Chapter 1

Introduction

This thesis develops a statistical model to predict breast tissue pathology based upon coefficients produced by an adaptive multi-scale transform of small-angle x-ray scattering (SAXS) images. The adaptive transform devised in this thesis provides a multi-scale, multi-directional representation of a SAXS image. An entire library of filter functions is used within this transform to allow the best functions to be selected for classification. For the diagnosis of breast cancer using SAXS images, the functions selected from this library (for inclusion in the transform) are those that provided the best separation of the data associated with the different tissue pathologies. Statistical models are subsequently developed from the coefficients of the transformed SAXS patterns. These models estimate the probability of normal, benign and malignant tissue pathology using coefficients across scales and locations of the transform. The results of this model provide insight into those coefficients most indicative of a particular tissue pathology. The methodology is extended to include a logistic model that infers the tissue pathology associated with a particular SAXS image based upon *all* of the coefficients of the adaptive transform. This model proves itself to be capable for the detection and diagnosis of breast cancer. The analytical techniques presented in this thesis contribute to the task of pattern recognition using the multi-scale paradigm, as they can also be used as a general modeling framework that is applicable to a range of other SAXS image data.

1.1 Motivation for the Model

The need for a diagnostic model is driven by two priority areas in cancer research.
These are:

i) The analysis of large amounts of data collected in SAXS experiments and,

ii) The understanding of the differences in SAXS image produced by different tissue pathologies.

1.1.1 Analysis of Large Quantities of SAXS Images

X-ray scattering experiments often involve the collection of a large number of scattering patterns in the form of digital images. The intensity as a function of the position in the SAXS image provides insight into the structure of the specimen under investigation. A typical SAXS image contains hundreds of thousands of data records (262,144 records for a 512 x 512 pixel image), which contain important information on specimen structure. A review of the literature (Chapter 2) reveals several applications of such techniques to SAXS imaging. All approaches reviewed suffer from the same fundamental limitation, the researcher must propose a feature (which is application dependent) and then extract this feature at the sake of all of the other image information. For each application, the researcher must determine if a set of features is useful for the subsequent analysis. Even when the objectives of a research project have been achieved, a range of other features are available to explore that may or may not result in better performance. The problem could be considered open-ended unless all of the information contained within the image is included in the diagnostic model. Despite the fact that modern image processing methods can extract key features and compress the most pertinent information, they have limited capability for SAXS image analysis. Almost every single pixel of a SAXS image contains information concerning the structure under investigation. The only part of the image that does not contain structural information is that corresponding to the ‘shadow’ of the beam-stop, a region deliberately obscured by the experimental design to ensure the x-ray beam that is used to image the sample does not saturate (and possibly destroy) the electronic detectors that are used in these experiments. Furthermore, features may exist in the image that are not visible and that require sophisticated mathematical analysis in order to be understood. Reliance on visual identification of features might be useful for the completion of a specific task in the analysis of a large data set, but it may completely miss important information and result in sub-optimal or poor classification performance. Subtle or complex dependencies may exist amongst image pixels and simple extraction of an image feature without consideration of this other information might also produce poor classification results. A model that provides a coherent, objective approach to analysing all of the information in a SAXS image would alleviate the problem of subjective feature selection and expedite the process of data analysis.

1.2 SAXS Imaging in Cancer Research

There is scope to apply models of the adaptive transform coefficients to the diagnosis of breast cancer using SAXS images. A precedent has been set by researchers who have applied digital image processing and statistical techniques to analyse SAXS images of cancer tissue (Lewis *et al* 2000; Butler *et al* 2003; Erickson 2005). The objectives of these researchers was to use the SAXS images to gain new insights into the different pathologies of cancer tissue. Scattering patterns of breast cancer tissue have been analysed using clustering, data mining and wavelet techniques (Lewis *et al* 2000; Butler *et al* 2003; Erickson 2005). Brain cancer tissue has also been analysed using both hierarchical clustering and independent component analysis (Siu *et al* 2005; Falzon *et al* 2007). Further applications to images of other tissue pathologies are envisaged in the future and the challenge of feature detection, selection and extraction will need to be re-addressed. The wide range of new statistical models available has provided bewildering choice for practitioners and it is evident that a variety of approaches to cancer diagnosis have been considered in the literature. The choice of model has to be well considered in light of the data set and an inappropriate technique is likely to produce poor or misleading results. There is a pressing need for consistency and rigour among the different approaches adopted. The adaptive image transform analyses images at a variety of resolutions (scales) and locations with the coefficients encodings the results of this analysis. It is these coefficients that are input into a statistical model that produces a classification as output. One approach is to produce a statistical model that mimics the structure of the input data. That is, a statistical model with multiple levels is proposed.

1.2.1 SAXS Imaging of Breast Tumours

The patterns produced in SAXS experiments are due to the diffraction of radiation from the molecular structure of the sample under investigation. Nano-scale resolution (10-100s nm) is possible because of the wavelength of the radiation used, as such, SAXS has been used to provide insight into the nano-structure of human breast tissue (Lewis *et al* 2000; Ferńandez *et al* 2002; 2004; Round 2006). These studies found that the organisation of collagen molecules differs between normal, benign and malignant breast tissue. Lewis *et al* (2000) proposed an accurate diagnostic model of breast cancer by relating tumour type to the specific features of the scattering images. The images were processed using a method known as radial integration which is a semi-automated technique that transforms the two-dimensional image data into a one-dimensional format. An interactive peak fitting routine was then used to fit a profile to this one-dimensional data. This profile was created by the summation of a quadratic trend and several smooth peaks. The diagnostic features of the Lewis *et al* (2000) model were based on the fraction of scattering and the position of the smooth peaks. These features were derived from a model that essentially fits an arbitrary number of pre-specified smooth components in an attempt to accurately represent the one-dimensional profile. Fit accuracy was based on maximising the Pearson correlation coefficient between the model and the profile. The combined process of a one-dimensional transformation and smoothing may have removed other useful information in the images. The model proposed by Lewis *et al* (2000) might be enhanced if more advanced mathematical and statistical methods were employed to capture the information in the SAXS image.

Although the model of Lewis *et al* (2000) was ground-breaking it was limited in the following aspects:

a) It was based on a limited number of samples and,

b) Feature extraction is time-consuming and requires manual processing,

c) It was exploratory data analysis in that it did not perform any objective tests of classification performance and,

d) A model was used that smoothed the data and therefore it may have removed important diagnostic properties from the data.

Nonetheless, the model of Lewis *et al* (2000) motivated further investigations by Butler *et al* (2003) and Erickson (2005) aimed at producing improved diagnostic models.

1.2.2 Previous Image Analysis of SAXS data

Data mining and wavelet transformations have been used to analyse scattering images of breast tissue (Butler *et al* 2003; Erickson 2005). These models searched for features in the images useful for automated diagnosis of breast cancer. Both methods were very successful and identified features useful for classification. The wavelet model stood out because it gave perfect predictions of tumour type, verified using cross-validation (Erickson 2005). Despite these perfect results, there were three key limitations of Erickson's (2005) study:

a) A limited number of samples (n = 49) were used.

b) A number of models were developed using pairwise comparisons, when all three of the tissue groups were included the best model performance dropped to 86-90%. Erickson (2005) assumed that optimisation of the classifier to separate each pair of groups independently would translate into optimal separation of all three tissue groups, yet little evidence was provided to support this belief.

c) The model provided little insight into which features of the image were influencing classification. The '*energy*' (sums of squares of wavelet coefficients) of several (up to 5) bands were used as input features in a naive Bayes'ian classifier, the overall diagnostic performance was assessed but the impact of each feature in the model was not. The probability density functions for the data from each tissue group (the group-conditional densities) was not compared, the interactions and dependencies between features was not examined and penalty for model complexity was not implemented.

Despite the limitations, these results were very encouraging, suggesting that multi-scale (wavelet) models may be a powerful method to analyse SAXS image data.

1.3 Thesis Structure

1.3.1 Thesis Objectives

The objectives of this research are:

(a) To unite the capabilities of the adaptive transform in describing digital images and statistical inference within the framework of a mathematical model.

(b) To develop a unified, comprehensive, objective and semi-automated method with which to analyse SAXS images.

(c) To apply both the transform and the associated model to the diagnosis of breast cancer using SAXS images.

The model produced in this thesis provides good results when assessed by an independent test data set, supporting the notion that multi-scale modeling is a useful method to analyse SAXS image data.

1.3.2 Thesis Outcomes

SAXS images contain many features which are not intuitive for which specialised mathematical functions as well as the associated computing are proposed in order to facilitate extraction and analysis. The specific achievements of this thesis include:

a) An adaptive image transformation that allows analysis of the image across scales and locations (Section 7.2).

b) Specification of five new filter functions that might be useful for image analysis (Section 7.3).

c) Selecting the most suitable functions for use in the adaptive transform via the probability of misclassification (Section 7.4).

d) Specifying a mathematical model that succinctly describes the large number of probability density function estimates that are used in analysis of the SAXS images (Section 8.2.2).

e) Development of the Mexican hat contourlet transform: a multi-scale, multi-directional transformation of image and bivariate data (Section 8.2.4).

f) Non-parametric regression (smoothing) in the presence of heteroskedastic rather than white noise using the Mexican hat contourlet transform (Section 8.2.6.1).

g) Modifying the magnitude of Mexican hat contourlet transform coefficients using the statistical theory of extreme values (Section 8.2.8).

h) Probability density function estimation using the Walsh wavelet packet transformation (Section 8.3.4).

These achievements satisfy both methodological and applied needs, the models developed in this thesis casts multi-scale modeling into the realm of pattern recognition and at the same time provide useful tools to the SAXS imaging community.

1.3.3 Thesis Organisation

The thesis is organised into ten chapters. The original contributions are contained within Chapters 7 to 9. The preceding chapters present principles of SAXS image analysis so that the context of the new work can be appreciated. These chapters may be skimmed and referred back to as references to them are made during Chapters 7 to 9.

This introductory chapter concludes with an overview of the contents of Chapters 2 through to 10. Chapter 2 provides the background science to the SAXS imaging technique with a particular emphasis on the structure of breast tissue collagen. Chapter 3 reviews and critiques the previous literature on the analysis of SAXS patterns of cancer tissue. Image analysis techniques are surveyed in Chapter 4 and it is in this chapter the grounds for treating a SAXS pattern as a digital image are established. A range of existing image analysis transformations is examined in Chapter 5. The impacts of the SAXS experimental methodology on the subsequent analysis are examined in Chapter 6. Chapter 7 develops image analysis tools for use in this thesis. Specific results include the adaptive image transformation, the Witch's Hat filter function and methodology to select the best filter functions to include in the transform using the probability of misclassification. Statistical models are described in Chapter 8, including the Mexican hat contourlet transform and probability density function estimation using the Walsh wavelet packet transform. The methodology developed in the previous two chapters is applied to the diagnosis of breast cancer using SAXS images in Chapter 9. The final chapter, Chapter 10, reviews and summarises the main findings of this research with the future directions of this research being identified and discussed.

Chapter 2

The Scientific Background to the Diagnosis of Breast Cancer using Small-Angle X-ray Scattering

This chapter describes the scientific background related to the diagnosis of breast cancer using SAXS. Section 2.1 discusses the basics physics behind the SAXS technique, providing an understanding of how the observed SAXS pattern relates to the structure of the specimen under investigation. This will be a foundation for formulating statistical models to interpret the images and infer the state of the tissue as expressed in the SAXS pattern. The ultra-structure of collagen is reviewed in Section 2.2 and is related to specific features observed in SAXS patterns. The relationship between collagen structure and breast cancer is examined in Section 2.3, motivating the use of SAXS imaging to investigate breast cancer tissue structure. An overall summary in Section 2.4 finalises the chapter.

2.1 Small-Angle X-ray Scattering

Small-angle x-ray scattering refers to an imaging technique used to study the ultra-structure (10-100 nano-metres) of a material. The physics that produces these patterns needs to be explained so that the features present in the SAXS images (which are digital representations of the SAXS patterns) can be understood and related back to the structure of the tissue.

