For example, consider a mountain range. The distribution of softer and harder components of the rock combined with the, generally, non-laminar flow of water over its surface leads to a random looking surface. However the randomness is not uniformly distributed. Once a shallow path is formed for the water to drain off, more water tends to follow that path in preference to others. This type of behaviour has resulted in fractal dimensions derived from power spectra.

If the power spectrum[4] was calculated for a property with a uniformly distributed random characteristic, there would be equal contributions to all frequencies and the spectrum would be flat. However, a random fractal such as a mountain range or the path of a particle undergoing Brownian motion[5] has a power spectrum

$$P(f) \propto f^{-(2D_T+3-2D)} \tag{A.3}$$

where P is the power, f the frequency, D_T the topological dimension and D a fractal dimension. The inertia of the particle resists very rapid changes in direction which suppresses high frequency components and the power decreases as the frequency increases. Hence, the observed fractal dimension for traditional Brownian motion is

[4] The magnitude of the Fourier transform for some characteristic.

[5] Brownian motion is the characteristic behaviour of fine particles under the collective effect of random collisions with the molecules of the surrounding medium.

$D = \frac{3}{2}$ [Schroeder, 1991] and is considerably different from a constant power spectrum ($D = \frac{5}{2}$). For most random fractals, including those that are created from components that undergo random walks, such as Brownian motion, D is not constrained to any particular value and it may even attain integral values. The fractal dimension is characterised by the approach used to find the dimension and is not limited to only non-integer values. These fractal dimensions, derived from a power spectrum, are common in many medical applications as well as in chaotic systems where the behaviour of a system appears to be partly random but there also appear to be some systematic trends. These fractal dimensions along with topics related to the behaviour of the power spectrum can be found in the literature under "coloured noise".

A.3 Summary

The discussion throughout this entire section on fractals and fractal dimensions has been of an elementary nature. A more comprehensive overall discussion can be found in Schroeder [Schroeder, 1991], Peitgen *et al.* [Peitgen et al., 1992] and Addison [Addison, 1997] while a discussion on power spectrum dimensions with particular attention to their application to natural objects appears in Petland [Petland, 1984]. Finally, a description of many fractal dimensions common in medical applications may be found in Veenland *et al.* [Veenland et al., 1996].

A.4 Multifractals

The traditional fractal objects, such as a Sierpinski gasket, a Koch curve or a Peano curve (Figure A.2–A.3), which appear in many references, [Peitgen et al., 1992] or [Schroeder, 1991], are examples of strictly self similar objects. That is, if a portion of the object is re-scaled by the appropriate amount, only a rigid body transformation is necessary to make it exactly match the original. However, many physical objects with fractal-like behaviour are created by processes with a certain amount of randomness and are statistically rather than strictly self-similar. Common examples of random fractals are coastlines, mountain ranges and Brownian motion. For these objects it is not possible, in general, to make any sub-region exactly correspond to the original but the general character of the sub-regions does resemble the full object. Indeed, if an image of the sub-region was viewed without reference to the original it is difficult to judge whether it is a sub-region or the full object.

Multifractals are generally random fractals and can be thought of as consisting of many random fractals, with possibly different dimensions, which are intricately intertwined. Then, when different approaches to calculate the fractal dimension are applied, a different dimension may result depending on the "fractal component" to which the method is most sensitive. Because of this, when the fractal dimension is calculated for any natural object the method of calculation for the fractal dimension is critically important.

In order to further elucidate the difference between the conventional fractal dimension and multi-fractal dimensions, consider the situation of several fields with different types of ore scattered over its surface by some natural process[6]. The fields

[6] This same example was also discussed in Chapter 2.2.1

are approximately the same size but of vastly different composition and value. Suppose we wish to buy one of the fields, the most valuable, but it is too difficult to estimate the total value of the ore over the entire area of each field. In that case we may be interested in the distribution of ore over a relatively small sample of each field (and assume it is typical for the entire region). It is likely that the distribution has a fractal character and one approach which is often used to evaluate the dimension is to use what is frequently called the box counting dimension (or Hausdorff mesh). In this approach a regular grid with a side length of ε is superimposed over the field and the number of cells, N_ε, which contain any type of ore are counted. The process is then repeated with many different sized meshes. The value of the fractal dimension, d, is then related to the slope of the regression fit of $\log N_\varepsilon$ as a function of $\log \varepsilon$. A dimension closer to two indicates a greater amount of ore but this process ignores the type of ore in each cell. Further, if the net value of a collection of ore is desired, the composition of the samples in each cell is very important.

If we now consider the distribution as a multi-fractal, the process of calculating the dimensions starts with the same regular grid but we assign a weight to each cell, μ_{ij}, where ij specifies a location within the mesh. In this case, the total value of the ore in the cell may be used for this purpose. The distribution can then be characterised by the set of fractal dimensions for the various collection of cells with the same μ_{ij}. The difficulty with this is that an integral part of the calculation of the fractal dimension requires changing the mesh size and the value of μ_{ij} will change as the size of the cells change. The logical remedy would be to scale μ_{ij} by the mesh size, ε^2 in this case. Unfortunately, for a fractal or multifractal distribution where we expect μ_{ij} to scale as $\varepsilon^{\alpha_{ij}}, \alpha_{ij} \in \mathbf{R}$, then $\lim_{\varepsilon \to 0} \frac{\mu_{ij}}{\varepsilon^2} = \lim_{\varepsilon \to 0} \varepsilon^{\alpha_{ij}-2}$ is not finite if $\alpha_{ij} < 2$. This can be remedied by using α_{ij} directly rather than μ_{ij}/ε^2 and α_{ij} is often called the coarse Hölder exponent so that

$$\alpha_{ij} = \frac{\log \mu_{ij}}{\log \varepsilon} \tag{A.4}$$

$$\mu_{ij} = \varepsilon^{\alpha_{ij}} \tag{A.5}$$

The frequency distribution of α_{ij}, given by f_ε is defined by

$$f_\varepsilon = -\frac{\log N_\varepsilon(\alpha_{ij})}{\log \varepsilon} \tag{A.6}$$

$$N_\varepsilon(\alpha_{ij}) = \varepsilon^{-f(\alpha_{ij})} \tag{A.7}$$

where $N_\varepsilon(\alpha_{ij})$ is the number of cells[7] that are needed to cover the regions of the multifractal with coarse Hölder exponent, α_{ij}, at a resolution (mesh size) characterised by ε. From this point there are different approaches which can be used. In this work, the technique known as the method of moments was used. What follows is a brief overview of the approach. A detailed description of the method can also be found in [Peitgen et al., 1992].

The name "method of moments" comes from the use of a partition function, $\chi_q(\varepsilon)$, for the q^{th} moment where

$$\chi_q(\varepsilon) = \sum_{i,j}^{N(\varepsilon)} \mu_{ij}^q \,, \quad q \in \mathbf{R} \tag{A.8}$$

Since $N_\varepsilon(\alpha)d\alpha$ represents the number of cells of the total $N(\varepsilon)$ which have $\alpha \in$

[7] N_ε is still the total number of cells of size ε that cover the entire multifractal; i.e. for *all* values of α_{ij}.

