FTIR Imaging: A Route Toward Automated Histopathology

Benjamin L. Bird, BSc

Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy, July 2006

Declaration

I, Ben Bird, hereby certify that this Thesis was composed by myself and is a record of my own work over the period September 2002 to September 2005.

Where work has been carried out in collaboration, the relevant researcher or researchers have been acknowledged.

This Thesis has not been accepted in partial or complete fulfilment of any other degree or professional qualification.

Benjamin Bird

January 2007

Acknowledgements

Firstly, I would like to give a special thanks to my supervisors, Prof. Mike George and Prof. Mike Chesters, for all their help and advice during my time in Nottingham. Without their invaluable support I doubt I would have ever been able to complete this degree.

I must also give a huge thanks to John Chalmers. His advice, support and friendship shall always be remembered.

None of the cervical cancer work in thesis would have been possible without the support and help of the brilliant medical staff at both Derby City and Derby Infirmary Hospitals. I especially thank Dr. Andy Hitchcock and Mr. Ian Symonds.

In addition, I need to thank Prof. Hugh Barr and Drs Nick Stone and Jenny Smith for their fruitful collaboration researching lymph node cancer.

A big thank you must also go to Xiao-Ying Wang and Dr. Jonathan Garibaldi for their expertise and collaboration using Fuzzy Logic algorithms. I am also grateful to Tom Hancewicz and Paul Pudney from Unilever for their help using multivariate curve resolution analysis techniques.

Many thanks must also go to Mark Tobin for his help conducting experiments whilst at the Daresbury SRS Laboratory.

I would also like to thank Prof. Max Diem and Dr. Melissa Romeo for their collaboration using dispersion correction algorithms.

For financial support, I thank the EPSRC and The University of Nottingham.

I must also thank all the members of the research group, past and present, who have made working in Nottingham so much fun.

Finally, but most importantly, I thank my friends, family and Andrea for all their love and support.

Contents

Abbreviations

Below are listed the abbreviations commonly used throughout this Thesis. Any abbreviations that are not listed are given in the text at their first occurrence.

IR	infrared
FTIR	Fourier transform infrared spectroscopy
SRS	synchrotron radiation source
FTT	fast Fourier transform
BaF_2	barium fluoride
PBS	phosphate buffered saline
ESS	elastic scattering spectroscopy
PET	positron emission tomograpgy
CAT	computed axial tomography
MRI	magnetic resonance imaging
DMPC	dimyristoylphosphatidylcholine
FPA	focal plane array
HgCdTe/Si	mercury cadmium telluride silicon
HgCdTe	mercury cadmium telluride
MCT	mercury cadmium telluride
CCD	charge-coupled device
LED	light emitting diode
CPU	computer processing unit
H&E	haematoxylin and eosin
ALND	axillary lymph node dissection
PAP	papanicolauo

CIN I	mild cervical intraepithelial neoplasia
CIN II	moderate cervical intraepithelial neoplasia
CIN III	severe cervical intraepithelial neoplasia
HPV	human papilloma virus
CIS	carcinoma *in situ*
HSIL	high grade intraepithelial lesion
LBC	liquid based cytology
DNA	deoxyribonucleic acid
RNA	ribonucleic acid
PCA	principal component analysis
PFA	principal factor anlaysis
LS	least squares
PC	principal component
KM	k-means clustering
FCM	fuzzy c-means clustering
LDA	linear discriminant analysis
HCA	hierarchical cluster analysis
SAFCM	simulated annealing fuzzy c-means clustering
VFC-SA	variable string length simulated annealing clustering
PCA-FCM	principal component analysis coupled with fuzzy c-means clustering
MCR	multivariate curve resolution analysis
ANN	artificial neural network

All references are numbered and listed at the end of each individual Chapter.

Abstract

The focus of this study is the potential use of FTIR imaging as a tool for objective automated histopathology. The Thesis also reports the use of multivariate statistical techniques to analyse the FTIR imaging data. These include Principal Component Analysis (PCA), Hierarchical Cluster Analysis (HCA), Multivariate Curve Resolution (MCR) and Fuzzy C-Means Clustering (FCM). The development of a new PCA-FCM Clustering hybrid that can automatically detect the optimum clustering structure is also reported.

Chapter 1 provides a brief introduction to the use of vibrational spectroscopy to characterise biomolecules in tissues and cells for medical diagnosis.

Chapter 2 details the basic histology of a lymph node before proceeding to present imaging results gained from the analysis of both healthy and diseased lymph node tissue sections. The ability of each multivariate technique to discriminate different tissue types is discussed. In addition, the spectral features that are characteristic for each tissue type are reported. The development and application of a new PCA-FCM Clustering algorithm that can automatically determine the best clustering structure is also described in full. The results indicate that cellular abnormality provides changes to both the protein and nucleic acid vibrations. However, similar spectral profiles were identified for highly proliferating cells that were contained within reactive germinal centres of the lymph node.

Chapter 3 provides a short introduction to the histology of the cervix before presenting imaging results that were gained from the analysis of both healthy and diseased cervical tissue sections. The ability of each multivariate technique to discriminate different tissue types is discussed. In addition, the spectral features that are characteristic for each tissue type are described in detail. Novel imaging experiments upon exfoliated cervical cells are also presented. It would appear that cellular abnormality in cervical tissues and cells affects both the protein and nucleic acid features of the spectra. Glycogen and glycoprotein contributions that are prevalent in healthy tissues are also absent.

Chapter 4 details sample preparation methods, the instrumentation and procedures used for data acquisition, and the subsequent data processing and multivariate techniques applied to analyse the collected spectral datasets.