X-rays incident on a material undergo a variety of interactions including scattering and a loss of energy in the material in a process known as absorption. SAXS is based upon the physical mechanism of *Rayleigh scattering*. In contrast, the absorption mechanism is the basis of radiological imaging technology currently used in medical practice. The scattering of x-rays have been viewed as detrimental to the quality of images formed using the x-ray absorption mechanism such as mammography (Bushberg *et al* 2002). This is a considerable problem as Rayleigh scattering is a significant x-ray interaction in the mammographic energy range (20-30 keV) (Bushberg *et al* 2002). But unlike x-ray absorption, scattered photons convey information on the structure of the material with which they have interacted (Royle *et al* 1999). This makes x-ray scattering a useful technique when trying to understand the structure of the material, such as breast cancer, under investigation. Rayleigh scattering involves an incident photon interacting with an atom within the sample under investigation. The basic mechanism of Rayleigh scattering is displayed in Figure 2.1. The electric field of the incident photon induces oscillations in the electrons of the atom causing all of the electrons to vibrate in phase. Energy is immediately released in the form of a photon that is of the same energy, and hence wavelength, as the incident photon.

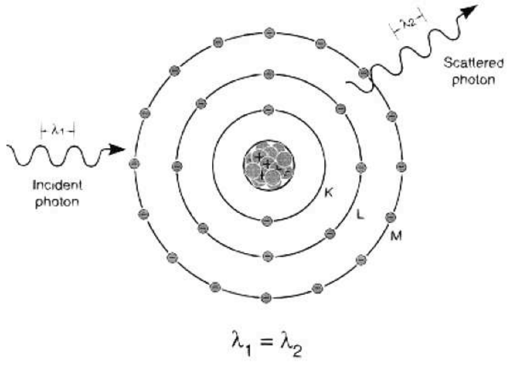

Figure 2.1: The Mechanism of Rayleigh scattering involves an incident photon of wavelength, λ_1 impinging on the outer shell of an atom. The photons energy is absorbed by the atom and released in the form of another photon of the same wavelength, λ_2 (Figure 3-6 Bushberg *et al* 2002).

SAXS experiments involve the use of a high intensity, high flux beam of photons incident on a sample. This sample can be described by an electron density distribution. The term $\rho^{\text{sample}}(\mathbf{r})$ is used to denote this electron density distribution, where $\mathbf{r}$ is a vector describing position within the sample. The sample is contained in a medium which also has an electron density distribution denoted by $\rho^{\text{ medium}}(\mathbf{r})$. When photons are scattered from the electron charge distribution of the form, $\rho'(\mathbf{r}) = |\rho^{\text{ sample}}(\mathbf{r}) - \rho^{\text{ medium}}(\mathbf{r})|$, a detector can be used to record the intensity, $I(\mathbf{h})$. The intensity is related to the structure of the material under investigation via the Fourier transform and its' conjugate (denoted $\mathcal{F}(\mathbf{h})$ and $\mathcal{F}^\star(\mathbf{h})$ respectively):

$$I(\mathbf{h}) = \mathcal{F}(\mathbf{h})\mathcal{F}^\star(\mathbf{h}) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \rho'(\mathbf{r}_1)\rho'(\mathbf{r}_2)\exp\big(i\mathbf{h}\cdot(\mathbf{r}_1 - \mathbf{r}_2)\big)d^3\mathbf{r}_1 d^3\mathbf{r}_2 \tag{2.1}$$

The scattering vector is defined as $\mathbf{h} = \mathbf{k}_2 - \mathbf{k}_1$ for x-rays that are incident on the electron charge distribution and traveling in the direction specified by the vector $\mathbf{k}_1 = (k_{1,x}, k_{1,y}, k_{1,z})$ and scattered x-rays traveling in the direction specified by vector $\mathbf{k}_2 = (k_{2,x}, k_{2,y}, k_{2,z})$ (page 578, Fernández *et al* 2002). Therefore the observed SAXS pattern is a Fourier transform representation of the electron density distribution of the specimen being studied. Because phase information is not recorded by the detector, the Fourier transform cannot be directly back-transformed to yield the electron density distribution $\rho'(\mathbf{r})$ of the specimen minus the medium. Nonetheless, the SAXS pattern provides valuable information about the structure of the sample under investigation. Synchrotron radiation has been used in many SAXS experiments because it provides high-flux electro-magnetic radiation at sub-nanometre wavelengths, such wavelengths allow the required resolution of the pattern in the tens to hundreds of nanometres range.

2.2 Collagen Structure & SAXS

Section 2.1 highlighted the links between physics, mathematics and the structure of the specimen involved in SAXS research. Equation (2.1) demonstrated how the Fourier transform of the structure of the specimen under investigation is directly related to SAXS pattern intensity. Therefore, SAXS imaging is a tool that can be used in the investigation of the structure of an object including those of a biological origin. The collagen family of proteins are an example of one such structure amenable to SAXS experiments.

2.2.1 Collagen Structure

The hierarchical arrangement of the fibril-forming collagens (Types I, II, III, V, XI) give rise to characteristic features in SAXS patterns. Human breast tissue contains collagen types I and III in the greatest proportions, making them of great interest in this work (Kauppila *et al* 1998). The fibril-forming collagens are composed of key units of amino acids, which in turn form into large chains of approximatley 1000 amino acids (Bigi & Roveri 1996). Three of these chains wrap together in a tight helix to form a collagen molecule of 300 nanometres in length. The molecules order into higher levels of organisation to form a collagen fibril. Three key structures of interest are the :

a) The longitudinal ordering of collagen molecules along and within a collagen fibril (denoted ‘d’ in Figure 2.2).

b) The lateral organisation between adjacent collagen fibrils (symbol ‘D’ in Figure 2.2) .

c) The radius of the fibrils (symbol ‘R’ in Figure 2.2).

Figure 2.2 displays these features as well as the relationship between collagen structure and the SAXS pattern features. The longitudinal arrangement of over-lapping collagen molecules within a collagen fibril is termed the axial D-repeat. This structure results in regions of high and low electron density along the fibril with a period of 64 to 67 nanometres. Constructive and destructive interference of x-rays that have interacted with this structure give rise to a series of sharp peaks which are colloquially termed the '*axial peaks*'. The symbol 'd' in Figure 2.2 denotes the axial D-repeat structure and one of the associated axial peak features along the meridian of the scattering pattern. Random orientation of the fibrils with respect to the incident x-ray beam results in the axial peaks being smeared into large arcs in the SAXS pattern that are termed scattering rings. These features are evident in the SAXS pattern of Figure 2.2. The lateral organisation of collagen fibrils is not completely resolved. When viewed end-on so that the only the circular cross-sections are visible, the fibrils appear to be arranged in a 'quasi-hexagonal' shape with each fibril separated by approximately 100 nm (Fernández *et al* 2002). The spacing between the lateral arrangement of the fibrils gives rise to scattering peaks along the equator of the SAXS pattern. This is evident when examining the structure denoted by the symbol 'D' in Figure 2.2. Finally, the mean radius of the fibrils (denoted by symbol 'R') is also of interest and it can be extracted from the central region of the SAXS pattern. In practice, a breast tissue specimen contains many such collagen fibrils as well as other components such as adipose tissue. The contributions of the other components to the SAXS pattern can be filtered out by examining only those x-rays scattered at an angle of $(\pm 5^0)$ from the direction of the incident beam. Within this range the contributions of scatter from the majority of other tissue structures is assumed to be minimal. The main additional contributors to the SAXS pattern of breast tissue collagen are scatter from amorphous substances within the tissue, scatter from air and scatter from the capillary tube in which the sample is placed during the experiment. In the ideal setting, the scatter produced by these components does not produce visible ring features in the pattern but rather adds additional counts (noise) to the pattern intensity.

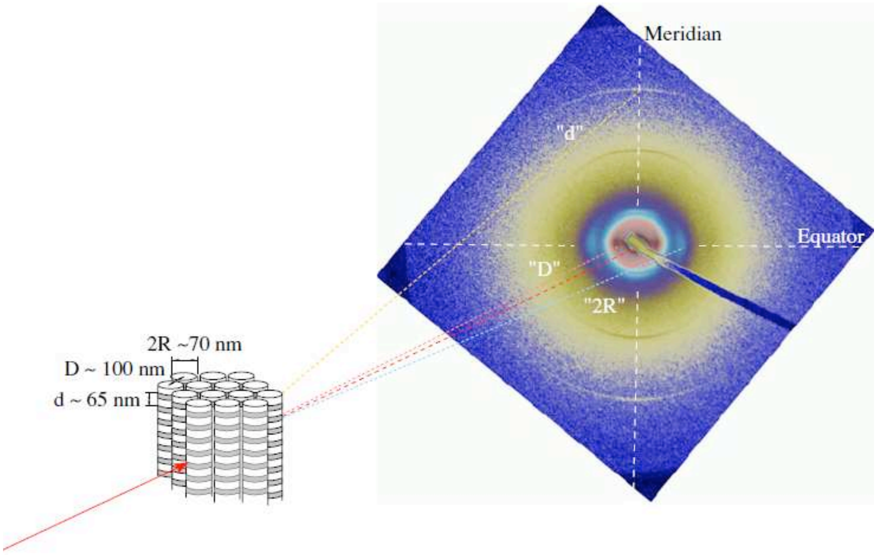

Figure 2.2: SAXS from oriented collagen fibrils, the symbol ‘d’ corresponds the longitudinal arrangement of collagen molecules, ‘D’ the distance between adjacent fibrils and ‘R’ the fibril radius (Fernández *et al* 2002)

2.2.2 Summary

Physical models of fibrillar collagens have many levels of hierarchical structure. These structures result in features that can be identified in their corresponding SAXS pattern. The same fibril contributes information to a number of different regions in the SAXS pattern. The very order of collagen suggests the interpretation of SAXS pattern features using multi-scale models. In such a framework, the logical flow between biological structure, the physical SAXS pattern and the statistical model is quite evident. To understand how this all relates to cancer research, it is necessary to understand the relationship between collagen structure and cancer.

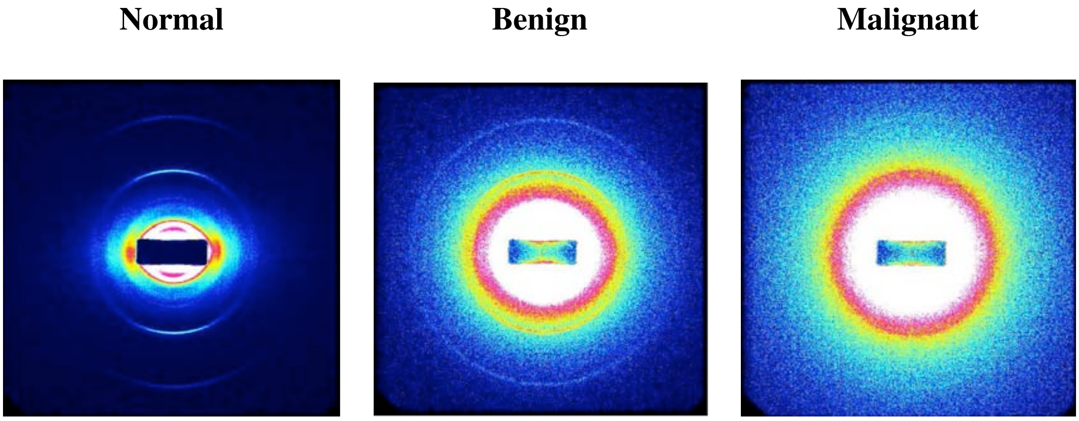

Figure 2.3: SAXS images of normal healthy, fibroadenoma (benign) and invasive carcinoma (malignant) pathologies (Falzon *et al* 2006).