$(\alpha, \alpha + d\alpha)$. We can then convert the sum over $N(\varepsilon)$ to an integral over α to give

$$\chi_q(\varepsilon) = \int N_\varepsilon(\alpha)\mu^q(\alpha)d\alpha \qquad \text{(A.9)}$$

χ_q effectively takes the place of the number of cells which cover the object when calculating the box counting dimension and since a multifractal dimension was desired, each cell is weighted by $\mu_{ij}(\alpha)$ — the probability of finding the object in a cell with Hölder exponent of α. Then by using Equations (A.5), (A.7) and (A.9) we obtain

$$\begin{aligned} \chi_q(\varepsilon) &= \int \varepsilon^{-f(\alpha)}\varepsilon^{\alpha q}d\alpha && \text{(A.10)} \\ &= \int \varepsilon^{[\alpha q - f(\alpha)]}d\alpha && \text{(A.11)} \end{aligned}$$

Define $\tau \equiv \alpha q - f(\alpha)$ and the generalised fractal dimension, D_q from

$$\tau(q) = (q-1)D_q \qquad \text{(A.12)}$$

Since the partition function is analogous to the number of cells needed to cover the object, for a fractal object, χ_q scales with the characteristic length as

$$\chi_q(\varepsilon) \propto \varepsilon^\tau \qquad \text{(A.13)}$$

The D_q is known as the generalised fractal dimension since specific values of q correspond to more commonly known dimensions, for example $q = 0$ gives the usual Hausdorff dimension while $q \to 1$ corresponds to the information dimension [Peitgen et al., 1992, Schroeder, 1991].

Appendix B

Basic Principles of Genetic Algorithms

In many applications there is a need to reduce the number of variables needed for some function important to the application. For example, for this thesis we would like the minimum number of features that can reliably be used to distinguish one class of mammograms (abnormal) from another (normal). The traditional approaches, such as an exhaustive search, are impractical as the number of potential features to be considered increases. Other conventional techniques such as stepwise refinement and steepest descent can easily get caught in local extrema, particularly as the dimensionality or the number of potential features increases[1]. An alternative technique that tries to circumvent these difficulties is the genetic algorithm (GA). The technique is a general approach to optimization and is not constrained to any individual field of study. Additional details may be found in the "classic" works in the field, such as

[1] A classification problem can be viewed as clustering in an abstract space where each potential feature is used as a separate dimension. Then the goal is to select the orientation of the viewpoint so that the projection of the clusters are sufficiently separated to some desired degree.

Holland [Holland, 1975].

In this method a possible solution consists of a subset of features that is encoded onto a "chromosome" as described below. A population of chromosomes is generated at random and evaluated for their ability to correctly classify the images. The chromosomes in the population that give the best results are used to form the next generation and as the process progresses a good, but not necessarily best solution, is found. In principle, a global minimum or maximum will be found although it may require an indeterminate amount of time.

The basic ideas behind a genetic algorithm are easily described although many variants have been introduced to accommodate special aspects of different problems. The basic requirements fall into a small number of categories:

1. A mechanism is needed to map the variables under consideration onto genes in the chromosome. A description of the conventional approach, using the position in a bit string for each variable, can be found in Holland [Holland, 1975] or in Prakash and Narasimba Murty [Prakash and Narasimha Murty, 1995].

   As an example, this thesis used a genetic algorithm to choose regions in a histogram that can be used to classify mammograms into several groups. The histogram consists of the number of pixels in the mammogram for each possible grey level. The chromosome was represented by a bit string and each bit position in the string corresponds to a collection of grey levels. Specifically, the i^{th} bit corresponds to the region containing the i^{th} to $(i+1)^{\text{th}}$ grey levels. Then, to select the i^{th} to $(i+1)^{\text{th}}$ region a 1 is placed in the i^{th} position in the chromosome and to indicate that the same interval was not selected, a 0 is placed in the i^{th} bit in the chromosome.

2. To be able to change the population and explore the feature space, a mechanism must be provided to change the chromosomes. The most typical techniques are crossover and mutation:

   Crossover As in the biological form of crossover, this genetic operator exchanges information between two chromosomes and as in natural selection, the "fittest" chromosomes reproduce more readily. This effect can be achieved in different ways. In [Siedlecki and Sklansky, 1989] two chromosomes were selected at random and the crossover was performed with a probability dependent on a function of the fitness for each chromosome. On the other hand, Prakash and Narasimba Murty [Prakash and Narasimha Murty, 1995] used the fitness function values to bias the probability of selecting a particular pair of chromosomes for applying the crossover operator. Once they were selected the crossover was guaranteed to occur. The crossover procedure itself is performed by choosing a point along the chromosomes, at random, and exchanging the chromosome pieces at the selected point.

   Mutation The procedure for this genetic operator starts with a randomly selected chromosome. Then each gene in the chromosome is considered. The gene is switched from 0 to 1 or 1 to 0 with a predefined probability, the mutation rate[2]. For this operator the chromosome's fitness is not taken into account. This enhances the likelihood that the genetic algorithm will find the global extremum. At any time during the procedure a new chromosome

[2] There is an alternative approach where two random numbers are selected. The first is to determine if the gene should have the opportunity to be changed and the second is to determine the actual value that should be given to that gene.

has a non-zero probability of forming the necessary genes to place it near the global extremum regardless of the previous history of *any* chromosome in the population.

Miscellaneous The aforementioned genetic operators are the ones that appear to be common to most studies employing a genetic algorithm. However, there are variations depending on the application. For example, Srikanth *et al.* [Srikanth et al., 1995] allowed the size of the chromosomes to vary. To allow changes in the chromosome length they implemented insertion and deletion of small sequences of genes as additional genetic operators.

3. Since the algorithm randomly changes the population, the procedure will only converge to a solution of the desired problem under the appropriate evolutionary pressure. This in turn requires a way to evaluate each chromosome to determine its fitness to the problem under consideration. The function can be very simple, such as a count of the number of correctly classified cases, but often modifications are made for various purposes. In particular, a fitness function that is too "severe" will make the chromosomes converge too quickly (stagnation) and the algorithm will likely get caught in a local extremum. On the other hand a function that allows too many unfit chromosomes to survive will require an excessively long time to converge to a result.

4. An important aspect of natural selection as used in a genetic algorithm is the removal of poor solutions. The central idea is to retain the best chromosomes and remove the worst. Often the retention and removal is constrained by the total number of chromosomes in the population. Of course, the exact approach of how to achieve the evolution of the population can vary. One common technique

is to completely remove the "parent" chromosomes and replace them by the new chromosomes (after crossover, mutation, etc.) In this case, the convergence to an optimal solution may be strictly asymptotic in that if a chromosome is created that corresponds to a global extremum, it is very likely the *exact* configuration will be lost in the creation of the next generation by the genetic operators. An alternative approach retains a number of the best chromosomes intact from the current generation (the elite population) and fills the remaining members of the population with the reproduced chromosomes. Alternatively, the parent and children can be placed in separate populations and the best, according to the fitness function, from either population is used for the next generation. This variation does not require a fixed number of members in the elite population and was the approach used in [Srikanth et al., 1995]. The formation of a new population ends the current generation and the procedure repeats with the evaluation of each chromosome in the population for reproduction. Typically, the population is allowed to evolve for a fixed number of generations.

Since the basic method is simple many variants have been used depending upon the nature of the problem under consideration. Indeed, many subtle differences from the basic technique appear throughout the literature.

Appendix C

Introduction to the Wavelet Transform

The usefulness of a transform that can be used to analyse the frequencies present in a signal cannot be understated but the typical approach, a Fourier transform, has some difficulties. Most notably, a Fourier transform has poor spatial resolution, which makes it inconvenient for analysing nonstationary signals (i.e. a large number of nonzero coefficients will always be required.). There are modifications that improve the situation such as the windowed Fourier transform which performs a conventional Fourier transform over a small region (window) of the signal at a time. However, the fixed size of the region may be ill suited for some applications:

1. Where the appropriate choice for the size of the window is not known *a priori.*

2. That do not have a single characteristic length [Aldroube and Unser, 1996].

An alternative approach is to use a wavelet transform. A wavelet transform has the flexibility to provide both good frequency and spatial localization. In addition, it provides a straightforward approach to providing multi-resolution analysis. Further, the discrete wavelet transform can be implemented very efficiently.