Chapter One

Introduction

The focus of this study stems from the great need to reliably improve the diagnosis of life threatening diseases, in particular to this research, the progression of cancers. Histopathologic evaluation of human tissues (histology) and exfoliates (cytology) are now well established techniques for disease identification, and have remained relatively unchanged since their clinical introduction [1]. Excised or exfoliated material is initially formalin-fixed to prevent its degradation, and subsequently prepared onto glass substrates for light microscopic analysis. The identification of cellular and extra-cellular components within the tissues and cells are enhanced by the addition of dyes that stain different components different colours. These staining patterns provide the basis for morphological pattern recognition, allowing a trained observer to distinguish healthy and diseased tissue. However, conventional histology remains a subjective technique, with significant problems often encountered. These include missed lesions, perforation of samples, and unsatisfactory levels of inter- and intra-observer discrepancy [2 – 7]. With the unfortunate recent decline in recruitment of qualified pathologists and cyto-technicians, there is a strain to complete an ever increasing and demanding workload. A less operator-dependent and more automated analysis of clinical samples is therefore highly sort. The use of Fourier transform infrared (FTIR) spectroscopy as this tool shows promise, with the distinct potential to highlight small biochemical changes occurring at the cellular level that could predispose cancerous change [8,9].

FTIR spectroscopy is a non-destructive photonic technique that can provide a rapid measure of sample chemistry. Covalent bonds within molecules absorb infrared (IR) light at different wavelengths dependent upon the atoms in a bond, the type of bond, the type of vibration and any inter- and intra-molecular interactions present. The intensity of light absorption is further directly related to the concentration of molecules [10]. The IR spectrum collected from a sample can therefore provide detailed information about the chemical composition of that material. Consequently, an IR spectrum collected from human tissues or cells can provide a direct indication of cellular biochemistry [11 -13]. Differences within the biochemistry of cells that accompany the onset of disease could therefore be characterised by changes within the IR spectrum. These spectroscopic differences would be related to changes in the concentration and conformational orientation of functional groups associated with lipids, proteins, nucleic acids and carbohydrates, the basic building blocks of mammalian cells.

In principle, the application of IR spectroscopy for diagnostic purposes has a number of advantages over more established techniques such as PET, CAT and MRI scanning. These include speed, sensitivity, flexibility, comparatively low cost and no dependence upon the physical state. A number of sampling techniques exist that enable spectra to be obtained from a large variety of biological samples, which include solids (e.g. bones, teeth), liquids (e.g. body fluids) and tissues [9,14]. The introduction of an endogenous chromophore that may potentially disturb the sample characteristics is also not required. Furthermore, changes that occur within the biochemistry of cells precede any morphological or symptomatic manifestation, thus

IR spectroscopy could probe for earlier stages of disease not presently detectable via conventional histopathology.

1.1 IR Spectral Characteristics of Mammalian Cells

The complexities of biological systems have meant that interpretation of IR spectra collected from mammalian cells is not always straight forward. A detailed understanding of the infrared-active constituents of these samples is therefore required. In principle, the IR spectrum of most tissues can be closely estimated to the summation of lipids, proteins and nucleic acid contributions. These individual components have many characteristic absorptions that span the mid-infrared region (4000 – 400 cm^{-1}), with an overlapping of several bands often evident. Only very small contributions from carbohydrates are observed within the mid-infrared region for most human tissues. However, mammalian tissues that store large quantities of carbohydrates in the form of glycogen can give rise to strong absorptions bands below 1200 cm^{-1}. Investigations to evaluate the contribution of these individual constituents to the IR spectrum have been undertaken by several research groups [9,14,15,16]. Additional studies that utilise a synchrotron source for enhanced spatial resolution have also been reported [17,18]. Spectral bands characteristic for these constituents will therefore be discussed utilising spectra collected from a typical protein, lipid and polynucleotide [19]. These spectra were collected from dry films of material to reduce interference from water and are shown in Figure 1.

The first spectrum (Figure 1a) was collected from a typical phospholipid, dimyristoyl phoshatidylcholine (DMPC). The most intense bands observed in this spectrum are