2.3 Collagen Structure & Breast Cancer

2.3.1 Overview

Considerable scientific evidence exists for the alteration of collagen structure within malignant breast tissue. A new form of collagen, called OF/LB (onco-fetal laminin binding) collagen has been found to exist only in embryonic fetal and tumour tissue (Pucci-Minafra *et al* 1993). In fact, OF/LB collagen was found in both human breast and colon carcinoma tissue (Pucci-Minafra *et al* 1993). Metastasis is the spread of cancer cells from the primary tumour, and in breast cancer it may involve collagen structure. Wang *et al* (2002) investigated the relationship between metastasis and tissue structure using a murine (mouse) model. Multi-photon microscopy was used to study the growth of breast tumours derived from metastatic and non-metastatic cell lines. The cells from tumours undergoing metastasis were observed to not only adhere to, but also travel along collagen fibres (aggregations of fibrils) as they traveled towards blood vessels. In contrast, cells from tumours that were not undergoing metastasis were completely obstructed by the fibres. The amount, size and integrity of collagen fibres was greatly reduced in tumours derived from metastatic cell lines as compared to those that were not derived from such cell lines (Wang *et al* 2002).

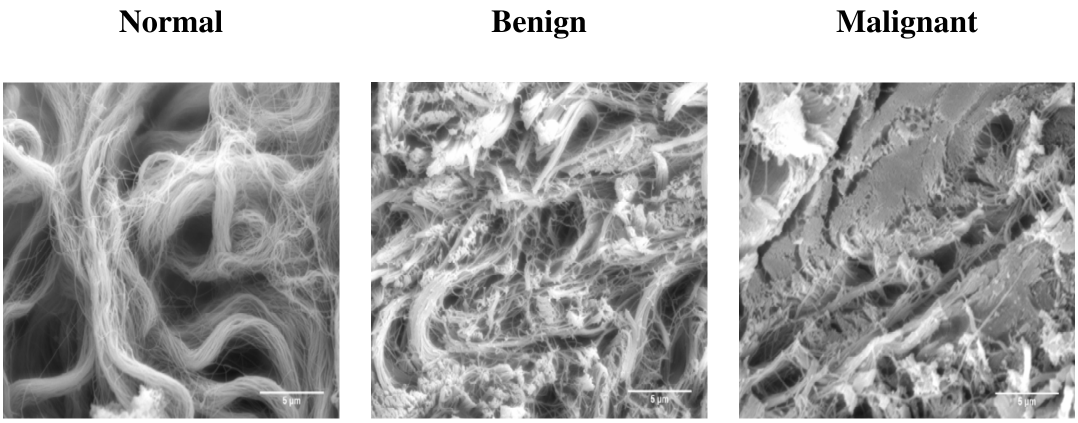

Figure 2.4: Electron microscopy images of collagen within breast tissue for the normal healthy, fibroadenoma (benign) and invasive carcinoma (malignant) pathologies, note the differences in the structure of collagen in each case, scale bar 5 μm (Reid 2006).

Lewis *et al* (2000) and Fernández *et al* (2002) conducted independent studies into the structure of cancerous breast tissue using the SAXS technique. Both groups have reported distinct differences in the SAXS images of normal healthy, benign and malignant breast tissue pathologies. Examples of such differences are displayed in Figures 2.3(a)-(c), observe that the SAXS image of malignant breast tissue appears to lack distinctive scattering ring features and has a much wider central disk (which is associated with the fibril radius R) than either the SAXS images of the normal or benign tissue pathologies. Such differences in the SAXS images suggest changes in the tissue structure that is associated with malignancy. This hypothesis was supported by an electron microscopy study that examined breast tissue structure on the micro-metre (μm) scale (Reid 2006). This length scale is 10-100 times greater than the scales probed by the SAXS technique and therefore the experimental results cannot be directly compared. Nonetheless, the electron microscopy images of Figures 2.4(a)-(c) also indicate a distinct change in the order and integrity of collagen within the sample. Reid (2006) found that normal healthy breast tissue appeared to have the greatest structural integrity and malignant breast tissue the least. It is evident from the SAXS and electron microscopy studies that any changes occurring in breast tissue structure as a result of malignancy occur over a range of length *scales*.

Both Lewis *et al* (2000) and Fernández *et al* (2002) have reported changes in the axial D-repeat structure in the presence of malignant breast tissue conditions. The change in the magnitude of the axial D-repeat can be understood in terms of Bragg's condition for constructive interference from an array. This condition relates the angle of the scattering to the period of a repetitive structure in the specimen. Bragg's condition states,

$$2d \sin\theta = n\lambda \tag{2.2}$$

where d is the scattering plane separation distance (the period),
n is the maxima order,
λ is the wavelength of incident radiation and,
2θ is the scattering angle, which is related to the scattering vector by
$|\mathbf{h}| = 4\pi \sin\theta/\lambda$ (equation 38-16, page 1181, Young & Freedman 1996).

Measurements of the period d of the structure can be made using diffraction (or scattering) experiments. In experiments involving fibrillar collagen, a periodic structure is produced by the alignment of collagen molecules within the axial D-repeat. The distance between scattering planes in this case is then the period of the axial D-repeat (labeled 'd' in Figure 2.2). According to Bragg's condition, a change in the magnitude in the axial D-repeat period relates to a shift in the angle 2θ of scattered x-rays. Hence the position of a peak (due to an intensity maximum) as recorded on a detector placed at a fixed distance from the sample will shift. Movements in peak position in SAXS images allowed both Lewis *et al* (2000) and Fernández *et al* (2002) to infer changes in magnitude of the axial D-repeat period for collagen from breast tumour tissue. The SAXS image features associated with the fibril centre-to-centre distance (labeled 'D' in Figure 2.2) have also been reported to be 'much broader' for breast tumour tissue (Fernández *et al* 2002). A physical explanation for this observation might be due to an increased variability in the centre-to-centre distance ('D') parameter for fibrils within the sample. A greater level of amorphous scatter has also been reported in SAXS images produced by malignant breast tissue samples (Round 2006). Greater proportions of adipose tissue or a general decrease in the overall structural order of the tissue might be responsible for this observation. Another possibility is that the collagen molecules are 'peeling-off' the collagen fibrils as the tissue is being invaded by cancer (Fernández *et al* 2004). This mechanism would be consistent with the observation of an increase in the level of amorphous scatter between the third and fifth order axial rings of SAXS images produced by malignant breast tissue (Fernández *et al* 2004). The proportion of scattering intensity occurring in the axial rings of the SAXS images has also been observed to decrease in malignant breast tissue conditions (Lewis *et al* 2000). The intensity within these rings is believed to be proportional to both the amount of order in the axial D-repeat structure and the total number of collagen fibrils in the sample. A change in either, or both, parameter would alter the observed image intensity.

2.4 Chapter Summary

The scientific background on the supra-molecular structure of fibrillar collagen and the evidence for alterations to its structure in breast cancer have been discussed. The fibrillar collagens found in human breast tissue possess a hierarchical structure that can be studied with SAXS imaging. Breast cancer produces distinct changes in SAXS image features that are known to be related to either collagen structure or quantity in the tissue. This knowledge establishes the grounds for further investigation and provides insight into some of the changes that occur in the images. Cancer diagnosis based upon SAXS images is a natural avenue of research to pursue. Statisticians wishing to produce diagnostic models of cancer using these images have the advantage of knowing those features worthy of pursuit. Furthermore, the very nature of the hierarchical structure of collagen suggests that a multi-scale modeling strategy might be a useful way to analyse this data. By understanding how the SAXS images were produced, significant gains can be made in both accurate modeling and interpretation.

Chapter 3

A Review and Critique on Interpreting Data from SAXS Images of Breast Cancer

This chapter reviews and critiques the relevant literature concerning the previous attempts to diagnose breast cancer using SAXS. Key diagnostic models are identified and categorised into two groups, those based upon a physical model of collagen structure and those based upon more general purpose image processing techniques. The models based upon the more general purpose image processing techniques do not necessarily have a ready association with the structure of collagen, but they have allowed the use of more rapid and automated diagnostic techniques. Section 3.1 surveys the broad literature concerning the medical and biological research relating to x-ray scattering. Section 3.2 describes the models of Lewis *et al* (2000), Round (2006) and Sidhu *et al* (2008) that extract those classification features that are related to a physical model of collagen. These models all reduce the image data to one-dimension and frequently employ clustering techniques to produce diagnostic models. Section 3.3 describes the models of Butler *et al* (2003), Erickson (2005) and Falzon *et al* (2006) that use more abstract image analysis methods (such as the wavelet transformation) to extract useful classification features. These techniques then use the extracted image features to build diagnostic models. The merits, disadvantages and limitations of each approach are examined with constructive criticism provided. The chapter concludes in Section 3.4 with an overall summary of the lessons learned from the previous research and an assessment of where the diagnosis of breast cancer using SAXS is at and where it can head to from that point.

3.1 A survey of biomedical research using x-ray scattering

The characterisation of biological tissue using scattered x-rays has received considerable interest over the past thirty years. Investigations include applications in radiology (Johns & Yaffe 1983; Muntz *et al* 1983), computed tomography (Harding *et al* 1987; Schlomka, Schneider & Harding 2000; Griffiths *et al* 2003), bone tissue analysis (Royle & Speller 1991), liver cirrhosis and heptocellular carcinoma (Elshemeny *et al* 2003), brain and prostate tissue (Lazarev *et al* 2000; Siu *et al* 2005, Falzon *et al* 2007; De Felici *et al* 2008), breast tissue and breast cancer (Evans *et al* 1991; Peplow & Vergese 1998; Kidane *et al* 1999; Poletti, Gonçalves & Mazzaro 2002; Ryan & Farquharson 2004; Castro *et al* 2004; 2005b), technical imaging (Leclair & Johns 1998; 1999; 2001; Westmore *et al* 1996) and a range of other biological tissues and biomedical materials (Kosanetzky *et al* 1987; Speller & Horrocks 1991; Tartari *et al* 1997; 2002; Elshemey *et al* 1999; Harding & Schreiber 1999; Royle *et al* 1999; Batchelar & Cunningham 2002; Poletti *et al* 2002). The SAXS imaging technique is one branch within this field that has been used by a number of authors for cancer research. Investigations have examined the structure of collagen in normal, benign and malignant breast tissue (Lewis *et al* 2000; Fernàndez *et al* 2002; 2004; 2005; Suhonen *et al* 2005; Round 2006), investigated the utility of specific techniques for diagnosis (Round *et al* 2006; Changizi *et al* 2006; Pearson *et al* 2006; Wilkinson *et al* 2006; 2007; Ooi *et al* 2008) and developed or discussed diagnostic models (Butler *et al* 2003; Erickson 2005; Changizi *et al* 2005a; 2008; Falzon *et al* 2006; Rogers *et al* 2006; Ryan & Farquharson 2007; Sidhu *et al* 2008). Related image processing and modeling tasks have also been investigated for use as a component in a diagnostic system (Wilkinson *et al* 2006; 2007). The structure of brain tissue has also been investigated using SAXS imaging and diagnostic models have been developed (Siu *et al* 2005; Falzon *et al* 2007; De Felici *et al* 2008). The majority of these studies of considerable overlap and fall into more than one of the assigned categories above. In addition, at least two review articles have been published that include the biological and medical applications of SAXS imaging within their scope (Suortti & Thomlinson 2003; Theodorakou & Farquharson 2008). The diagnosis of breast cancer using SAXS patterns from hair samples has also been an intensive, yet controversial area of research having both advocates (James *et al* 1999; 2005; James 2001; 2003; 2004; 2006; Meyer *et al* 2000; Meyer & James 2001) and skeptics (Amenitsch *et al* 1999; Briki *et al* 1999; Chu *et al* 1999; Evans *et al* 2001; Howell *et al* 2000; Laaziri *et al* 2002). This thesis is focused on the development of a statistical model of SAXS patterns of breast tissue collagen, with one of the desired outcomes being improved diagnosis. The next two sections review those studies (Lewis *et al* 2000; Round 2006; Sidhu *et al* 2008; Butler *et al* 2003; Erickson 2005; Falzon *et al* 2006) that are deemed by the candidate to be most relevant in the development of a diagnostic model of breast cancer using SAXS patterns.