Wavelet transforms have been found to be useful in many areas, including numerous medical applications. For example, the transform has been used in the analysis of EKG and EEG signals [Unser and Aldroubi, 1996] as well as in image enhancement [Giger and MacMahon, 1996, Zhang et al., 1998]. A general introductory discussion can be found in [Strang, 1994], [Morgan, 1996] and [Langi, 1996] while an example-based description can be found in [Press et al., 1992]. Discussions that give a more mathematical formulation of the transform can be found in [Aldroube and Unser, 1996], [Cohen and Kovačević, 1996], [Harpen, 1998], [Jawerth and Sweldens, 1994], [Strang, 1989] and [Unser and Aldroubi, 1996] along with the classic reference [Daubechies, 1992]. In the discussion that follows it was assumed that the signal is discrete and one dimensional. The extension to multiple dimensions is straightforward.

The transform can be viewed in terms of a series of filters that were applied to the original signal. This interpretation is useful for efficient implementation of a discrete form of the wavelet transform but the transform itself has a substantial mathematical basis. Some of the fundamental mathematical concepts are described in the next section and further details may be found in the literature. Section C.2 describes the development from the fundamental mathematics to the filter bank description used in many of the transform's implementations.

C.1 Basic Wavelet Theory

The wavelet transform can be used to extract information in a hierarchial manner. The transform can be viewed as the projection of the signal onto many sets of basis functions which span various vector spaces in $\mathcal{L}_2$, the set of square integrable functions[1]. At each level of the hierarchial analysis, a subset of the original vector space was selected and broken into two smaller vector subspaces. In the most common approach, the bases for all vector spaces used in the transform can be created from the dilation and translation of two fundamental functions, the mother wavelet $\psi(x)$ and mother scaling function $\varphi(x)$. The basis functions are characterised by two parameters, a_j and $b_{j,k}$. This is in contrast to the Fourier transform where a single parameter is needed to identify each function in the basis, the frequency. The parameter a_j for a wavelet transform characterises the scaling of the mother function and is usually selected to have a form as given in Equations C.2–C.3 below. On the other hand, $b_{j,k}$ identifies the translation of the function relative to the mother function. The bases can then be written as

$$\begin{aligned} \psi_{a_j,b_{j,k}}(x) &= C_{a_j,b_{j,k}}\,\psi\left(\tfrac{x-b_{j,k}}{a_j}\right) \\ \varphi_{a_j,b_{j,k}}(x) &= D_{a_j,b_{j,k}}\,\varphi\left(\tfrac{x-b_{j,k}}{a_j}\right) \end{aligned} \quad (C.1)$$

where C_{a_j} and $D_{b_{j,k}}$ are normalization constants. With the typical choice for a_j and $b_{j,k}$ (for a discrete transform)

$$a_j = 2^j \quad (C.2)$$

[1] For all functions $f(x) \in \mathcal{L}_2$ the integral $\int f(x)f(x)dx = \kappa$, κ an arbitrary, but finite, constant and we define the inner product $\langle f(x), g(x)\rangle$ as $\int f(x)g(x)dx$.

$$b_{j,k} = 2^j k \tag{C.3}$$

Equation (C.1) can be written

$$\psi_{j,k}(x) = 2^{-j/2}\psi\left(2^{-j}x - k\right) \tag{C.4}$$

$$\varphi_{l,m}(x) = 2^{-l/2}\varphi\left(2^{-l}x - m\right) \tag{C.5}$$

The definition of the problem requires two sets of basis functions, $\{\psi_{j,k}\}$ and $\{\varphi_{l,m}\}$ which span two disjoint subspaces W_j and V_l respectively. The analysis can now be described as the projection of a function onto V_i which produces some lower resolution version (smoothed form) of the function while the projection onto W_i contains the information that is lost after the smoothing. Hence, the transformation is invertible.

One of the more useful aspects of the wavelet transform is that it can be made to perform a multi-resolution analysis of the input signal. This characteristic requires a series of embedded vector spaces, V_i (mathematically $V_i \subset V_{i-1} \forall\, i$). Each vector space, V_{i-1}, is further divided into the subspaces V_i and W_i in such a way that $\{\varphi_{i,j}\}$ continues to span V_i and $\{\psi_{i,j}\}$ spans W_i. As well, the subspaces are necessarily divided to satisfy the constraint that the combination of the subspaces V_i and W_i are equivalent to the subspace V_{i-1} for all i or

$$V_i \oplus W_i = V_{i-1}, \forall\, i. \tag{C.6}$$

For a wavelet transform with an infinite number of subspaces ($i \to \infty$), only the projections onto either the $\{V_i\}$ or $\{W_i\}$ are necessary. The "missing" projections can be calculated from the set that was retained. So, if the decomposition is carried out indefinitely on a function, $f(x)$,

$$f(x) = \sum_{j,k} c_{j,k}\, \psi_{j,k} \tag{C.7}$$

for some set of constants $c_{j,k}$. However, if the decomposition was stopped at the J^{th} subspace (hereafter referred to as "levels") then the transform becomes

$$f(x) = \sum_{j=0}^{J} \sum_{k} c_{j,k}\, \psi_{j,k} + \sum_{k} d_{J,k}\, \varphi_{J,k} \tag{C.8}$$

and $\sum_{k} d_{J,k}\varphi_{J,k}$ is the resulting smoothed signal on the last level of the decomposition. Calculating the wavelet transform then becomes the task of finding $\{c_{j,k}, d_{J,k}\}$ or more generally $\{c_{j,k}, d_{j,k}\}$.

Conceptually, the simplest form of wavelet transform is obtained when the bases are orthogonal

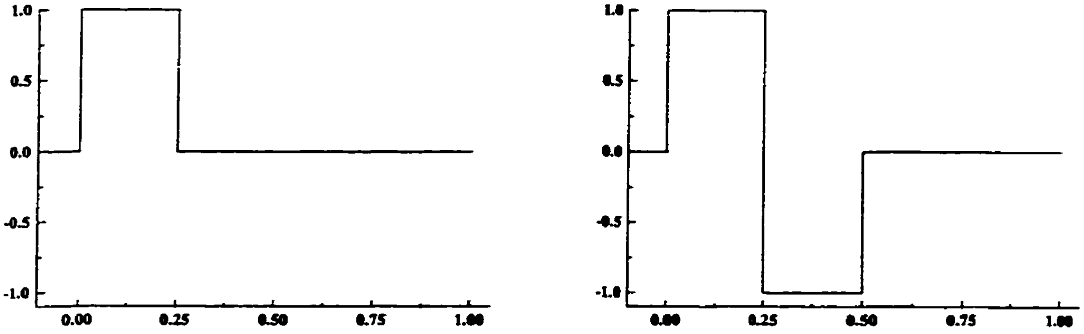
$$\begin{aligned} \langle \varphi_{j,k}, \varphi_{l,m} \rangle &= \delta_{j,l}\delta_{k,m} \\ \langle \psi_{j,k}, \psi_{l,m} \rangle &= \delta_{j,l}\delta_{k,m} \\ \langle \varphi_{j,k}, \psi_{l,m} \rangle &= 0 \end{aligned} \tag{C.9}$$

where $\langle f(x), g(x) \rangle = \int f(x)g(x)dx$ for arbitrary functions $f(x)$ and $g(x)$. Then, finding $c_{j,k}$ and $d_{j,k}$ is straightforward

$$\begin{aligned} c_{j,k} &= \langle f(x), \psi_{j,k} \rangle \\ d_{j,k} &= \langle f(x), \varphi_{j,k} \rangle \end{aligned} \tag{C.10}$$

These types of wavelet transforms are especially useful in various signal compression applications as it is easy to estimate the error introduced by ignoring terms with contributions less than a given amount.