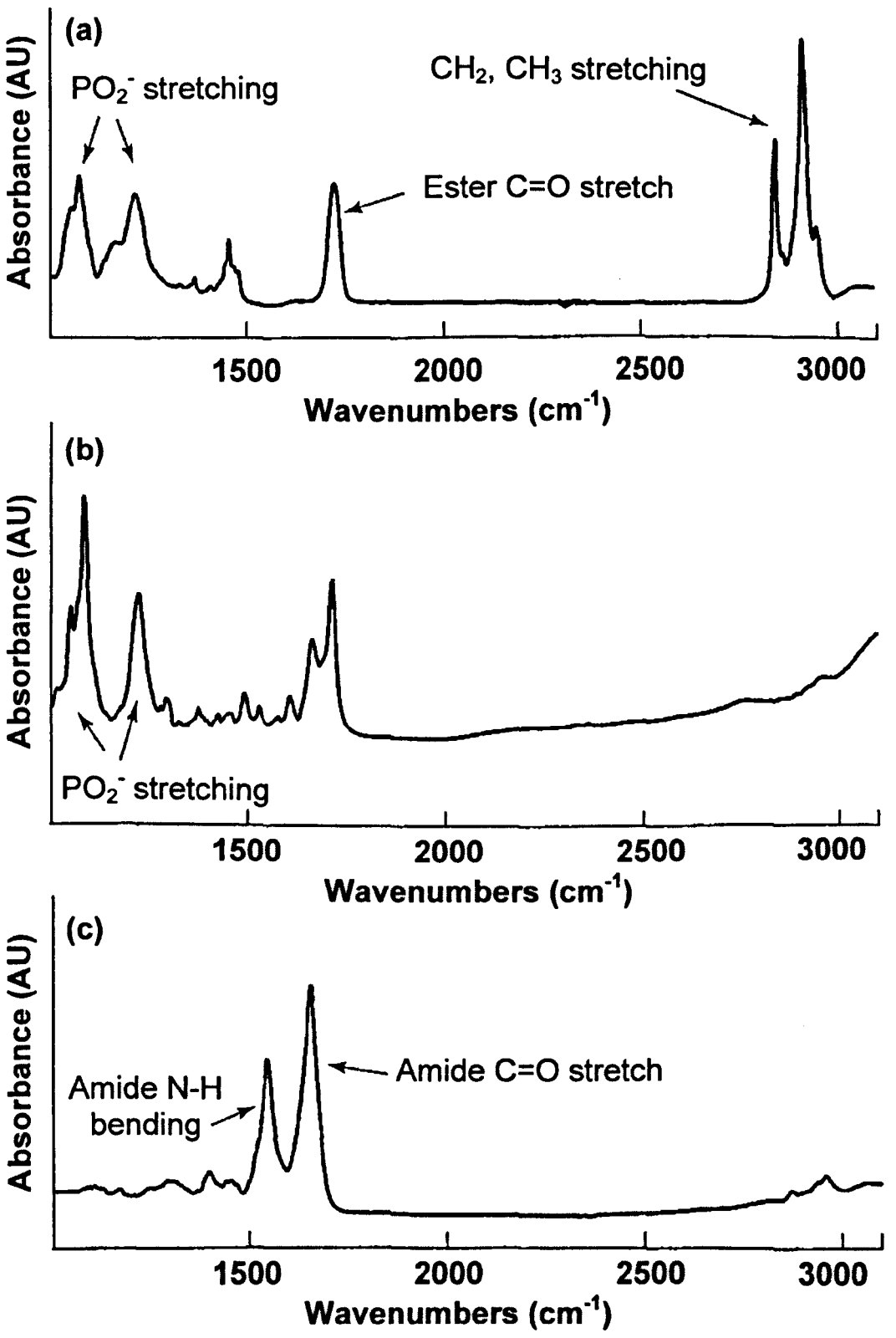

Figure 1: *IR spectra collected from dry films of (a) dimyrisolphosphaicdyclholine (DMPC), (b) a polynucleotide and (c) haemoglobin. Adapted from ref [19]. Note all subsequent IR spectra are reported from high to low frequency in wavenumber units (cm^{-1}).*

those found within the 3000 – 2800 cm^{-1} region. Such bands are analogous to the IR spectra of alkanes, and can be attributed to the symmetric and antisymmetric stretching vibrations of CH_2 (2852 and 2922 cm^{-1}) and CH_3 (2874 and 2956 cm^{-1}) respectively [20]. The symmetric and antisymmetric stretches for the CH_2 groups are in the order of 10 – 20 times more intense than those observed for the CH_3 stretches. This reflects a distinctly larger concentration of CH_2 groups present within lipids. The position and band width of the CH_2 and CH_3 stretching absorptions can also reveal information regarding the packing of acyl side chains [21,22]. The strong absorption band located within the 1800 – 1600 cm^{-1} region is associated with the C=O stretching vibration of ester groups. This band is normally located at c.a. 1735 cm^{-1}, but its frequency can be strongly affected by hydration [21,22]. Between 1500 – 1250 cm^{-1} a small number of very weak bands are present. The most distinct is that located at c.a. 1468 cm^{-1} which is characteristic of the CH_2 scissoring vibration. Below 1250 cm^{-1} two intense bands are noticeable and arise due to the symmetric (c.a. 1085 cm^{-1}) and antisymmetric (c.a. 1225 cm^{-1}) vibration modes of phosphate groups (PO_2^-) respectively. The frequency of these bands can additionally provide insight into head-group hydration [9].

The second spectrum displayed in Figure 1b was alternatively collected from a polynucleotide. When initially scrutinising the spectrum it is clearly noticeable that absorptions above 1800 cm^{-1} are greatly reduced. The strong CH_3 stretching absorptions previously observed for phospholipids are no longer present and only weak bands for CH_2 stretching absorptions are distinguishable. These are likely to be characteristic of CH_2 vibrations from carbohydrate residues and the C – H stretching vibrations of the nucleotide bases. For nucleic acids, two distinct bands

can be observed within the 1800 – 1600 cm^{-1} region located at 1717 and 1666 cm^{-1}. These absorptions are characteristic of the C=O stretching vibrations of purine and pyrimidine bases respectively. Below 1500 cm^{-1} a number of sharp but weak absorptions are apparent. The major absorption bands again occur from the vibrational modes of phosphate (PO_2^-) groups. These are associated with the phosphodiester linkages of the polynucleotide chain and are assigned to symmetric (1085 cm^{-1}) and antisymmetric (1225 cm^{-1}) phosphate stretches.