3.2 Diagnostic models based upon parameters related to the structure of collagen

Considerable effort has ben expended on the development of diagnostic models of breast cancer based upon parameters related to the physical model of collagen. This includes parameters such as the axial D-repeat (see Figure 2.2), the radius of gyration (a parameter related to the size and shape of the collagen fibril) and the levels of amorphous scatter that describe the breast tissue structure. Three notable studies include that of Lewis *et al* (2000) (Section 3.2.1) one of the initial studies into the diagnosis of breast cancer using SAXS, Round (2006) who confirmed the findings of Lewis *et al* (2000) (Section 3.2.2) and investigated a large range of other potential diagnostic features and that of Sidhu *et al* (2008) (Section 3.2.3) who further developed multi-variate classification models of breast cancer. These models are very useful because they can be interpreted in light of our current understanding of collagen structure, but they are still constrained by fundamental limitations. Wess (2005) highlighted the limited knowledge on the structure of collagen, which indicates that one should be cautious when developing diagnostic models of breast cancer. The SAXS breast cancer classification models developed by physical scientists impose our current scientific understanding of the problem and does not incorporate all of the information available in the image. Important diagnostic information in the image might be inadvertently discarded by focusing on models derived from the physical structure of collagen. Whilst the models may work for the application at hand (diagnosis of three breast tissue pathologies), they may or may not work for future applications such as the diagnosis of a greater range of breast cancer pathologies, the identification of the grade of malignant breast tissue and the diagnosis of other cancers (such as colo-rectal cancer). This section reviews the major works of three key authors in the field and justifies the claims that all aspects of the image data should be considered for analysis rather than just extracting those features which describe the current physical model for collagen.

3.2.1 Exploratory Analysis Based upon a Physical Model: Lewis *et al* 2000

3.2.1.1 Overview

Lewis *et al* (2000) proposed one of the first models to diagnose normal healthy, benign (fibroadenoma), malignant (invasive carcinoma) and mammoplasty (healthy tissue near the margins of a tumour) tissue using SAXS. The model was based on two features, 3rd order axial scattering peak position and the proportion of scattering intensity contained within all of the scattering peaks in the one-dimensional profile of the image. The 3rd order axial scattering peak position is related to the magnitude of the axial D-repeat structure of the collagen fibrils by Bragg's condition (equation 2.2), while the proportion of intensity contained within all the scattering peaks feature is related to both the structural integrity and the amount of collagen within the sample. These features were extracted from the data via a process known as *radial integration* which transforms a two-dimensional image into a one-dimensional vector. This process is illustrated in Figures 3.1(a) and (b).

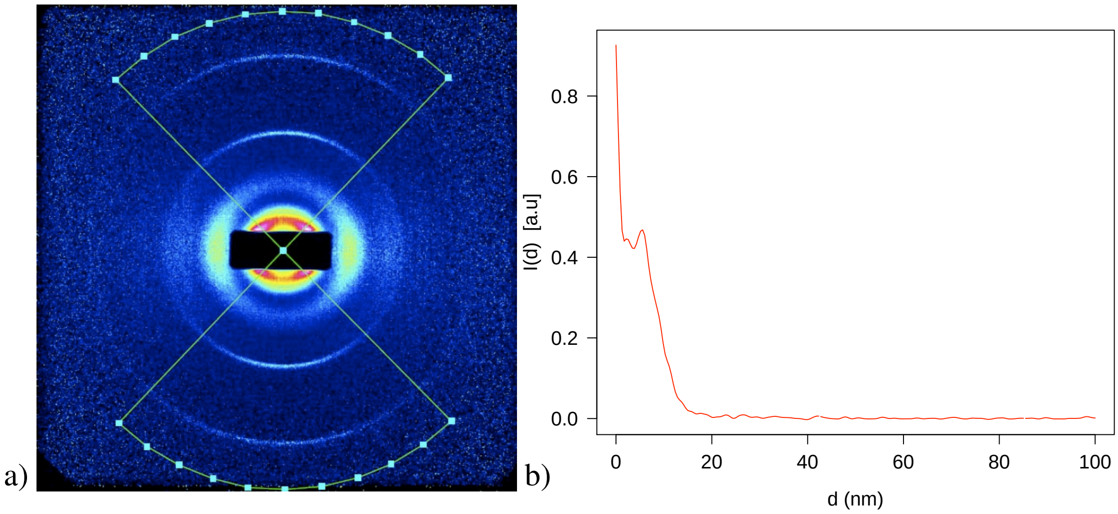

Figure 3.1: The radial integration technique that was applied to the SAXS patterns: (a) the image data contained within the image mask (green) is averaged to yield a function of intensity with respect to position in the image mask, (b) the resulting one-dimensional function $I(d)$.

The two-dimensional SAXS images were reduced to a one-dimensional function of intensity, $I(d)$ using the radial integration technique by summing and averaging the intensity as a function of object size (which can be calculated using the location of the scattering vector, $\mathbf{h}$ on the image). The $I(d)$ function describes the intensity of scattering as a function of object size, d, which can be determined from the SAXS images using Bragg's condition in equation (2.2). This technique is acceptable for SAXS images of collagen in breast tissue because they have a point of symmetry. Identification of this point of symmetry was performed manually using a graphical user interface (Lewis *et al* 2000). This might have resulted in some inaccuracies in the estimate of the $I(d)$ function. The parameters of the diagnostic model were extracted from these curves using a peak-fitting routine (Lewis *et al* 2000). Therefore, slight errors in the estimates of the $I(d)$ function result in inaccurate values of the diagnostic parameters. Exploratory data analysis allowed Lewis *et al* (2000) to identify the two features (3rd order axial scattering peak position and total fraction of scattering intensity in the peaks) that separated the data into distinct clusters. Tissue pathology was then associated with each cluster. The model was not developed any further and a statistical model was not specified. This makes it difficult to assess the accuracy of the technique in diagnosing future cases. Nonetheless the research reported in Lewis *et al* (2000) provided the framework from which other diagnostic models were developed.

3.2.1.2 Constructive Criticism

There were some important limitations to the research of Lewis *et al* (2000). These include:

a) Reduction of the image from two to one dimension.

b) Manual processing using the radial integration technique.

c) Using only a very limited amount of image information to separate the tissue groups.

d) Not developing a classification model.

The use of only two features for the separation of the tissue groups is of particular interest. At the first glance, the discovery of two features that allow tissue group identification appears to have solved the problem. On further consideration, the use of such a limited amount of image information is not so desirable. The SAXS images contain a wealth of information on the structure of breast tissue and greater separation might have been achieved using other features. Furthermore, there is no guarantee that these two features (3rd order axial scattering peak position and the fraction of total scattering intensity) will be useful for future tasks that are envisaged such as prediction of tumour grade or pathology. Good scientific practice demands a deeper exploration of the data to ensure any useful classification features are not missed.

3.2.2 Cluster & Principal Component Analysis: Round 2006

3.2.2.1 Overview

Round (2006) conducted a comprehensive investigation on the SAXS images of breast cancer tissue. A total of 225 SAXS images from 82 patients and 3 different tissue pathologies (corresponding to the labels used in this thesis of 'normal-healthy', 'benign' and 'malignant') were examined. The work followed on from the model of Lewis *et al* (2000) and investigated a range of topics including:

a) The effects of the experimental techniques on the SAXS images.

b) The effect of enzymatic (collagenase) degradation on the tissue structure.

c) Diagnosis of breast cancer using SAXS images corresponding to the structure of adipose tissue.

d) Diagnosis of breast cancer using SAXS images corresponding to the structure of collagen, as in the work of Lewis *et al* (2000).

e) Diagnosis of breast cancer using SAXS images obtained using a conventional (rather than a synchrotron) x-ray source.

f) The study of the changes in breast cancer collagen structure with increasing distance away from the centre of the tumour.

Round's (2006) analysis of the SAXS images corresponding to the collagen structure (item d) above) is of primary interest in this literature review. This analysis is comparable to the work of Lewis *et al* (2000) and corresponds to the images analysed in this thesis. Round (2006) radially integrated the images to produce a one-dimensional profile of intensity as a function of distance from the centre of the image. The distance from the centre of the image was then indexed to physical size of the structures under investigation to obtain the $I(d)$-curve. A model was then fit to these profiles by fitting an exponential trend and a series of Gaussian peaks. Features related to a physical model of tissue structure were extracted including the radius of gyration, relative intensity ratio (RIR) of the average intensity near the beam-stop (d = 230-429 nm) and the average intensity between the 3rd-order and the 5-th order axial scattering peaks (d = 42-58 nm), peak position, power law behaviour of the scattering profile (Round 2006), scattering peak area and, full-width half maximum of the scattering peaks. Two-tailed t-tests were applied to each univariate parameter (Table 10.3.2 Round (2006)) and potential features identified for further analysis using the magnitude of the p-value of the test. Exploratory data analysis was used to identify clusters corresponding to tissue pathology in bivariate scatter plots of the features. The position and peak area of the 3rd order axial peak displayed in Figure 3.2(a) provided a certain amount of separation of the three tissue groups, but there was a lot of overlap and many outliers that did not fit a particular cluster (Figure 10.3.4 Round (2006)). As displayed in Figure 3.2(b), limiting the comparisons to data collected within only a single experiment improved the clustering and separation considerably (Figure 10.3.5 Round (2006)). This result indicated that data collected from different experiments was difficult to compare directly. A certain amount of separation was also achieved using position and the peak area of the 2nd-order equatorial scattering peak (Figure 10.3.8 Round (2006)) . The scatterplot of the relative intensity ratio and the radius of gyration also produced very good separation of the tissue groups (Figure 10.3.9 Round (2006)) but only when malignant samples containing less then 80 $\%$ tumour volume in the sample where excluded in the study. Round (2006) manually partitioned the data in these scatterplots into regions corresponding to different tissue pathologies (Figure 10.3.7 Round (2006)). Principal component analysis (PCA) was also applied to the $I(d)$ profiles to develop a classifier with 85.1852 $\%$ accuracy when assessed using leave-one-out cross-validation. Similar techniques were applied to both the adipose tissue and conventional camera data.

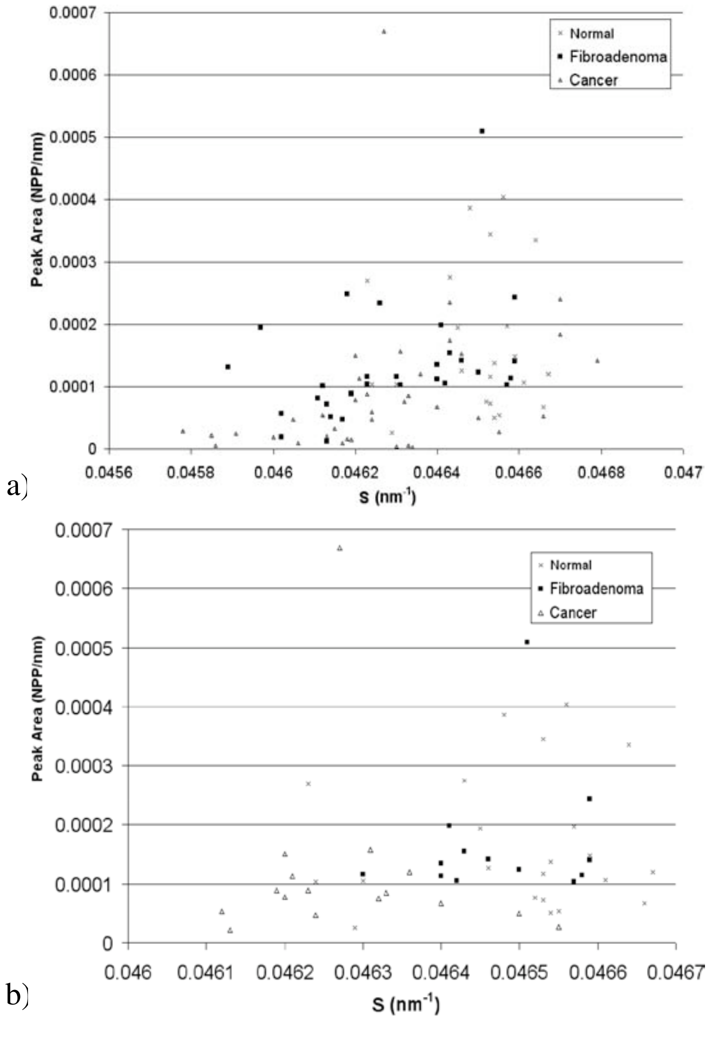

Figure 3.2: Exploratory data analysis plots of the 3rd-order axial scattering peak position and area for SAXS images of breast tissue: (a) collected across multiple experiments, (b) data from a single experiment. The parameter s, is defined by $s = n/d$, where the symbol d refers to the parameter used the describe collagen structure and the labels ‘normal’ corresponds to normal healthy breast tissue, ‘fibroadenoma’ to benign tumour tissue and ‘cancer’ to malignant breast tissue (reproduced from Figures 10.3.4 and 10.3.5 Round 2006). Note the separation of the data into clusters that are associated with tissue pathology.