An orthogonal basis may not have the desired characteristics for some applications and for these it may be more judicial to use bi-orthogonal wavelets. For instance, if the desired properties are orthogonality (for ease in finding $\{c_{j,k}, d_{j,k}\}$ and to prevent redundancies in the coefficients), compact support (for convenience when transforming a finite sized signal) and symmetry (for convenience when transforming symmetric signals) there is only one possible choice, a Haar basis (Figure C.1) [Daubechies, 1992].



(a) Scaling function, φ

(b) Wavelet function, ψ

Figure C.1: Example of a one dimensional scaling function (C.1(a)) and a one dimensional wavelet function (C.1(b)) in the Haar basis.

The drawback with a Haar basis is that is not smooth so that it is *inconvenient* for transforming a smooth signal. That is, attempting to decompose a smooth signal through a combination of functions with sharp corners would require an excessively large number of components. A bi-orthogonal wavelet transform can be used for greater flexibility in the choice of basis functions with only a small additional effort in the calculation of the transform itself.

A bi-orthogonal basis loses the very restrictive properties in Equation (C.9). This modification allows the freedom to select functions with other desirable properties, such as smoothness. However without the properties in Equation (C.9) it could be difficult to find $c_{j,k}$ and $d_{j,k}$, in general. This difficulty is avoided by a judicious selection of basis functions. The specific bases are selected such that for each level i there exists a pair of vector spaces $\widetilde{V_i}$ and $\widetilde{W_i}$ that are dual to V_i and W_i, respectively. The duals can be viewed as an alternative subdivision of the same subspace spanned by V_i and W_i combined. The dual bases still possess the properties $\widetilde{V_i} \oplus \widetilde{W_i} = \widetilde{V_{i-1}}$ and the alternative subdivision of the subspace is performed such that the following conditions hold (Equations C.11).

$$\begin{aligned} \langle \varphi_{j,k}, \widetilde{\varphi_{l,m}} \rangle &= \delta_{j,l}\delta_{k,m} \\ \langle \psi_{j,k}, \widetilde{\psi_{l,m}} \rangle &= \delta_{j,l}\delta_{k,m} \\ \langle \varphi_{j,k}, \widetilde{\psi_{l,m}} \rangle &= 0 \\ \langle \psi_{j,k}, \widetilde{\varphi_{l,m}} \rangle &= 0 \end{aligned} \tag{C.11}$$

See for example [Jawerth and Sweldens, 1994]. These requirements restore the simplicity in calculating the $\{c_{i,j}, d_{i,j}\}$ coefficients. The constants are found by taking the inner product of the function with the dual bases rather than the original basis.

$$\begin{aligned} c_{j,k} &= \langle f(x), \widetilde{\psi_{j,k}} \rangle \\ d_{j,k} &= \langle f(x), \widetilde{\varphi_{j,k}} \rangle \end{aligned} \tag{C.12}$$

The introduction of the dual spaces effectively separates the functions used for the forward and inverse transform and allows the wavelet and scaling function bases to be independent. This in turn allows the properties for the wavelet and scaling functions to be selected independently. For example, the smoothness of a wavelet competes with its compactness, a smoother function tends to be less compact. Since the projection of the function onto V_i produces a result which is like a smoothed form of the function, a smooth basis for V_i would be desirable. On the other hand, the projection onto W_i reflects the information lost from the projection onto V_i and contains the higher frequency components so that a more compact basis is more convenient. The bi-orthogonal transform allows both requirements to be fulfilled while retaining the efficiency of an orthogonal transform.

Of course, the difficult aspect of a bi-orthogonal transform is in the selection of the bases that satisfy all the requirements. Quite fortunately, Sweldens [Sweldens, 1994, Sweldens, 1995] has developed a method, called lifting, where the appropriate bases with the desired properties can be generated automatically. Specifically, the procedure starts with a known basis, such as a Haar basis, and takes linear combinations of the functions (lifts it) in such a way to form a new basis that has the desired properties while maintaining the constraints necessary for a wavelet transform. The precise linear combinations that are taken are complex. Details can be found in [Sweldens, 1994] and [Sweldens, 1995].

C.2 Wavelet Transforms as Filter Banks

The wavelet transform was described, in the previous section, in terms of inner products with many sets of basis functions and it may appear difficult to implement the transform efficiently. In practice, a simple and fast procedure to perform the calculation has been found: the fast wavelet transform. The following discussion is based on that in the thesis of Langi [Langi, 1996]. We assume the orthogonal transform is used but the analogous results can be derived for a bi-orthogonal case.

The coefficients $c_{j,k}$ and $d_{j,k}$ are given by

$$\begin{aligned} d_{j,k} &= \langle f(x), \varphi_{j,k}\rangle & \text{(C.13)} \\ c_{j,k} &= \langle f(x), \psi_{j,k}\rangle & \text{(C.14)} \end{aligned}$$

First consider $d_{j,k}$. Since $V_i \subset V_{i-1}$, $\varphi_{j,k}$ can be written as

$$\varphi_{j,k} = \sum_l \langle \varphi_{j,k}, \varphi_{j-1,l}\rangle \varphi_{j-1,l} \qquad \text{(C.15)}$$

and the $\langle \varphi_{j,k}, \varphi_{j-1,l}\rangle$ are a set of constants, h_l. In addition, the functions

$$\varphi_{j,k}(x) = 2^{-j/2}\varphi(2^{-j}x - k) \qquad \text{(C.16)}$$

are orthogonal. Therefore, $\langle \varphi_{j,k}, \varphi_{j-1,l}\rangle \neq 0$ iff

$$\begin{aligned} 2^{-j}x - k &= 2^{-(j-1)}x - l & \text{(C.17)} \\ 2^{-j}x &= l - k & \text{(C.18)} \\ x &= 2^j(l-k) & \text{(C.19)} \end{aligned}$$

and the argument for $\varphi_{j,k} \neq 0$ occurs when

$$\begin{aligned} (2^{-j}x - k)\ |_{2^j(l-k)} &= 2^{-j}[2^j(l-k)] - k & \text{(C.20)} \\ &= l - k - k & \text{(C.21)} \\ &= l - 2k & \text{(C.22)} \end{aligned}$$

Finally,

$$\langle \varphi_{j,k}, \varphi_{j-1,l} \rangle = h_{l-2k}\delta[x, 2^j(l-k)], h_{l-2k} \in \mathbf{R} \tag{C.23}$$

Combining equations (C.13), (C.15) and (C.23) gives

$$\begin{aligned} d_{j,k} &= \langle f(x), \varphi_{j,k} \rangle && \text{(C.24)} \\ &= \langle f(x), \textstyle\sum_l h_{l-2k}\delta[x, 2^j(l-k)]\varphi_{j-1,l} \rangle && \text{(C.25)} \\ &= \sum_l h_{l-2k}\delta[x, 2^j(l-k)]\langle f(x), \varphi_{j-1,l} \rangle && \text{(C.26)} \\ &= \sum_l h_{l-2k}\delta[x, 2^j(l-k)] d_{j-1,l} && \text{(C.27)} \end{aligned}$$