The third spectrum displayed in Figure 1c was collected from a typical globular protein, human haemoglobin. Again within this spectrum, absorptions above 1800 cm^{-1} are negligible. Absorption bands arising from CH_2 and CH_3 stretching vibrations are relatively weak and are likely to reflect small contributions from the amino acid side chains. When directly comparing these bands to those observed for a phospholipid (Figure 1a), it can be noted that the relative band intensity ratio for these peaks is greatly reduced. This spectral change would substantiate a more equal proportion of CH_2 and CH_3 groups within the protein side chains. Below 1800 cm^{-1} a number of strong bands can be observed. The most intense absorption is located at c.a. 1650 cm^{-1} and is more commonly termed the amide I band. This absorption is characteristic for the C=O stretching vibration of the amide C=O group. The frequency of this band can also be a sensitive marker for the conformation of the protein secondary structure [23 – 25]. Two additional amide modes are also observed within protein spectra. The amide II absorption band is normally located between 1500 – 1560 cm^{-1}, and is predominately associated with the N – H bending vibrations and the C – N stretching vibrations of proteins. In contrast, the amide III band is attributed to a complex vibration involving C – N stretching, N – H in plane

bending and a significant contribution from CH_2 wagging vibrations. These bands normally occur between 1350 – 1250 cm^{-1}. In addition, further important absorption bands assigned to the COO^- symmetric and antisymmetric stretching vibrations can be located at 1580 and 1400 cm^{-1} respectively. These absorptions are associated with the amino acids aspirate and glutamate. Collagen, an important structural protein found in most connective tissues, also provides a number of characteristic peaks within the spectrum [26,27]. Two distinct peaks located at 1030 and 1080 cm^{-1} can be observed that are attributed to the C – O stretching vibrations of the carbohydrate moieties attached to collagen. A further triad of peaks located at 1280, 1240 and 1204 cm^{-1} are also often observed and can be used to monitor the relative collagen concentration within these tissue types [26,27].

The spectral region between 2800 – 1800 cm^{-1} is generally free from absorptions due to lipids, nucleic acids, proteins and carbohydrates found within mammalian tissue. The only exception would be contaminating bands originating from atmospheric water and CO_2. However, if the spectrometer and surrounding sample area are adequately purged these contributions should be negligible. A table displaying the main absorption bands found within mammalian tissue is further displayed in Table 1. These reported frequencies can only be used as a rough guide since several factors can cause variation, including sampling type, preparation method, data collection procedure and instrumentation sensitivity. A detailed knowledge of both spectroscopy and histology is required before an assignment of IR absorptions to specific chromophores can be made within mammalian tissue.

Absorption Peak (cm^{-1})	Assignment	Cellular Constituent
3290	Amide A N- H stretch	Protein
3050	Amide B N – H bending 1^{st} overtone	Protein
3010	Olefinic C – H stretch	Lipids
2960 – 2930	CH_3 antisymmetric stretch	Lipids, proteins
2925 – 2920	CH_2 antisymmetric stretch	Lipids, proteins
2874 – 2870	CH_3 symmetric stretch	Lipids, proteins
2855 – 2850	CH_2 symmetric stretch	Lipids, proteins
1735	Ester C=O stretch	Lipids
1717	Purine C=O stretch	Nucleic acids
1666	Pyrimidine C=O stretch	Nucleic acids
1655 – 1650	Amide I C=O stretch	Proteins (α-helical secondary structure)
1640 – 1630	Amide I C=O stretch	Proteins (β-sheet secondary structure)
1580	COO^- antisymmetric stretch	Proteins
1560 – 1500	Amide II N – H bending	Proteins
1470 – 1405	CH_2 symmetric and aymmetric bending	Proteins, lipids
1400	COO^- symmetric stretch	Proteins
1380 – 1250	CH_3 symmetric and antisymmetric bending	Proteins, Lipids
1280	Amide III of collagen	Protein
1245 – 1220	PO_2^- antisymmetric stretch	Nucleic acids, lipids
1240	Amide III of collagen	Proteins
1204	Amide III of collagen	Proteins
1155 – 1150	C=O stretch of glycogen	Carbohydrate
1085 – 1075	C=C stretch of glycogen	Carbohydrate
1028 – 1020	C – O – H deformation of glycogen	Carbohydrate
1080	PO_2^- symmetric stretch	Nucleic acids, lipids

Table 1: *Representative frequencies of the major absorptions bands found within mammalian tissues and cells. Adapted from ref [19].*

1.2 FTIR Microspectroscopy with Multichannel Detectors

Infrared (IR) spectroscopy is a proven and powerful analytical tool for spectrochemical analyses [10,28]. Over the past few decades, advances in instrument technology have enabled the rapid acquisition of IR spectra using interferometers in reasonably uncomplicated configurations. However, until relatively recently, IR spectroscopy has largely been a bulk material technique, since the collection of spectra from microscopic sample volumes proved problematic. Early IR analyses upon mammalian tissues and cells utilised similar macroscopic techniques to assess their biochemical compositions [29-37]. Although these proved that to some extent healthy and diseased tissue could be characterised by observed spectral differences, a number of complications remained. Spectroscopic results could not be directly correlated to cellular pathology since it was impossible to establish the types and numbers of cells that were being scrutinised by the technique. Since different tissue and cell types provide significantly different spectral profiles, contamination from non-diagnostic cells could manifest themselves in the average spectrum acquired [35,38,39,40,41]. The correct interpretation and classification of these spectra therefore proved problematic.

The advent of instruments that couple IR spectroscopy and optical microscopy, however, now permit the collection of IR spectra from sample volumes in the order of 20 x 20 x 5 μm [10,42]. This would describe an approximate sample thickness of 5 μm at the focal point of the microscope, scrutinising a 20 x 20 μm sample area that is defined by knife edges or a fixed aperture. The introduction of such instruments thus allowed the collection of individual spectra from pure tissue components within

sectioned material that could be subsequently correlated to histology, free from averaging contaminations. Furthermore, spectral mapping techniques could be used to examine larger regions of the sample and allow more precise characterisation of different cell types present in the tissue matrix. Such IR maps are collected by scanning a sample in a raster pattern through the focal point of a single detector, using steps that are the same size as the x and y dimensions of the pixel element [43]. Although such experiments are time consuming, work to date using spectral mapping of tissue sections, coupled to some form of statistical analysis, has clearly proven that FTIR microscopy can discriminate alternative tissue types and disease states comparable to conventional histology [44 – 46].