3.2.2.2 Constructive Criticism of the work of Round (2006)

Round's (2006) research extended and identified the limitations of the Lewis *et al* (2000) model. The problems produced when comparing the data across experiments were identified. Many of the experimental techniques were refined and a much larger data set (than that of Lewis *et al* (2000)) was studied. The PCA classifier produced good results which were verified using leave-one-out cross-validation. There are several issues with the study that are worthy of further consideration. These include:

i) The model fit to the $I(d)$ curves smoothed the true profile.

ii) Feature extraction was a lengthy, manual process.

iii) A large number of statistical tests were used to identify features worthy of further investigation.

iv) The exploratory data analysis was biased to malignant samples that had greater than 80 $\%$ tumour volume

v) The diagnostic regions were identified manually.

vi) The PCA classification model was not fully specified.

Selected items from points (a)-(f) above are selected for further discussion.

3.2.2.3 Model Fitting

The radial integration technique reduced the data from a two-dimensional image into a one-dimensional curve. The model was fitted as a mixture of Gaussians (to describe the scattering peaks) and a quadratic trend of the form, $I(d) = \frac{(d-p_0)^2-p_1}{4p_2}$ (where p_0, p_1 and p_2 are constants) (Round 2006). The use of Gaussian profiles enforced an assumed amount of *smoothness* on the scattering peaks. The very high intensity scattering peaks might not have fit a Gaussian profile well and other techniques using triangular or wavelet basis functions might have been better suited to capture the structure of the scattering peaks. The potential loss of information by fitting a Gaussian profile to each scattering peak was not investigated. Key diagnostic information could have been lost by adopting this model fitting approach which smooths out the characteristics of the scattering peaks.

The precise mathematical model to describe the data was not specified and little information was provided on the model estimation routine. The total fit of the scattering profile was assessed using the r^2-statistic, unfortunately a high value of r^2 was interpreted as indicating a good fit of the model (Round 2006). This is a common misunderstanding, a high-value of r^2 does not necessarily indicate a good fit of the model to the data (Neter *et al* 1996). The r^2-statistic measures the fit of a linear model, but curves with quadratic or other non-linear trends can also produce a high r^2 value. The models fit by Round (2006) contain such non-linear trends and the use of the r^2-statistic could have been very misleading. More sophisticated alternatives to assess model fit (such as the χ^2 or likelihood ratio tests) could have been applied to provide more robust assessments.

3.2.2.4 Statistical Tests for Feature Identification

Promising diagnostic features were identified with a barrage of two-tailed t-tests. Round (2006) did not assess the adequacy of these tests to the data, the parametric assumptions were not assessed nor were the data screened for potential outliers (which are evident in the plots of Figures 3.2(a)-(b)). Furthermore, no consideration was given to the false discovery rate of the tests and the combined effects of the features was not considered. An alternative strategy would have been to perform exploratory data analysis, identify potential classification features using scatterplots and then develop and refine a classification model. Statistical inference could have then been performed using this model.

3.2.2.5 Biased Analysis

Round (2006) developed classification models using a subset of the ‘malignant’ group. This subset was defined as those samples that contained more than 80$\%$ of tissue volume corresponding to malignant tumour tissue. Whilst this approach is reasonable for the development of initial models it is not for the full development of a classification model. In practice it is important that malignant tissue is detected even if it corresponds to 1$\%$ of the volume of the sample, 80 $\%$ of tumour volume per sample is a very arbitrary level. Samples encountered in clinical practice could contain any proportion of tumour tissue. Selecting only the most suitable samples for classification creates an unbalanced view of reality and wastes precious samples.

3.2.2.6 The Scatterplot Classification Model

Round (2006) developed a diagnostic model based upon the the position and area of the 3rd order axial scattering peak (as plotted in Figure 3.2(b)). The decision regions in the scatterplots of the features were found empirically by determining the maximum number of correctly labeled data points ('normal-healthy', 'benign', 'malignant') in each region. A *specificity* of 81.8 $\%$ and a *sensitivity* of 100 $\%$ was reported (Round *et al* 2005). A statistical model was not used to define these regions and the errors rates were not assessed with cross-validation. The sensitivity and specificity rates appear to be estimated from the training data, it would have been more appropriate to estimate these errors using independent test data.

3.2.2.7 PCA classification model

The PCA classification model used by Round (2006) was poorly described, making it hard to assess what was actually done. Round (2006) presumably fitted a spline function to the $I(d)$ curve for each sample, performed PCA on the coefficients of these splines, followed by linear discriminant analysis on the eigenvectors of the PCA model. Twenty-five principal components and twenty-five linear discriminants were identified but specific equations were not provided. The PCA model was probably fit in a 25-dimensional feature space, yet only 108 samples (22 normal healthy, 30 benign and 56 malignant) were used and at most only 82 samples could have been independent. Inter-sample correlation was not accounted for, which might have produced optimistic cross-validation results. Furthermore, sparse amounts of data in a high-dimensional feature space can lead to substantial problems. Ripley (1996) states that N patterns *randomly* selected from any continuous distribution in $\mathcal{R}^p$ (where $p \,\epsilon\, \mathbb{R}$ is the dimension of the feature space) can be randomly divided into two groups with probability,

$$2^{1-N} \sum_{i=0}^{\min(N-1\,,p)} \left(\begin{array}{c} N-1 \\ i \end{array} \right) \sim \Phi\left(\frac{2p-N}{\sqrt{N}}\right) \tag{3.1}$$

for large N and where Φ is the standard normal cumulative density function (page 119, Ripley 1996). Equation (3.1) means that good classification results might just have been a result of insufficient data for the number of features used. According to equation (3.1), the 'normal healthy'-'benign' contrast had a probability that a linear partition could have been found that just happened to separate the two groups *perfectly* of 0.5. This probability is unacceptably high and a meaningful linear discriminant may not have been found.

3.2.3 Multivariate Analysis: Sidhu *et al* 2008

3.2.3.1 Overview

Sidhu *et al* (2008) also conducted a large study on the diagnosis of breast cancer using SAXS image data. Samples from a total of 80 patients from normal healthy, benign, malignant and mammoplasty breast tissue were analysed. Radial integration was used to obtain one-dimensional profiles $I(d)$ from the two-dimensional SAXS images. The data were detrended using a quadratic term that was a function of object size, $q \propto d$. Gaussian peaks were also fit using a method similar to that used by Lewis *et al* (2000) and Round (2006). The final curves are displayed in Figure 3.3(a) for selected samples, 95% confidence intervals are displayed (shaded) around the average (solid line) where possible (Sidhu *et al* 2008). Features extracted from profiles included the position, area, full-width half-maximum and the amplitude of the axial and Bessel scattering peaks as well as the parameters of a quadratic curve fit to the scattering profiles which included the total area under the curve (the amorphous scatter feature). The axial D-repeat of the structure of collagen fibrils was then calculated from these features. The azimuthal distribution of the scattering rings was also extracted, which is defined as the angular extent of the scattering rings as measured from the centre of the image. Sidhu *et al* (2008) investigated two different tissue scattering techniques, developed regression models to understand the relationships between different physical parameters and created classification models. The classification models are the main part of Sidhu's *et al* (2008) work that is of interest in this review.

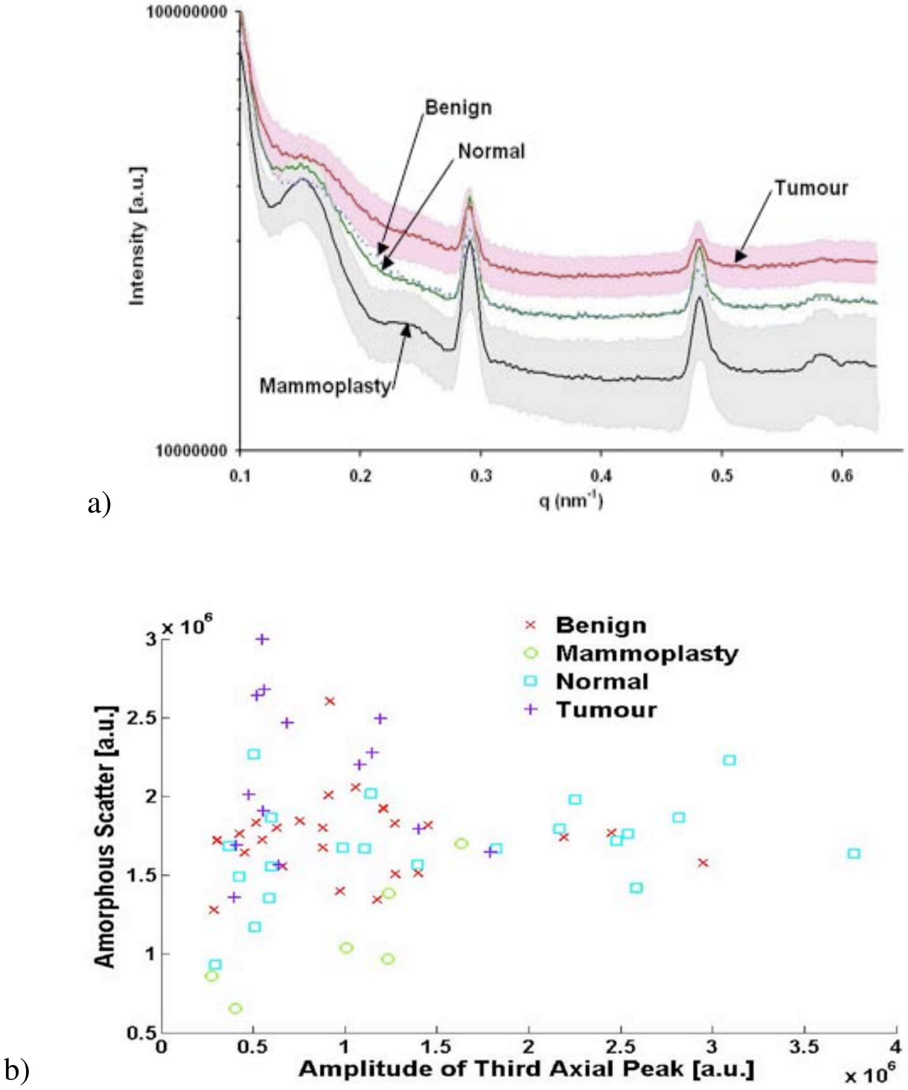

Figure 3.3: Analysis of SAXS images of breast tissue using the Sidhu *et al* (2008) model: (a) Intensity, $I(q)$, profiles of the four tissue groups (normal healthy, mammoplasty, benign and tumour (malignant)) with the 95$\%$ confidence intervals (shaded), (b) the scatterplot of the 3rd-order axial scattering peak amplitude and the amorphous scatter features separates some of the tissue contrasts (Sidhu *et al* 2008).