Similarly, for the other set of coefficients, $c_{j,k} = \langle x, \psi_{j,k} \rangle$, recall $V_j \oplus W_j = V_{j-1}$ which implies $W_j \subset V_{j-1}$ so, again, it is possible to write

$$\begin{aligned} \psi_{j,k} &= \sum_l \langle \psi_{j,k}, \varphi_{j-1,l} \rangle \varphi_{j-1,l} && \text{(C.28)} \\ &= \sum_l g_l \varphi_{j-1,l} && \text{(C.29)} \end{aligned}$$

and we define $g_l \equiv \langle \psi_{j,k}, \varphi_{j-1,l} \rangle$. Similar to the derivation outlined for $\langle \varphi_{j,k}, \varphi_{j-1,l} \rangle$, it can be shown that

$$\langle \psi_{j,k}, \varphi_{j-1,l} \rangle = g_{l-2k}\delta[x, 2^j(l-k)], g_{l-2k} \in \mathbf{R} \tag{C.30}$$

and

$$c_{j,k} = \sum_l g_{l-2k}\delta[x, 2^j l - k] c_{j-1,l} \tag{C.31}$$

The results (C.27) and (C.31) are quite important and several observations can be made.

1. The coefficients for each succeeding level can be calculated using strictly the coefficients from the previous level.

2. The equations ((C.27) and (C.31)) can be viewed as the application of a filter onto the input signal [Antoniou, 1979]. From this point of view, the constants h_{l-2k} and g_{l-2k} function as the kernel of the filters and since the equations are recursive, the wavelet transform is often described as a "tree" of filter banks applied to a signal. (See [Langi, 1996].)

3. Not all the coefficients on the $(j-1)^{\text{th}}$ level are needed to form the coefficients on the j^{th} level. A dyadic sequence was used for a_j and $b_{j,k}$ and for this case only every other coefficient is used.

A wavelet transform is usually constructed such that the filter with the h kernel acts as a low pass filter or a projection onto V_i and the filter with the g kernel acts as a high pass filter (or a projection onto W_i), followed by sub-sampling by two. The output from the low pass filter is then analysed again by another pair of filters followed by sub-sampling. The process can be repeated as often as desired. From this point of view a fast implementation for the forward and inverse wavelet transform is not difficult.

The tree-of-filter-banks interpretation is important since it provides an intuitive interpretation of the meaning behind a wavelet transform and it also allows for some modifications to the classic wavelet transform that are difficult to conceptualise mathematically. In the conventional wavelet transform only the output of the low pass filters are input into the next level of the analysis but there is no fundamental reason for this configuration to be used exclusively. In principle, an arbitrary branch of the tree of filter banks can be used for the next level of the analysis. For example, a configuration as shown in Figure C.2 would not be forbidden. The arrangement of the filter banks can be varied depending on the desired characteristics of the signal which is to be captured. Indeed, such an altered transform has been used in many applications. A general description of trees of filter banks and an application to signal compression can be found in [Langi, 1996]. These arbitrary trees of filter banks are in fact the "variation of a wavelet transform" used in [Clarke et al., 1994, Qian et al., 1995] for microcalcification segmentation and described in Chapter 1.

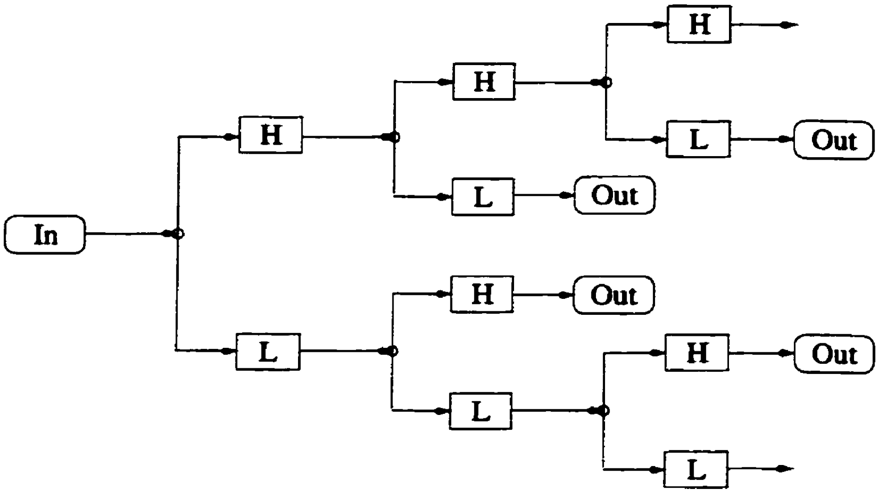


Figure C.2: Arbitrary tree of filter banks.

Glossary

ACR American College of Radiology

AEC Automatic Exposure Control. A device in a mammography unit which monitors the exposure and halts the beam of x-rays when the exposure reaches an upper limit.

ASCO American Society of Clinical Oncology

BiRADS Breast Imaging — Reporting and Data System [ACR, 1993]

CAD Computer Aided Diagnostics

CC Cranial Caudal, literally head to tail.

Compact The domain over which the function is non-zero is finite.

Confusion Matrix A two dimensional histogram of *a priori* classification (performed by the clinician in this case) and *posteriori* classification (performed by the program). Perfect classification produces a confusion matrix with elements only on the principal diagonal.

Covariance Matrix A matrix of calculated properties common in statistics. The matrix contains variances along the principal diagonal and covariances in the off diagonal elements.

Dysplasia Alteration in the size, shape and organization of cells forming the mammographic ducts.

DDSM Digital Database for Screening Mammography: A publicly available database of digitised screening mammograms from the University of South Florida.

HD Hurter and Driffield curve. The curve that describes the optical density of film as a function of log exposure.

Hyperplasia Abnormal increase in the number of normal cells in the duct epithelium.

kVp kiloVoltage peak. The power supply to a mammography unit is not strictly DC and the peak voltage across the x-ray tube is characterised by this quantity.

Laws' texture energy Laws' textures are calculated by filtering the image with filters defined by Laws and finding some statistic for a window around each pixel. The Laws' texture energy is found by filtering the image with a filter that enhances spots and lines, then calculating the standard deviation in a window centred over each pixel. Taylor [Taylor et al., 1990] normalised the values by the local contrast map. The procedure to find the local contrast map was the same except that the image was filtered with a smoothing filter (also created by Laws).

mAs milliAmpere seconds. A measure of the charge transported through the x-ray tube during an exam.

MTF Modulation Transfer Function. A measure of the relative magnitude of a signal after propagating though a system as compared to the original. For a perfect system the MTF is 1 for all frequencies contained in the original signal.

Nulliparity Never having carried a pregnancy.

Objective Function A function that quantifies the how well the selected properties correctly classify the sample cases.

ROC Receiver Operating Characteristic curve. A plot of the true positive probability as a function of the false positive probability.

ROI Region of Interest

SCC Six-Category Classification scheme of the mammographic density due to Byng et. al [Byng et al., 1994]

SGLD or SGLD$_{d,\theta}$ Spatial Grey Level Dependence matrix. A 2 dimensional array which is a function of 2 variables, d and θ. Each entry in the array consists of the probability for finding a pair of pixels with grey levels i and j separated by a distance d and with an orientation θ.

Unsharp Mask A method of enhancing the high frequency components of an image. First a lowpass filtered version of the image is formed. Then, the original image is weighted by a user defined amount (amplification factor) and the lowpass image is subtracted from the weighted original image.