Despite the advantages mentioned above for FTIR microscopic mapping using a single detector, the technique is still limited in its applicability for automated pathology. The collection of FTIR spectral maps from large sample areas is very time consuming and can require long computation times for the analysis of the data. If we consider the method by which suspicious lesions are conventionally screened for disease, a better understanding of the requirements necessary for a spectroscopic diagnosis become clear. An example tissue section cut from a diseased prostate will be used to help demonstrate these requirements and is shown in Figure 2.

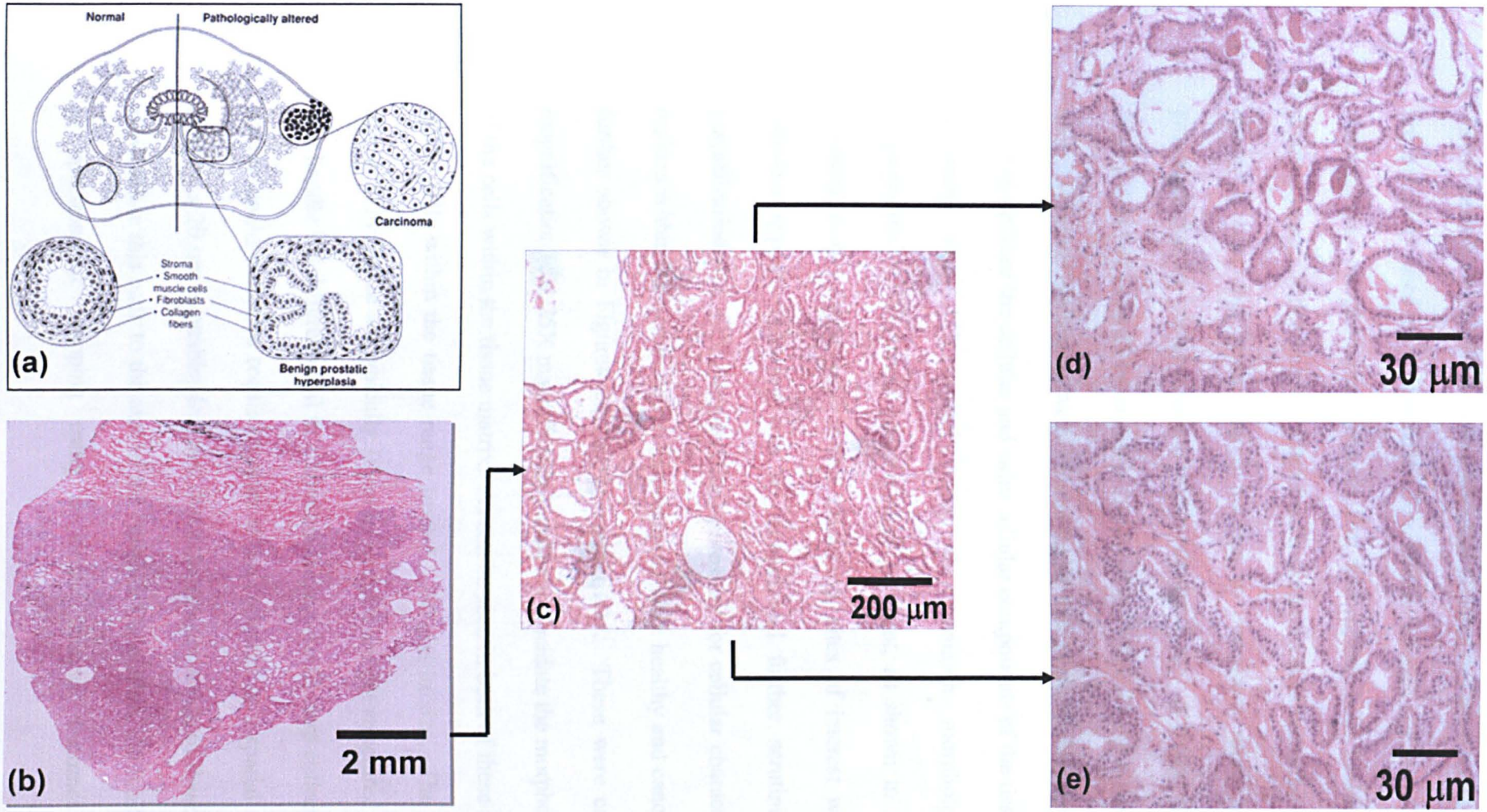

Figure 2: *Conventional screening method used in histopathology. A prostate tissue section displaying an adenocarcinoma is used as an example. (a) Schematic describing morphological features apparent within healthy and diseased tissue. (b) Photomicrograph captured from a H&E stained prostate tissue section c.a. 10 x 10 mm in size. (c) Magnified region from within the tissue section (c.a. 1000 x 1000 µm in size). The morphological patterns within the tissue matrix become more apparent at this resolution. (d) High magnification image of region displaying healthy morphological features (c.a. 200 x 200µm in size). (e) High magnification image of region displaying malign morphological features diagnostic for prostate adenocarcinoma (c.a. 200 x 200µm in size). Adapted from ref [68].*