Significant differences (p $<<$ 0.01) were identified in the intercept component, C of the quadratic curve fit to the scattering profiles $I(q)$ (Sidhu *et al* 2008). The greatest differences appeared to be between the mammoplasty and malignant groups and suggested that this feature might be useful for incorporation in a classification model of tissue group (Sidhu *et al* 2008). Significant differences between tissue groups were also identified for the amplitude of the 3rd order axial scattering peak feature. Figure 3.3(b) displays the scatterplot of the third-order axial scattering peak and the amorphous scatter features with respect to each tissue group. The benign, tumour (malignant) and mammoplasty groups appear to be separated on the amorphous scatter feature, while some of the normal tissue samples are identified using a combination of both features. A quadratic discriminant analysis classification model was fit using these two features and was very effective at separating the mammoplasty and malignant tissue groups (Sidhu *et al* 2008). Detailed cross-valiated classification results are reported in Table 3.1, excellent separation is achieved for some contrasts (malignant-mammoplasty) but not with others (benign-normal) (Sidhu *et al* 2008).

Contrast	**Sensitivity**	**Specificity**	**Accuracy**
Benign-Mammoplasty	95.8	71.4	90.3
Benign-Normal	83.3	47.6	66.7
Normal-Mammoplasty	95.2	57.1	85.7
Malignant-Benign	50.0	87.5	73.7
Malignant-Mammoplasty	92.9	100.0	95.2
Malignant-Normal	64.3	85.7	77.1

Table 3.1: Quadratic discriminant analysis classification results
(From Table 2, Sidhu *et al* 2008).

3.2.3.2 Constructive Criticism of the work of Sidhu *et al* (2008)

Sidhu *et al* (2008) conducted a comprehensive investigation on the diagnosis of breast cancer using SAXS imaging. A large number of features were explored and improvements made to the SAXS imaging technique. Changes in the relationship between the 2nd-order equatorial (Bessel) scattering peak area and amplitude were also observed for the diseased (benign and malignant) tissue groups. A challenging four group classification was attempted, with good results in some cases. Sidhu *et al* (2008) recognised that the width (variability around the mean) of the scattering peaks contained important diagnostic information. Nonetheless there are some limitations to the model proposed by Sidhu *et al* (2008). The central issue is the reduction of the SAXS image from two to a one dimensions, followed by the smoothing of the $I(d)$ profile by fitting Gaussian peaks. The features derived are very application specific and can be very labour-intensive and time-consuming to extract. The feature extraction techniques used by Sidhu *et al* (2008) are subject to inaccuracies, but this was not taken into account in the analysis. Similar to the approach of Round (2006), a number of statistical tests were applied, but in this case after careful inspection using exploratory data analysis. Analysis of variance model assumptions were very carefully checked, the data was screened for potential outliers and robust statistical methods were used where possible. The main issue is not so much a problem with the analysis performed by Sidhu *et al* (2008) but the desire for an improved methodology in the analysis of SAXS images.

3.2.4 Section Summary of the Diagnostic models related to Collagen Structure

Breast cancer diagnosis using features related to the structure of collagen has been throughly explored by researchers in the SAXS imaging community. This approach has yielded insights into how the structure of collagen changes in the malignant breast tissue. The reduction of the image to a one-dimensionl vector, followed by smoothing by fitting a linear mixture of Gaussians and a quadratic trend has distinct disadvantages. Spatial information and key spectral (and possibly fractal) characteristics might be lost in this process. A wide range of features have been proposed and little consensus has been reached among authors on the best features to use. The structure of collagen in breast tissue not yet fully understood, which limits the interpretation of some of these features. The feature extraction process is user-intensive and has inherent inaccuracy. In general, this inaccuracy has not been accounted for in the classification models that have been proposed. Good classification results have been achieved and valuable scientific understanding has been provided but this approach might be superseded by more automated and powerful mathematical methods dedicated to the task of classifying breast cancer using the SAXS images.

3.3 Diagnostic Models of SAXS Images of Breast Cancer Using Image Analysis Techniques

Alternative approaches exist to the diagnosis of breast cancer using SAXS imaging that are based upon statistical and image analysis methodology. The wide range of possible methods that could be implemented includes stochastic (random field), fractal and functional models. Two main methods have been used to date, data mining and multi-scale analysis using the wavelet transform. This section will review and critique the three main works that use image analysis techniques on SAXS images of breast cancer. The first study by Butler *et al* (2003) used data mining to identify features useful for classification of the three breast tissue groups. In contrast, Erickson (2005) used the discrete wavelet transform to develop similar classification models. The work of Falzon *et al* (2006) extended the wavelet technique by extracting and identifying diagnostic features in the transform coefficients related to the axial scattering rings in the SAXS images. All three approaches allowed the identification of features that were used in rapid, semi-automated diagnostic algorithms that produced very good classification results. The features appeared to focus on different aspects of the images and subsequently extract different information than those parameters related to a physical model of collagen (refer to Section 3.2). Interpretation of these features is challenging and in most cases it is difficult to understand how they relate to the structure of collagen. Regardless of these challenges, the results reported in these studies provide a strong case for the use of statistical and image analysis techniques in the diagnosis of breast cancer using the SAXS technique.

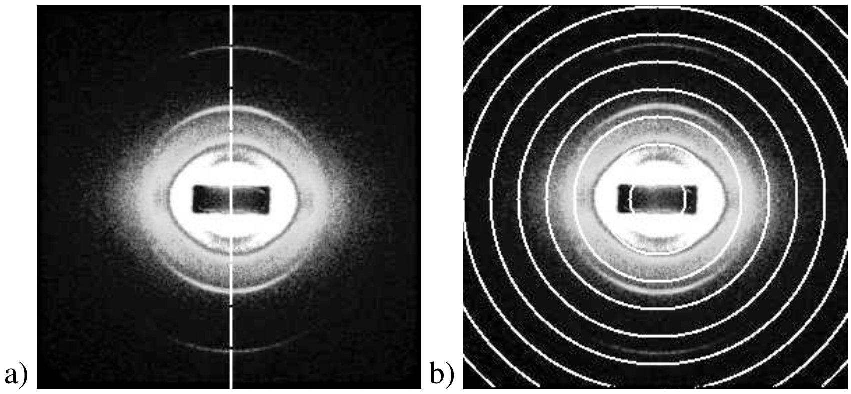

Figure 3.4: Feature Extraction method used in the analysis of SAXS images by Butler *et al* (2003): (a) the vertical slice and (b) radial segment feature extractors as indicated by white lines (Figure 2 Butler *et al* 2003).

3.3.1 Data Mining: Butler *et al* 2003

3.3.1.1 Overview

Butler *et al* (2003) applied data mining techniques to search for features in the SAXS images that were useful for automated classification. The SAXS images of 20 normal healthy and 22 malignant breast tumour samples were examined using data mining methods. Over 100 features were proposed and explored. These features were based upon information extracted along a vertical slice through or in radial segments out from the centre of the image. Figures 3.4 (a) and (b) display the regions of the image used to extract the vertical slice and radial features.

Features extracted included:

a) Total intensity of the scattering image (the sum of all pixel values in the image).

b) The maximum intensity in each of 5 or 10 equally sized divisions along a nine-pixel moving average smooth of the vertical slice of the image.

c) The y-coordinate of the maximum intensity in each of these divisions along the smoothed vertical slice of the image.

d) The sum of the intensity in each of 5 or 10 radial segments of the image.

e) The maximum intensity magnitude in each of these radial segments.

f) The radial distance and the x- and y- coordinates of the intensity maximum in each of the circular regions.

A naive Bayes'ian classification model combined with the equal frequency discretisation method was used to assess the classification potential of each feature (Butler *et al* 2003). The most promising features were combined and assessed further using leave-one-out cross-validation. The estimates of the accuracy of these models range from 97.62 % for a combination of vertical slice and radial segment features to 45.24 % for the sum of the whole image intensity feature (Butler *et al* 2003). These results are very impressive and motivated further research into the automated diagnosis of breast cancer using SAXS images. The models proposed put the diagnosis of breast cancer using SAXS images in a statistical modelling/machine learning framework.

3.3.1.2 Constructive Criticism of the work of Butler *et al* (2003)

Despite the positive contributions of the work of Butler *et al* (2003), there is room for constructive criticism. Specific issues include:

i) Producing a model that is too dependent on the data set at hand (*over-fitting*).

ii) Arbitrary partitioning of the regions in the SAXS image used to extract the features.

iii) Uncertainty in the estimation of feature values.

iv) Smoothing of the intensity profiles of the features extracted from the vertical transects.

3.3.1.3 Model Over-fitting

The cross-validated model results reported by Butler *et al* (2003) might be over-optimistic because the models might have been over-fit to the data. The most accurate classification model had an accuracy of 97.62 % (Table 2, Butler *et al* 2003). This model contained 70 features but used only 42 samples (20 normal healthy, 22 malignant) in the fitting process. Therefore, each data point could have been associated with at least one feature. An alternative model had the same accuracy (97.62 %), but used 60 features (Table 2, Butler *et al* 2003). In fact all models that had greater than 95% accuracy were fit with between 5-105 features (Tables 1-4, Butler *et al* 2003). An accuracy of 97.62 % for 42 samples corresponds to exactly one sample misclassified when the model was assessed using leave-one out cross-validation. It seems very likely that the one sample held out from the model fitting process was incorrectly classified, which strongly suggests model over-fit. Multiple runs of the cross-validation algorithm, combined with random selection of the samples to 'hold-out' would have allowed the detection of this problem.

3.3.1.4 Feature Extraction

Little explanation was provided for the use of the features incorporated into the classification models by Butler *et al* (2003). The set of features studied was not exhaustive and a range of other features could have been proposed. For instance, a horizontal transect could have been used to extract features related to the packing of the fibrils ('D'). Rigorous criterion to determine the optimum number of partitions of the image was not provided. Too few partitions could have missed important diagnostic information, while too many partitions would have captured many random fluctuations in image intensity. An adaptive image partition method would have been more appropriate, allowing the detection and separation of finely spaced features while having larger partitions in regions that were more homogenous.

3.3.1.5 Feature Estimation

Several of the features used by Butler *et al* (2003) require precise determination of the SAXS pattern centre based upon the data provided by the digital SAXS image. The centre of the SAXS pattern was often difficult to determine with precision because it did not always coincide with the centre of the digital image (Butler *et al* 2003). Both the vertical slice and radial segment features were dependent on the accurate estimation of pattern centre and inaccuracies in this estimate would have resulted in poor estimates of feature magnitude. This error could have also lead to instability in some estimates. For instance an error in the estimate of pattern centre could have resulted in poor partitioning of the image. Two relevant local maxima (corresponding to scattering rings) could have been assigned to one segment, while only noise was contained in a neighbouring segment. Features based upon the position of the local maximum in each segment would ignore one of the relevant scattering rings and incorrectly extract noise as a diagnostic feature. Butler *et al* (2003) did not account for the uncertainty in the estimates of the feature values nor did they account for the 'false-discovery' of misleading features. Models that incorporate Bayes'ian statistics might have been useful in both cases.

3.3.1.6 Smoothing of Intensity Profiles

A moving average filter was used to smooth data extracted from the vertical transects before further analysis (Butler *et al* 2003). No justification for this approach was provided, but it might have been performed to reduce image noise thereby assisting in the detection of local maxima. Little consideration was given as to how the smoothing operation could have biased the results of the study. The very sharp, narrow, high intensity scattering peaks present in the vertical intensity transects could have been completely smoothed out. Even if the size of the moving average filter (9 pixels) was appropriate in this case, there is no guarantee that it will be for other cases. A range of other techniques exist, such as local, wavelet and kernel regression that are adaptive to the structure of the data set (Loader 1999; Ruppert, Hand & Carroll 2003; Vidakovic 1999; Wand 1995; Simonoff 1996). These techniques could have been explored and would have been more useful (than a fixed size filter) in preserving the structure of the transect.