Bibliography

[ACR, 1993] ACR, editor (1993). *Breast Imaging — Reporting and data system (Bi-RADS).* American College of Radiology.

[Addison, 1997] Addison, P. (1997). *Fractals and Chaos: An Illustrated Course.* Institute of Physics Publishing.

[Adler and Wahl, 1995] Adler, D. and Wahl, R. (1995). New methods for imaging the breast: Techniques, findings and potential. *Am. J. Radiol.*, 164:19–30.

[Aldroube and Unser, 1996] Aldroube, A. and Unser, M., editors (1996). *Wavelets in Medicine and Biology*, chapter 1. CRC Press.

[Anastasio et al., 1998] Anastasio, M., Yoshida, H., Nagel, R., Nishikawa, R., and Doi, K. (1998). A genetic algorithm-based method for optimizing the performance of a computer-aided diagnosis scheme for detection of clustered microcalcifications in mammograms. *Med. Phys.*, 25(9):1613–1619.

[Antoniou, 1979] Antoniou, A. (1979). *Digital Filters: Analysis and Design.* McGraw-Hill.

[Arthur et al., 1990] Arthur, J., Ellis, I., Flowers, C., Roebuck, E., Elston, C., and Blamey, R. (1990). The relationship of "high risk" mammographic patterns to histological risk factors for development of cancer in the human breast. *Br. J. Radiol.*, 63(755):845–849.

[Atkinson et al., 1999] Atkinson, C., Warren, R., Bingham, S., and Day, N. (1999). Mammographic patterns as a predictive biomarker of breast cancer risk: effect of tamoxifen. *Canc. Epidemiol. Bio. Prev.*, 8:863–866.

[Bernstein et al., 1988] Bernstein, I., Garbin, C., and Teng, G. (1988). *Applied Multivariate Analysis.* Springer-Verlag.

[Bevington, 1969] Bevington, P. (1969). *Data Reduction and Error Analysis for the Physical Sciences.* McGraw-Hill Book Co.

[Boone et al., 1998] Boone, J. M., Lindfors, K. K., Beatty, C. S., and A., S. J. (1998). A breast density index for digital mammograms based on radiologists' ranking. *J. Digit. Img.*, 35(2):235–247.

[Boyd et al., 1997] Boyd, N., Greenberg, C., Lockwood, G., Little, L., Martin, L., Byng, J., Yaffe, M., and Tritchler, D. (1997). Effects at two years of a low-fat high-carbohydrate diet on radiologic features of the breast: Results from a randomized trial. *J. Natl. Cancer Inst.*, 89(7):488–496.

[Boyd et al., 1995] Boyd, N. F., Byng, J. W., Jong, R. A., Fishell, E. K., Little, L. e., Miller, A. B., Lockwood, G. A., Tritchler, D. L., and Yaffe, M. J. (1995). Quantitative classification of mammographic densities and breast cancer risk: Results from the canadian national breast screening study. *J. Natl. Cancer Inst.*, 87:670–675.

[Boyd et al., 1982] Boyd, N. F., O'Sullivan, B., Campbess, J. E., Fishell, E., Simor, I., Cooke, G., and Germanson, T. (1982). Mammographic signs as risk factors for breast cancer. *Br. J. Cancer*, 45:185–193.

[Breitenstein and Shaw, 1998] Breitenstein, D. S. and Shaw, C. C. (1998). Comparison of three tissue composition measurement techniques using digital mamograms — a signal-to-noise study. *J. Digit. Img.*, 11(3):137–150.

[Brisson et al., 1982a] Brisson, J., Merletti, F., Sadowsky, N., Twaddle, J., Morrison, A., and Cole, P. (1982a). Mammographic features of the breast and breast cancer risk. *Am. J. Epidemiol.*, 115(3):428–437.

[Brisson et al., 1984] Brisson, J., Morrison, A., , Kopans, D., Sadowsky, N., Kalisher, L., Twaddle, J., Meyer, J., Henschke, C., and Cole, P. (1984). Height, weight, mammographic features of breast tissue and breast cancer risk. *Am. J. Epidemiol.*, 119(3):371–381.

[Brisson et al., 1982b] Brisson, J., Sadowsky, N., Twaddle, J., Morrison, A., Cole, P., and Merletti, F. (1982b). The relation of mammographic features of the breast to breast cancer risk factors. *Am. J. Epidemiol.*, 115(3):438–443.

[Burattini et al., 1995] Burattini, E., Cossu, E., Di Maggio, C., Gambaccini, M., Indovina, P., Marziani, M., Pocek, M., Simeoni, S., and simonetti, G. (1995). Mammography with synchrotron radiation. *Radiology*, 195:239–244.

[Byng et al., 1994] Byng, J. W., Boyd, N. F., Fishell, E., Jong, R. A., and Yaffe, M. J. (1994). The quantitative analysis of mammographic densities. *Phys. Med. Biol.*, 39:1629–1638.

[Byng et al., 1996a] Byng, J. W., Boyd, N. F., Fishell, E., Jong, R. A., and Yaffe, M. J. (1996a). Automated analysis of mammographic densities. *Phys. Med. Biol.*, 41:909–923.

[Byng et al., 1996b] Byng, J. W., P., C. J., Boyd, N. F., Little, L., Lockwood, G., Jong, R. A., Fishell, E., Tritchler, D., and Yaffe, M. J. (1996b). Analysis of digitized mammograms for the prediction of breast cancer risk. In Doi, K., Giger, M. L., Nishikawa, R. M., and Schmidt, R. A., editors, *Digital Mammography '96*, pages 185–190. Elsevier Science B. V.

[Byng et al., 1997] Byng, J. W., Yaffe, M. J., Lockwood, G., Little, L., Tritchler, D., and Boyd, N. (1997). Automated analysis of mammographic densities and breast carcinoma risk. *Cancer*, 80(1):66–74.

[Byng et al., 1999] Byng, J. W., Yaffe, M. J., Lockwood, G. A., Little, L. E., Tritchler, D. L., and Boyd, N. F. (1999). Automated analysis of mammographic densities and breast cancer risk. *to be published.*

[Caldwell et al., 1990] Caldwell, C. B., Stapleton, S. J., Holdsworth, D. W., Jong, R. A., Weiser, W. J., Cooke, G., and Yaffe, M. J. (1990). Characterisation of mammographic parenchymal pattern by fractal dimension. *Phys. Med. Biol.*, 35(2):235–247.

[Chaikin and Lubensky, 1995] Chaikin, P. and Lubensky, T. (1995). *Principles of Condensed Matter Physics*. Cambridge University Press.

[Chan et al., 1998] Chan, H. P., Sahiner, B., Lam, K. L., Petrick, N., Helvie, M., Goodsitt, M., and Adler, D. (1998). Computized analysis of mammographic microcalcifications in morphological and texture feature spaces. *Med. Phys.*, 25:2007–2019.

[Chan et al., 1995] Chan, H. P., Wei, D., Helvie, M. A., Sahnier, B., Adler, D. D., Goodsitt, M. M., and Petrick, N. (1995). Computer-aided classification of mammographic masses and normal tissue: Linear discriminant analysis in texture feature space. *Phys. Med. Bio.*, 40:857–876.

[Chen et al., 1996] Chen, Y., Thompson, W., Semenciw, R., and Mao, Y. (1996). Epidemiology of contralateral breast cancer. *Ca. Epidemiol. Biomarkers and Prev.*, 34(3):565–596.