As mentioned previously, the morphological patterns or features within stained tissue are presently used as a descriptor for disease change. Figure 2a displays the characteristic morphological features that accompany both healthy and diseased prostate tissue. Clinical screening of such samples relies upon the identification of abnormal cells that feature large nucleus to cytoplasm ratios and have tissue structures that are uncharacteristic within healthy tissue. A typical tissue section cut from a diseased prostate specimen is displayed in Figure 2b and is c.a. 10 x 10 mm in size. This was stained using a haematoxylin and eosin (H&E) dye that enhances contrast between the cellular and extra cellular components of the tissue. At a higher resolution (c.a. 15X magnification) the common morphological features characteristic of this tissue become more apparent, as shown in Figure 2c. By scanning across the sample at this resolution, sites of interest where the tissue structure appears abnormal can be located and further scrutinised at a high magnification (c.a. 40X magnification) necessary for cellular characterisation. Two regions within the tissue section that comprise both healthy and cancerous tissue are further shown in Figures 2d and 2e respectively. These were captured at high magnification (c.a. 25X magnification) and help elucidate the morphological features of the cells within the tissue matrix. As can be seen in both of these images, the size of the cells within the tissue range from 5 – 50 μm in size. The malignant cells shown in Figure 2e are typically 10μm in size and are not structured tightly together in a bundle but differentiated in a circular shape. An IR map collected from such a region would therefore require spectra to be collected at a spatial resolution of at least 20 x 20 μm to enable clear and distinct characterisation of these tissue features. If we relate this back to the size of the entire tissue section, the limiting capabilities of spectroscopic mapping using a single detector becomes clear. The

characterisation of one abnormal cell would be equivalent to locating a house 20m wide in a 10km radius area. If an IR map was acquired from the entire tissue section using a 20 x 20 μm spatial resolution, c.a. 250,000 individual IR spectra would be collected. Such an analysis would be very time consuming and ultimately insufficient for automated pathology.

Around a decade ago, the first efforts to couple mulitchannel detectors with FTIR spectrometers were reported [47–50]. These instruments utilised detector technology originally available only to the military for missile guidance systems. The migration from single element detectors to focal plane array detectors (FPA) has born a new methodology often termed "chemical imaging". By use of FPA detector systems, the time frame previously required to map large spatial areas has been reduced by several orders of magnitude and permits the measurement of 4096 spectra in seconds to minutes (by use of a 64 x 64 FPA detector, for example). However, the noise level of a single spectrum collected from an FPA measurement is worse than that recorded from a single detector, since these systems utilise hybrid HgCdTe/Si detector technology. In comparison to mapping methods utilised by single detector systems, the sample is left in a fixed position under the detector array during data collection. Previous challenges of collecting spectra from small sample areas through small apertures (diffraction limit) are alleviated in FPA measurements since the spatial resolution is given by the effective pixel size of the FPA detector. Different pixel spatial resolutions are achieved in this system by changing the magnification cassegrain objectives within the microscope. Therefore, limitations of spatial resolution are dependent only on the nature of the light used. Schematics