3.3.2 Automated Feature Extraction using Wavelets: Erickson 2005

3.3.2.1 Overview

Erickson (2005) investigated the diagnosis of breast cancer using features extracted from the discrete wavelet transform of SAXS images. An accurate diagnostic model (perfect classification in some cases) was developed by examining the information present in the images over different scales of resolution (Erickson 2005). A total of 49 SAXS images of normal, benign and malignant breast tissue (20 normal healthy-labeled 'N', 7 benign-labeled 'F' and 22 malignant-labeled 'T' in the original work) were decomposed using the separable two-dimensional discrete wavelet transform into a set of coefficient matrices that described each image at different scales (resolutions) and directions.

The two-dimensional separable discrete wavelet transform is defined for a square digital image, $I(x, y)$, for a total of J resolution bands as,

$$I(x,y) = \sum_{k_1=1}^{n_J} \sum_{k_2=1}^{n_J} c_{J,k_1,k_2} \phi_{J,k_1,k_2}(x,y) + \sum_{l=1}^{3} \sum_{j=1}^{J} \sum_{k_1=1}^{n_j} \sum_{k_2=1}^{n_j} d^l_{j,k_1,k_2} \psi^l_{j,k_1,k_2}(x,y) \tag{3.2}$$

where n_1 is the number of rows of the image $I(x, y)$ being analysed and the number of rows and columns of the wavelet coefficient matrices at each resolution level is given by $n_j = \frac{n_1}{2^j}$ (adapted from page 157 Vidakovic 1999). The discrete wavelet transform described the image in terms of a set of coefficient matrices $\{\mathbf{C}_{J,k_1,k_2}, \mathbf{D}^1_{j,k_1,k_2}, \mathbf{D}^2_{j,k_1,k_2}, \mathbf{D}^3_{j,k_1,k_2}\}$ indexed by scale parameter $j = 1, 2, \ldots, J$, direction index $l = 1, 2, 3$ and location parameters (k_1, k_2). The direction indices are correspond to the $l = 1$ for vertical, $l = 2$ for horizontal and $l = 3$ for diagonal directions. The coefficients were found by projecting the image onto the basis functions, $\{\phi_{J,k_1,k_2}(x, y), \psi^1_{j,k_1,k_2}(x, y), \psi^2_{j,k_1,k_2}(x, y), \psi^3_{j,k_1,k_2}(x, y)\}$ using the inner product. In practice, the coefficients were calculated rapidly using a digital filter bank. (page 310-311 Mallat 1998). Selection of a particular set of filter bank coefficients corresponds to specifying a particular basis for the wavelet transform. The Haar, Daubechies (Db) and biorthogonal (Bior) bases were used in the discrete wavelet transform (equation 3.2) of pre-processed versions of the SAXS images (Erickson 2005). This pre-processing step involved the application of several binary circular masks of radii 50, 100, 150, 190 and 215 pixels respectively. These masks were designed to supress particular regions of the image (by setting them to zero) whilst preserving other regions (by leaving these regions unchanged).

Following the application of the discrete wavelet transform, a range of features were proposed including mean value ($\overline{\mathbf{D}}^{l}_{j,k_1,k_2}$), the standard deviation ($\sigma(\mathbf{D}^{l}_{j,k_1,k_2})$) and the energy of the power spectrum (A_l) of each coefficient matrix of each of the coefficient matrices $\mathbf{D}^{l}_{j,k_1,k_2}$, $(l \equiv v,\ h,\ d)$, as well as a feature based upon the sum of the intensities in the upper quadrant of the SAXS image (Erickson 2005). This feature was found by summing the sums of the intensity values along each x and y co-ordinates in the upper-quadrant of the image, that is, $F = (\sum_{i=1}^{N} I(x,y)_{x,y\,\epsilon\,\mathcal{O}})$ where $\mathcal{O}$ is a set specifying the N pairs of x and y indices of the intensity magnitudes to be summed in the calculation of this feature.

A preliminary experiment was performed to determine which one of these four features should be examined further (Erickson 2005). The wavelet basis, mask, tissue group contrast and combination of features were all constrained (the particular constraints were a biorthognal 3.7 filter, no mask, 'normal healthy'-'malignant' samples only), no optimisation of features used was allowed. A naive Bayes' classifier was fit to a range of combinations of mask, filter and classifier, with the leave-one-out cross-validated error rate used as an assessment of model fit. The classification rates of each feature were compared and the most accurate model was used to select the feature '*type*' for further study. The mean intensity of the wavelet coefficients ($\overline{\mathbf{D}}^{l}_{j,k_1,k_2}$) provided the greatest accuracy (98$\%$) and all subsequent analyses only considered this '*type*' of feature (Erickson 2005).

All combinations of wavelet filter and mask were then considered subject to the constraint that only samples from the 'normal healthy'-'malignant' groups were compared. Several basis-mask combinations were short-listed for further analysis based upon their classification performance. Similar experiments to determine the optimum wavelet basis-mask combinations were performed for the 'normal-healthy'-'benign' and 'benign-malignant' tissue group contrasts, again several filter-mask combinations were short-listed based upon their classification performance. The classification rates of the most promising basis-mask combinations were then compared for all three pairs of contrasts, 'normal healthy'-'malignant', 'normal-healthy'-'benign' and 'benign-malignant'. Three combinations were very competitive: (Daubechies wavelet Db4-50 pixel radius mask, Bior3.7-100 pixel radius mask and the Bior3.7-190 pixel radius mask) (Erickson 2005). The three-group classification rates were 90 $\%$ for the Db4-50 pixel, 88 $\%$ for the Bior3.7-100 pixel, 86$\%$ for the Bior3.7-190 pixel combination (Erickson 2005). A total of four features was used for Db4-50 pixel, five features for the Bior3.7-100 pixel and one for the Bior3.7-190 pixel models (Erickson 2005). In the final assessment the Bior3.7-190 pixel model based upon the $\overline{\mathbf{D}}^{d}_{7,k_1,k_2}$ feature was used because it minimised the number of features used in the classification model. The Db4-50 pixel model using the $\{\overline{\mathbf{D}}^{v}_{1,k_1,k_2}, \overline{\mathbf{D}}^{v}_{2,k_1,k_2}, \overline{\mathbf{D}}^{v}_{6,k_1,k_2}, \overline{\mathbf{D}}^{v}_{7,k_1,k_2}\}$ features was also recommended when the optimum diagnosis of samples belonging to the benign group was of paramount concern (Erickson 2005).

Despite the very good classification results of Erickson's (2005) study there are a number of issues worthy of constructive criticism:

a) There was a lack of a proper statistical design of the classification experiments.

b) Only a limited amount of information in each image was used for classification.

c) Model features were poorly specified.

d) Two group classification was used as a proxy for three-group classification performance.

e) A penalty was not used when comparing models with a different number of parameters.

f) The variability of the model accuracy estimates was not accounted for when comparing models.

A selection of the above points (a)-(f) is elaborated on in further detail.

3.3.2.2 Lack of Statistical Design

The model selection process in Erickson's (2005) work was empirical, a statistical experiment design was not used. A very large number of models were fit in Erickson's (2005) study and a carefully designed experiment might have reduced the number of models fit or provided a more robust understanding of the treatment (features, bases, masks) effects. The preliminary study conducted by Erickson (2005) in order to determine the '*type*' of feature to study is not convincing. No consideration was given to possibility that the cross-validated model results might be influenced by the constraints put in place. Erickson (2005) fixed a basis, the biorthogonal 3.7 basis and concentrated on the 'normal healthy-malignant' contrast. Later experiments focused on models designed to provide the best basis-mask combination. It was no surprise that the biorthogonal 3.7 basis featured in the top three models found and that best classification results were achieved for the 'normal healthy-malignant' contrast. A well planned statistical experimental design could have been used and might have negated the need for a preliminary experiment (Atkinson, Donev & Tobias 2007).

3.3.2.3 Poor Feature Specification

Erickson (2005) defined new diagnostic features, some of which may have been a poor choice. The mean of the wavelet coefficient matrix feature, $\overline{\mathbf{D}}^{l}_{j,k_1,k_2}$, was designed to describe the '*texture*' of each wavelet coefficient matrix (page 55, Erickson 2005). The notion of 'texture' was not defined and it can only be assumed that Erickson (2005) was referring to texture as the higher-order (2-4) moments of the probability density function of the intensity values in the image (pages 596-597, Rangayyan 2005). It is hard to understand how the mean value of the wavelet coefficients could be used to capture these higher-order correlations. In fact, it is surprising that this feature is useful for classification at all. The density of wavelet coefficients of an image within a band (matrix) of the transform is often symmetric and centered on zero (Simoncelli 1999). Therefore it is very surprising that differences were detected in each tissue group using this feature. Numerical or graphical summaries of the magnitude of each feature were not provided. The data used by Erickson (2005) was available and the confidence intervals of the wavelet coefficient energy statistics for the Db4 basis were calculated. Table 3.2 displays the estimates of the 95 $\%$ confidence intervals of the wavelet energy features for the three highest resolution levels of the transform, lower resolution levels were also found to be centered around zero. Inspecting these intervals across tissue groups for each feature suggests little if any diagnostic information is contained within the mean of the wavelet coefficient feature. Based upon these results it seems unlikely that accurate classification would be achieved when the model based upon the wavelet mean energy feature is applied to independent test data.

Feature	Normal	Benign	Malignant
$\overline{\mathbf{D}}_8^v$	[0.08,1.26]	[-0.34, 1.82]	[0.08, 1.30]
$\overline{\mathbf{D}}_8^h$	[-1.05, 1.32]	[-1.42, 2.19]	[-1.19, 1.49]
$\overline{\mathbf{D}}_8^d$	[-0.25, 0.25]	[-0.23, 0.20]	[-0.27, 0.22]
$\overline{\mathbf{D}}_7^v$	[-0.92, 1.85]	[-0.84, 2.38]	[-0.07, 3.26]
$\overline{\mathbf{D}}_7^h$	[-5.13, 3.64]	[-9.06, 7.27]	[-4.44, 2.24]
$\overline{\mathbf{D}}_7^d$	[-0.47, 0.53]	[-0.46, 0.10]	[-0.65, 0.93]
$\overline{\mathbf{D}}_6^v$	[-5.01, 3.32]	[-1.97, 0.72]	[-1.85, 5.42]
$\overline{\mathbf{D}}_7^h$	[-8.37, 4.65]	[-4.09, 3.00]	[-8.56, 2.14]
$\overline{\mathbf{D}}_6^d$	[-1.66, 0.80]	[-1.30, 0.42]	[-1.64, 0.78]

Table 3.2: 95 $\%$ confidence intervals of SAXS image mean wavelet coefficient features for scales j = (6,7,8) as determined using the same data as Erickson (2005) and the Daubechies wavelet basis with four vanishing moments.

The standard deviation of the wavelet coefficients, $\sigma(\mathbf{D}^l_{j,k_1,k_2})$, was designed by Erickson (2005) to represent the '*surface energy*' of each matrix, but again the term surface energy was not defined (page 55, Ericskon 2005). The notion of surface energy may refer to the integral of a two-dimensional function $E = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\,dx\,dy$, which in practice is approximated by the discrete sum of the squares of the pixels or coefficients (see page 10 and page 14 Mix & Olejniczak (2003) for a discussion of energy in relation to one-dimensional signals) . For the wavelet coefficient matrices, the surface energy would then be defined as $E^l_j = \sum_{k_1=1}^{N}\sum_{k_2=1}^{N}(\mathbf{D}^l_{j,k_1,k_2})^2$ for fixed fixed scale, j and direction parameter, l. The standard deviation $\sigma(\mathbf{D}_{j,k_1,k_2})$ of the wavelet coefficients feature clearly does not match well with the more conventional definition of surface energy E^l_j. Unfortunately, Erickson (2005) did not explain how the standard deviation of the wavelet coefficients can capture *surface energy* in sufficient detail for further comment. It is difficult to understand how the standard deviation of the wavelet coefficients can accurately describe the data. The literature states that the density of the wavelet coefficients within a scale and direction resembles a Laplace distribution (Simoncelli 1999). This distribution has a sharp central spike and very long tail. In general the Laplace distribution has non-zero higher-order $(n > 3)$ cumulants, yet the standard deviation only describes the second-order statistics of the data. Clearly, the standard deviation feature is insufficient to capture all of the information contained within the wavelet coefficients.