[Clarke et al., 1994] Clarke, L., Kallergi, M., Qian, W., Li, H. D., Clark, R., and Silbiger, M. (1994). Tree-structured non-linear filter and wavelet transform for microcalcification segmentation in digital mammography. *Canc. Lett.*, 77:173–181.

[Cohen and Kovačević, 1996] Cohen, A. and Kovačević, J. (1996). Wavelets: The mathematical background. *Proc. of IEEE*, 84(4):514–522.

[Daubechies, 1992] Daubechies, I. (1992). *Ten Lectures on Wavelets.* SIAM.

[Fernández et al., 1996] Fernández, G., Periaswamy, S., and Sweldens, W. (1996). LIFTPACK: A software package for wavelet transforms using lifting. In Unser, M., Aldroubi, A., and Laine, A. F., editors, *Wavelet Applications in Signal and Image Processing IV*, pages 396–408. Proc. SPIE 2825.

[Fisher, 1936] Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188.

[Flury and Riedwyl, 1988] Flury, B. and Riedwyl, H. (1988). *Multivariate Statistics: A Practical Approach.* Chapman and Hall.

[Gajalakshmi et al., 1998] Gajalakshmi, C., Shanta, V., and Hakama, M. (1998). Risk factors for contralateral breast cancer in chennai (madras), india. *Int. J. Epid.*, 27:743–750.

[Giger and MacMahon, 1996] Giger, M. and MacMahon, H. (1996). Image processing and computer aided diagnosis. *Radiol. Clin. N. Am.*, 34(3):565–596.

[Goldstein, 1980] Goldstein, H. (1980). *Classical Mechanics, Second Edition.* Addison-Wesley Publishing Company, Inc.

[Goodwin and Boyd, 1988] Goodwin, P. and Boyd, N. (1988). Mammographic parenchymal pattern and breast cancer risk: A critical appraisal of the evidence. *Am. J. Epidemiol.*, 127(6):1097–1108.

[Gupta and Undrill, 1995] Gupta, R. and Undrill, P. E. (1995). The use of texture analysis to delineate suspicious masses in mammography. *Phys. Med. Bio.*, 40:835–855.

[Haralick et al., 1973] Haralick, R. M., Shanmugam, K., and Dinstein, I. (1973). Texture features for image classification. *IEEE Trans. Syst. Man. Cybern.*, 3:610–621.

[Harpen, 1998] Harpen, M. (1998). An introduction to wavelet theory and application for the radiological physicist. *Med. Phys.*, 25:1985–1993.

[Heath and Bowyer, 1998] Heath, M. and Bowyer, K. W. Kopans, D. (1998). Current status of the digital database for screening mammography. *Digit. Mammography*, 1:457–460.

[Holland, 1975] Holland, J. (1975). *Adaptation in Natural and Artifical Systems.* The University of Michigan Press.

[Huo et al., 2000] Huo, Z., Giger, M., Wolverton, D., Zhong, W., Cumming, S., and Olopade, O. (2000). Computerized analysis of mammographic parenchymal patterns for breast cancer risk assessment: Feature selection. *Med. Phys.*, 27(1):4–12.

[Ingal et al., 1998] Ingal, V., Beliaevskaya, E., Brianskaya, A., and Merkurieva, R. (1998). Phase mammography — a new technique for breast investigation. *Phys. Med. Biol.*, 43:2555–2567.

[Jawerth and Sweldens, 1994] Jawerth, B. and Sweldens, W. (1994). An overview of wavelet based multiresolution analyses. *SIAM Rev.*, 36(3):377–412.

[Jones, 1992] Jones, C. (1992). Methods of breast imaging. *Phys. Med. Biol.*, 27:463–499.

[Karssemeijer, 1998] Karssemeijer, N. (1998). Automated classification of parenchymal patterns in mammograms. *Phys. Med. Biol.*, 43:365–378.

[Lado et al., 1995] Lado, M., Tahoces, P., Méndez, A., Souto, M., and Vidal, J. (1995). A wavelet-based algorithm for detecting clustered microcalcification in digital mammograms. *Med. Phys.*, 26(7):1294–1305.

[Langi, 1996] Langi, A. (1996). *Wavelet and Fractal Processing and Compression of Nonstationary Signals.* PhD thesis, University of Manitoba.

[Magnin et al., 1986] Magnin, I. E., Cluzeau, F., and L., O. C. (1986). Mammographic texture analysis: An evaluation of risk for developing breast cancer. *Opt. Eng.*, 25(6):780–784.

[Mallat, 1989a] Mallat, S. G. (1989a). Multifrequency channel decompositions of images and wavelet models. *IEEE Trans. Acoust. Speech signal Process.*, 37(12):2091–2110.

[Mallat, 1989b] Mallat, S. G. (1989b). Multiresolution approximations and wavelet orthonormal bases of $l_2(\mathbf{R})$. *Trans. Amer. Math. Soc.*, 315(1):69–87.

[Manly, 1986] Manly, B. (1986). *Multivariate Statistical Mehtods: A Primer.* J. W. Arrowsmith Ltd., Bristol.

[McLachlan, 1992] McLachlan, G. (1992). *Discriminant Analysis and Statistical Pattern Recognition.* Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons.

[Morgan, 1996] Morgan, D. (1996). Multiresolution signal analysis and wavelet decomposition. *Embedded Sys. Prog.*, pages 30–48.

[NCSC, 1999] NCSC (1999). *National Cancer Institute of Canada: Canadian Cancer Statistics 1999.* National Cancer Institute of Canada.

[Newman, 1999] Newman, J. (1999). Recent advances in breast cancer imaging. *Radiol. Technol.*, 71(1):35–54.

[Nikouline, 1998] Nikouline, A. (1998). *New Preprocessing Methods for Better Classification of MR and IR Spectra.* PhD thesis, University of Manitoba.

[Nishikawa et al., 1996] Nishikawa, R., Schmidt, R., Papaioannou, J., Osnis, R., Heusler, R., Giger, M., Wolverton, D., Comstock, C., and Doi, K. (1996). Performance of a prototype clinical "intelligent" mammography workstation. In Doi, K., Giger, M. L., Nishikawa, R. M., and Schmidt, R. A., editors, *Digital Mammography '96*, pages 435–438. Elsevier Science B. V.

[Peitgen et al., 1992] Peitgen, H. O., Jürgens, H., and Saupe, D. (1992). *Chaos and Fractals, New Frontiers of Science.* Springer-Verlag.

[Petland, 1984] Petland, A. P. (1984). Fractal-based description of natural scenes. *IEEE Trans. Patt. Anal. Mach. Intel.*, Pami-6(6):661–674.

[Prakash and Narasimha Murty, 1995] Prakash, M. and Narasimha Murty, M. (1995). A genetic approach for selection of (near-) optimal subsets of principal components for discrimination. *Pat. Recog. Lett.*, 16:781–787.

[Press et al., 1992] Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. (1992). *Numerical Recipes in Fortran, The Art of Scientific Computing*, chapter 13, pages 584–599. Cambridge University Press, 2 edition.

[Qian et al., 1995] Qian, W., Kallergi, M., Clarke, L., Li, H. D., and Venugopal, P. (1995). Tree structured wavelet transform segmentation of microcalcifications in digital mammography. *Med. Phys.*, 22(8):1247–1254.

[Qian et al., 1999] Qian, W., Li, L., and Clarke, L. P. (1999). Image feature extraction for mass detection in digital mammography: Influence of wavelet analysis. *Med. Phys.*, 26(3):402–408.

[Reynolds, 1999] Reynolds, H. (1999). Advances in breast imaging. *Hematol. Oncol. Clin. North Am.*, 13(2):333–348.