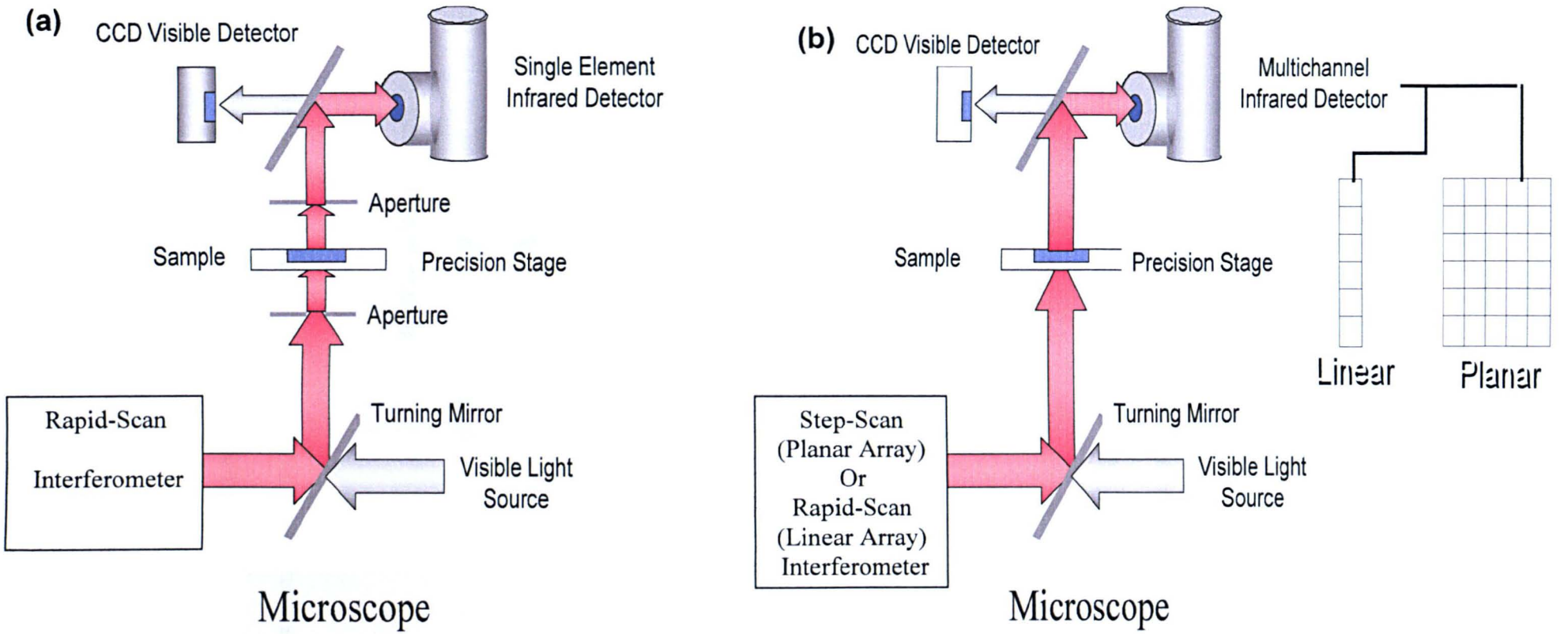

Figure 3: *Schematics describing the layout and principal of FTIR microscopes coupled to single-channel (a) and multi-channel (b) detectors. Single-channel detectors require apertures to delineate the sample area examined. Multi-channel detector systems utilise varying magnification cassegrains to achieve different spatial resolutions. Two types of multi-channel detection system are presently commercially available. Focal plane array (FPA) detector systems utilise hybrid HgCdTe/Si technology for their detection/readout hardware. These range in size from a 64 x 64 to a 256 x 256 array of detector elements. Linear array detectors alternatively use HgCdTe technology only, and therefore offer higher detector sensitivity. These usually consist of 16 or 32 detector elements arranged in a linear fashion. Adapted from ref [68].*

displaying the layout and principle of an IR microscope that are coupled to both a single channel and multichannel detector are displayed in Figure 3. In general, the acquisition of data using an FPA detector system does not vary greatly from single element measurements with a standard interferometer. The most important difference is that 4096 detector elements (using a 64 x 64 FPA detector) are read at the same time during the spectral acquisition. This process requires a greater amount of time than previously needed for a single detector and can be more directly related to the ability of the electronics used to read the elements within the FPA detector array [51,52]. Although continuous-scan interferometry is the most common form of data acquisition used in modern spectrometers, since it provides several advantages (Multiplex and Jaquinot) [10,53], the optical retardation is coupled to the time domain and thus requires rapid signal detection to prevent data acquisition errors [10,54]. Since the rates of data acquisition from FPA detectors are markedly longer, a rapid-scan configuration is no longer permissible. Hence a step-scan approach is required for accurate data collection with FPA detector systems [47]. In practice, this means the moving mirror within the Michelson interferometer no longer moves continuously during data acquisition, but conversely waits until each detector element has been read before moving to its next position. By collecting data in such a way, an interferogram from each pixel in the array is collected simultaneously and later transformed into an IR spectrum.

More recently, small linear array detector systems have been developed that provide a compromise between the multi-channel advantage of large FPA detector systems and the high fidelity features of rapid-scan FTIR spectrometry. By using a smaller number of detectors a number of advantages are realised. The individual detector

elements within the array are significantly smaller when compared to their single element counterparts, and employ HgCdTe detector technology alone. In addition, since the detector numbers are small, their uniformity is relatively high. Each element has its own gold connection used to perform its signal processing, and therefore allows all channels to be continuously sampled. Thus detector sensitivity is significantly high when compared to FPA systems. Within these systems, the spatial resolution is again determined by the optics of the microscope and does not require an aperture to delineate the sample area. But these systems offer a very limited number of magnifications. For example, the commercially available Perkin Elmer Spotlight Imager uses a 16 element linear array detector and can collect spectra from pixel sizes of either 25 x 25 μm or 6.25 x 6.25 μm. This is achieved by use of a Z fold tube that dips a 4X magnification mirror into and out of the beam. Similar to point mapping using a single detector, the sample is moved underneath the array in a raster pattern, building an image one linear element at a time. However, the instrument allows IR spectral maps to be collected from samples independent of their size or orientation, and rivals FPA-based instruments in data acquisition time with significantly lower cost.

The advent of both FPA and linear array detector systems now permits the rapid collection of IR spectroscopic data from spatial resolutions that are in the same order of magnitude as a mammalian cell, and from samples sizes close to that conventionally scrutinised by histology. Larger and more sensitive detector array systems could therefore be utilised for a spectroscopic, non-subjective route toward automated pathology.

1.3 Non Subjective Analysis of Microscopic IR Maps

As mentioned above, IR microscopic maps collected from large sample areas can contain many thousands of individual spectra. As the amount of pixels within a map increases and the spatial resolution is improved, the size and complexity of such datasets becomes astronomical. The interpretation of IR spectra collected from biological material can also be somewhat subjective and requires a detailed understanding of both spectroscopy and histology. Thus the extraction and presentation of useful information from these complex datasets presents a unique challenge. However, a number of statistical methods that remove this subjectivity can be applied to spectroscopic maps. These can provide insight into the individual components found within a sample and help elucidate descriptive spectral features.

A common method of analysing and presenting data from IR microscopic maps is "functional group" mapping. This method utilises parameters such as the peak intensity of an absorption band, the integrated intensity of an absorption band and the frequency at which an absorption band arises. In addition, an intensity map can also be calculated from the ratio of peak or integrated intensities of two separate absorption bands. Since the intensity of a functional group can be directly related to the concentration of the material giving rise to that absorption, the distribution of this species throughout the collected map can be visualised. This is achieved by plotting the recorded or calculated intensity at each pixel contained within the map. However, care must be taken if the original recorded absorption is used, especially in spectra of tissues, since large baseline fluctuations are often observed due to scattering effects. Such differences can introduce artefactual changes to the intensity

that more closely relate to the variation of the observed sloping baselines. Differences in the intensity of absorption bands can also be caused by irregularities in sample thickness and cellular density, which must also be taken into consideration. To negate such problems it is therefore necessary to pre-treat the data to compensate for these observed irregularities. A variety of pre-processing routines can be applied to the data that can correct for these changes. In our work, we adopted a 6 base point linear interpolation to compensate baseline distortions and subsequently vector normalised all spectra to reduce effects from irregular sample thickness (see section 4.4). By use of such routines, a more accurate distribution of chemical species found within analysed tissues can be observed.