The area under the power spectrum of the wavelet coefficient matrix feature, A^l_j, was designed to represent the spatial frequency information in the wavelet coefficient matrices (page 56 Erickson 2005). The mapping from the power spectrum to the feature value is not unique. Two completely different power spectra might have the same feature value. For instance, imagine two power spectra that consist of a series of 4 vertical lines to the left and to the right of the zero frequency $\{K_x, K_y = (0, 0)\}$. The first power spectrum contains vertical lines that increase in magnitude away from the centre, describing a image that has more dominant higher spatial frequency information. While the second power spectrum has vertical lines that decrease in magnitude away from the centre and thus has more dominant lower spatial frequency information. Both of these spectra will have the same 'area' feature value, yet they differ considerably in the spatial frequency information that they describe.

The sum of the intensities feature was used to compress but still represent the information in the SAXS images (page 59, Erickson 2005). This feature was used as a benchmark to compare the performance of the wavelet classification algorithms. Again the features do not correspond to a unique arrangement of the SAXS images which is undesirable for general purpose applications.

3.3.2.4 Classification Bias

The models in Erickson's (2005) research were developed with the 'normal healthy-malignant' contrast in mind, priority was given to features that separated these two tissue groups at the cost of the others. The importance of separating these two groups is well-founded but the 'benign-malignant' contrast seems equally if not more important. Erickson (2005) assessed model performance using two group contrasts assuming that this would translate to good representations of three group contrasts. This is not necessarily true, for instance a feature set could have been very good at separating 'normal healthy-malignant' groups but have been very poor at separating the 'normal healthy-benign' groups. The summary of classification results in Table 5.7, page 81 Erickson (2005) might indicate this problem. The Bior-190 pixel radius mask model had perfect classification of the normal and malignant groups, but failed to classify a single benign sample. Four out of the seven benign samples were classified as malignant and three as normal healthy. Similar problems exist for the other two competing models, for instance the Daub4-50 pixel radius mask model classified only two benign samples correctly. Models examining all three groups should have been compared or at the least a model should have been constructed by selecting features designed to provide optimum results for each contrast.

3.3.2.5 Variability in the Estimates of Accuracy

Leave-one out cross-validation produces estimates of model accuracy that are biased and have a large variance (pages 215-216, Hastie, Tibshirani & Friedman 2001). This problem did not appear to factor into Erickson's (2005) comparison of competing models. For instance in the preliminary experiment designed to select feature 'type', the mean of the wavelet coefficient matrix features produced a model with 98 % accuracy (Erickson 2005). In contrast, the area under the power spectrum of the wavelet coefficient matrix feature produced a model with 93 % accuracy (Erickson 2005). Therefore, the mean of the coefficients model classified two more samples correctly than the model based upon the power spectrum feature (the 42 samples result in a resolution of approximately 2 % per sample). There are a finite number of samples in the data, so that even if a perfect estimate of accuracy was found the estimates could only be determined to be within the intervals of 96-100 % and 91-95 % accuracy respectively. Incorporating the additional uncertainty associated with the leave-one out estimates would have made evidence of a difference between the models even less likely. Calculation of confidence intervals of model accuracy that account for all forms of uncertainty in the estimate would have been far more appropriate and would have provided a greater understanding of the performance of the competing models.

3.3.3 Generalised Linear Models & Wavelet Coefficient Energies: Falzon *et al* 2006

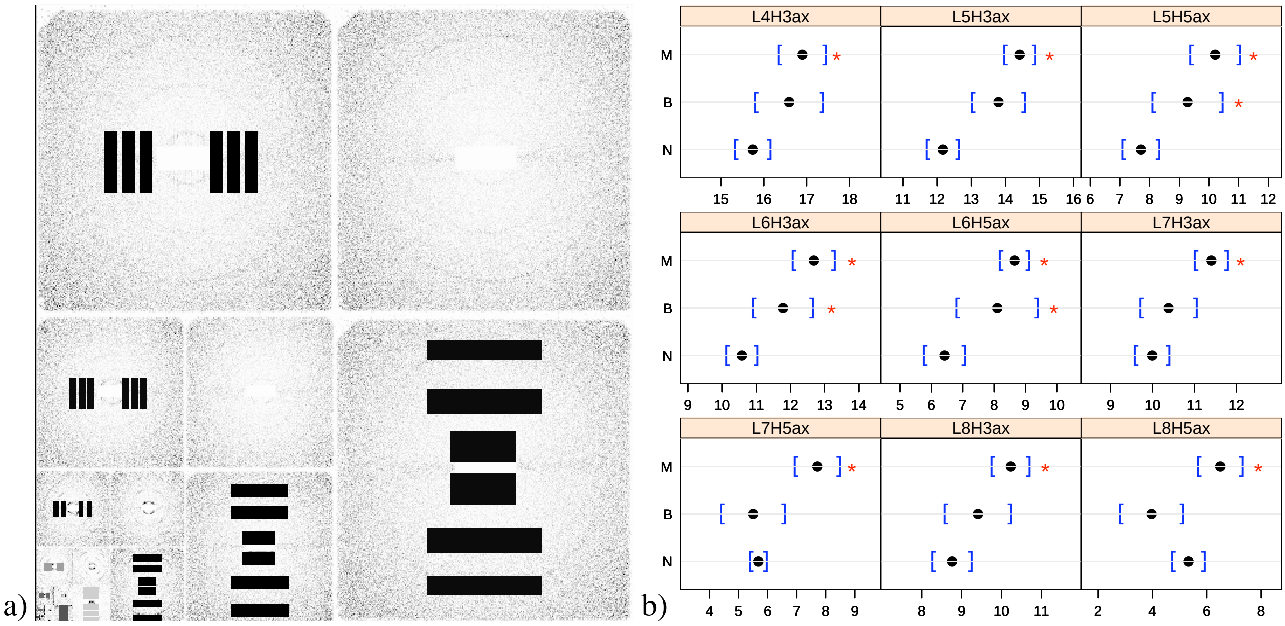

Figure 3.5: Analysis of Wavelet Coefficient Energies of SAXS Images: (a) Regions (black rectangles) extracted from the wavelet coefficient matrices (b) Confidence intervals of the features with respect the three tissue groups (Falzon *et al* 2006).

Falzon *et al* (2006) investigated the use of the wavelet transform of SAXS images further to diagnose breast cancer. The objective of the study was to determine the key regions in the wavelet coefficient matrices useful for diagnostic purposes. Exploratory data analysis combined with two-sample Kolmogorov-Smirnov tests identified that the majority of wavelet coefficient matrices contained features that separated the three tissue groups. Energy features (sums of squares of the wavelet coefficients) were extracted for those coefficients associated with collagen structure. A total of 49 features were extracted that were associated with the scattering rings along the meridian and equator of each SAXS image. Figure 3.5(a) displays the regions in the wavelet coefficient matrices that were extracted for this task as black rectangles. The confidence intervals for selected features with respect to the three tissue groups are displayed in Figure 3.5(b). The features are indexed by level, direction and location, for instance the symbol 'L6H5ax' indicates the sixth resolution level (L6) in the horizontal direction (H) for the locations associated with the 5th axial peak ‘5ax’. Significant differences between the tissue groups were evident in many features such as L6H3ax and L8H5ax. Projections were then found from a linear combination of the 49 features that mapped the data onto two discriminant vectors using a variant of discriminant analysis (termed projection pursuit discriminant analysis) that is described by Lee *et al* (2005). The three groups were separated into distinct clusters that were associated with tissue group and another pair of discriminant vectors was found that described a trend from ‘normal-healthy’ to ‘benign’ to ‘malignant’.

3.3.3.1 Constructive Criticism of the work of Falzon *et al* (2006)

The work of Falzon *et al* (2006) provided firm statistical evidence of a link between wavelet coefficient energy, collagen structure and tissue pathology. A rapid, automated feature extraction process was developed as an alternative to both the models of Butler *et al* (2003) and Erickson (2005). The model produced by Falzon *et al* (2006) had a considerable advantage over these other two models because it summarised a large amount of information across the scales of the wavelet transform and compacted it into two vectors. Despite the improvement over previous models there are inherent limitations in the study of Falzon *et al* (2006). These limitations include:

a) Only the Daubechies family of wavelet bases were included in the study.

b) Primitive feature extraction techniques were used.

c) A classification model was not developed and estimates of classification accuracy were not reported.

d) A limited amount of image information was extracted.

e) Non-linear dependencies and interactions between features were ignored when searching for useful discriminant vectors.

Selected issues from items (a)-(e) above are discussed in further depth.

3.3.3.2 Limited Family of Wavelet Bases

Only the Daubechies family of wavelet bases were included in the study. A large range of alternative bases exist that may have enhanced the separation of the tissue groups. Greater efficiency or discriminatory ability might have been achieved using alternative custom-designed bases tailored to the problem at hand.

3.3.3.3 Feature Extraction

Specific wavelet coefficient energy features were found by defining a 'box' around the location of the wavelet coefficients of interest (as in Figure 3.5(a)). Although the technique was effective, it was fundamentally a primitive method of feature extraction. Greater scope exists for the development of more sophisticated, targeted feature extraction techniques.

3.3.3.4 Linear Models

Projection pursuit discriminant analysis, was used to project the data from the 49 features onto two discriminant vectors (Lee *et al* 2005; Falzon *et al* 2006). These discriminant vectors clustered the data into three distinct groups associated with tissue pathology. The data appeared well separated visually but a classification model and the resulting accuracy estimates were not reported. The discriminant vectors manage to capture key linear dependencies in the data but they could not capture any non-linear dependencies that might have been present. Identification of any such non-linear dependencies could have provided additional insight into the structure of the data set.

3.3.4 Section Summary

Image analysis combined with statistical modeling is an alternative method to achieve breast cancer diagnosis using SAXS data. To date, this approach has produced some of the most sophisticated models and the most accurate results. Model selection, feature identification and the accuracy of image classification using these techniques are challenges that have been met but there is little consensus among researchers about the best approach to use. The three studies (Butler *et al* (2003); Erickson (2005) and Falzon *et al* (2006)) that were reviewed in this section suffer from fundamental limitations, the most concerning of which include model over-fit, biased model assessment and a poor choice classification features. Overall, there is still some evidence (from these studies) that a good diagnostic model of breast cancer can be produced using statistical image analysis techniques. The multi-scale approach using wavelets has been a component of models in two different projects and is believed to produce good results. A wide range of other techniques that model different aspects of the image will be reviewed in the next chapter.

3.4 Chapter Summary

The analysis of breast tissue SAXS images fall into two distinct camps. The first is the development of diagnostic models based upon the features related to a physical model of collagen structure, whilst the second focuses upon more automated techniques that incorporate a range of features that may or may not be readily identified with the structure of collagen. The first approach allows interpretation of the features in terms of the current understanding of collagen structure but this approach is very application dependent and often requires the reduction of the image data to one-dimension, followed by fitting of a model that smoothes the intensity profile and requires the laborious extraction of classification features. The second approach incorporates a wider variety of image information but it is (at present) difficult to interpret in terms of the structure of collagen and there is little consensus among researchers upon the most appropriate model to adopt. One of the greatest problems common to many of the research projects reviewed in this chapter is that a vast number of statistical tests were performed. This is not an economical or efficient approach and all of the features considered to date have lost some of the information present in the images. A unified approach to the analysis of SAXS images is needed and a image model might offer a solid framework to build on.