[Säbel and Aichinger, 1996] Säbel, M. and Aichinger, H. (1996). Recent developments in breast imaging. *Phys. Med. Bio.*, 41:315–368.

[Saftlas et al., 1991] Saftlas, A. F., Hoover, R. N., Brinton, L. A., Szklo, M., Olson, D. R., Salane, M., and Wolfe, J. N. (1991). Mammographic densities and risk of breast cancer. *Cancer*, 67:2833–2838.

[Sahiner et al., 1996] Sahiner, B., Chan, H. P., Wei, D., Petrick, N., Helvie, M., Adler, D., and Goodsitt, M. (1996). Image feature selection by a genetic algorithm: Application to classification of mass and normal breast tissue. *Med. Phys.*, 23(10):1671–1684.

[Salminen et al., 1998] Salminen, T., Hakama, M., Heikkilä, M., and Saarenmaa, I. (1998). Favourable change in mammographic parenchymal patterns and breast cancer risk factors. *Int. J. Canc.*, 78:410–414.

[Schmidt and Nishikawa, 1995] Schmidt, R. and Nishikawa, R. (1995). Clinical use of digital mammography: The present and the prospects. *J. Digit. Imag.*, 8(1):74–79.

[Schroeder, 1991] Schroeder, M. (1991). *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise.* W. H. Freeman and Company.

[Shadagopan et al., 1982] Shadagopan, A., Alcorn, F. S., Semmlow, J. L., and Ackerman, L. V. (1982). Computerized quantification of breast duct patterns. *Radiol.*, 143:675–678.

[Siedlecki and Sklansky, 1989] Siedlecki, W. and Sklansky, J. (1989). A note on genetic algorithms for large-scale feature selection. *Pat. Recog. Lett.*, 10:335–347.

[Simonetti et al., 1998] Simonetti, G., Cossu, E., Montanaro, M., Caschili, C., and Giuliani, V. (1998). What's new in mammography. *Eur. J. Radiol.*, 27:S234–S241.

[Srikanth et al., 1995] Srikanth, R., George, R., Warsi, N., D., P., Petry, F., and Buckles, B. (1995). A variable-length genetic algorithm for clustering and classification. *Pat. Recog. Lett.*, 16:781–787.

[Strang, 1989] Strang, G. (1989). Wavelets and dilation equations: A brief introduction. *SIAM Rev.*, 31(4):614–627.

[Strang, 1994] Strang, G. (1994). Wavelets. *Am. Sci.*, 82:250–255.

[Sweldens, 1994] Sweldens, W. (1994). The lifting scheme: A custom-design construction of biorthogonal wavelets. Technical Report 1994:7, Industrial Mathematics Initiative, Dept. of Mathematics, University of South Carolina, ftp://ftp.math.sc.edu.

[Sweldens, 1995] Sweldens, W. (1995). The lifting scheme: A new philosophy in biorthogonal wavelet constructions. In Laine, A. F. and Unser, M., editors, *Wavelet Applications in Signal and Image Processing III*, pages 68–79. Proc. SPIE.

[Tabár and Dean, 1982] Tabár, L. and Dean, P. (1982). Mammographic parenchymal patterns: Risk indicator for breast cancer? *JAMA*, 247(2):185–189.

[Tahoces et al., 1995] Tahoces, P. G., Correa, J., Souto, M., Gómez, M., and Vidal, J. J. (1995). Computer-assisted diagnosis: The classification of mammographic breast parenchymal patterns. *Phys. Med. Biol.*, 40:103–117.

[Taylor et al., 1990] Taylor, P., Hajnal, S., Dilhuydy, M.-H., and Barreau, B. (1990). Measuring image texture to separate "difficult" from "easy" mammograms. *Br. J. Radiol.*, 35(2):235–247.

[te Brake et al., 1998] te Brake, G., Karssemeijer, N., and Hendriks, J. (1998). Automated detection of breast carcinomas not detected in a screening program. *Radiol.*, 207:465–471.

[Unser and Aldroubi, 1996] Unser, M. and Aldroubi, A. (1996). A review of wavelets in biomedical applictions. *Proc. of the IEEE*, 84(4):626–638.

[Urbanski et al., 1988] Urbanski, S., Jensen, H., Cooke, G., McFarlane, D., Shannon, P., Kruikov, V., and Boyd, N. (1988). The association of histological and radiological indicators of breast cancer risk. *Br. J. Canc.*, 58:474–479.

[Ursin et al., 1996] Ursin, G., Pike, M., Spicer, D., Porrath, S., and Reitherman, R. (1996). Can mammographic densities predict effects of tamoxifen on the breast? *J. Natl. Cancer Inst.*, 88(2):128–129.

[Veenland et al., 1998] Veenland, J. F., Grashuis, J. L., and Gelsema, E. S. (1998). Texture analysis in radiographs: The influence of modulation transfer function and noise on the discriminative ability of texture features. *Med. Phys.*, 25(6):922–936.

[Veenland et al., 1996] Veenland, J. F., Grashuis, J. L., van der Meer, F., Beckers, A. L. D., and Gelsema, E. S. (1996). Estimation of fractal dimension in radiographs. *Med. Phys.*, 23(4):585–594.

[Velanovich, 1996] Velanovich, V. (1996). Fractal analysis of mammographic lesions: A feasibility study quantifying the difference between benign and malignant masses. *Am. J. Med. Sciences*, 311(5):211–214.

[Weber, 1998] Weber, B. (1998). Update on breast cancer susceptibility genes. In Senn, H., Gelber, R., Goldhirsch, A., and Thürlimann, B., editors, *Recent Results in Cancer Research: Adjuvant Therapy of Primary Breast Cancer VI*, pages 49–59. Sringer-Verlag.

[Wei et al., 1995] Wei, D., Chan, H. P., Helvie, M. A., Sahiner, B., Petrick, N., and Adler, D. D. Goodsit, M. M. (1995). Classification of mass and normal breast tissue on digital mammograms: Multiresolution texture analysis. *Med. Phys.*, 22(9):1501–1513.

[Wei et al., 1997] Wei, D., Chan, H. P., Petrick, N., Sahiner, B., and Adler, D. D. Goodsit, M. M. (1997). False-positive reduction technique for detection of masses on digital mammograms: Global and local multiresolution texture analysis. *Med. Phys.*, 24(6):903–914.

[Wolfe, 1976a] Wolfe, J. N. (1976a). Breast patterns as an index of risk for developing breast cancer. *Am. J. Roent.*, 126:1130–1139.

[Wolfe, 1976b] Wolfe, J. N. (1976b). Risk for breast cancer development determined by mammographic parenchymal pattern. *Cancer*, 37:2486–2492.

[Wolfe et al., 1986] Wolfe, J. N., Saftlas, A. F., and Salane, M. (1986). Mammographic parenchymal patterns and quantitative evaluation of mammographic densities: A case-control study. *Am. J. Roent.*, 148:1087–1092.

[Yaffe and Rowlands, 1997] Yaffe, M. and Rowlands, J. (1997). X-ray detectors for digital radiography. *Phys. Med. Biol.*, 42:1–39.

[Zhang et al., 1998] Zhang, W., Yoshida, H., Nishikawa, R., and Doi, K. (1998). Optimally weighted wavelet transform based on supervised training for detection of microcalcifications in digital mammograms. *Med. Phys.*, 25(6):949–956.

[Zwillinger, 1996] Zwillinger, D., editor (1996). *CRC Standard Mathematical Tables and Formulae, 30th Edition.* CRC Press.