As mentioned in section 1.1, a large number of absorptions can be observed in an IR spectrum collected from mammalian tissue. Of particular interest are those arising from proteins (e.g. amide I C=O stretching absorption band), lipids (e.g. ester C=O stretching absorption band) and nucleic acids of DNA (e.g. purine C=O stretching absorption band). By plotting the intensity of these bands for each spectrum at the co-ordinates from which they originate, the distribution of proteins, lipids and nucleic acids throughout the sample can be visualised. This type of analysis can therefore provide an outline of the general biochemistry found within tissues. The usefulness of such an analysis to characterise a lymph node tissue section is illustrated in Figures 4 and 5. This tissue section was cut from a malignant lymph node that had almost been completely infiltrated by fatty and fibrocollagenous scar tissues. However, a few small pockets of remnant cancerous tissue could still be found. A white light image of the entire tissue section is displayed in Figure 4a. Unfortunately a parallel H&E stained tissue section was not made available for this

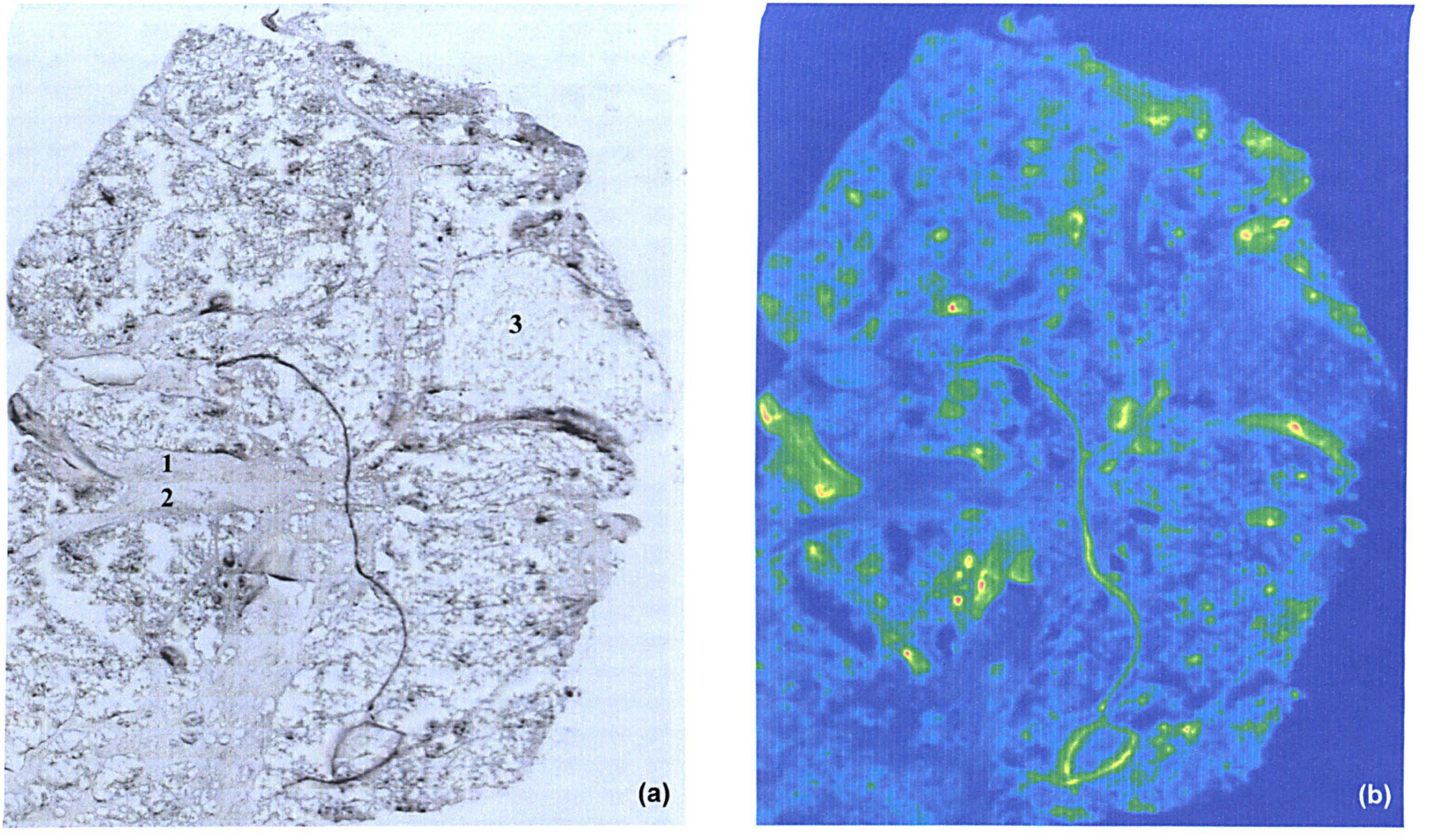

Figure 4: *Spectroscopic analysis of a malignant lymph node. (a) White light image collected from entire lymph node. Tissue types found within the mapped area include cancerous cortex (1), collagenous scar (2), and fatty (3) tissues. (b) Total absorbance IR image of mapped lymph node.*

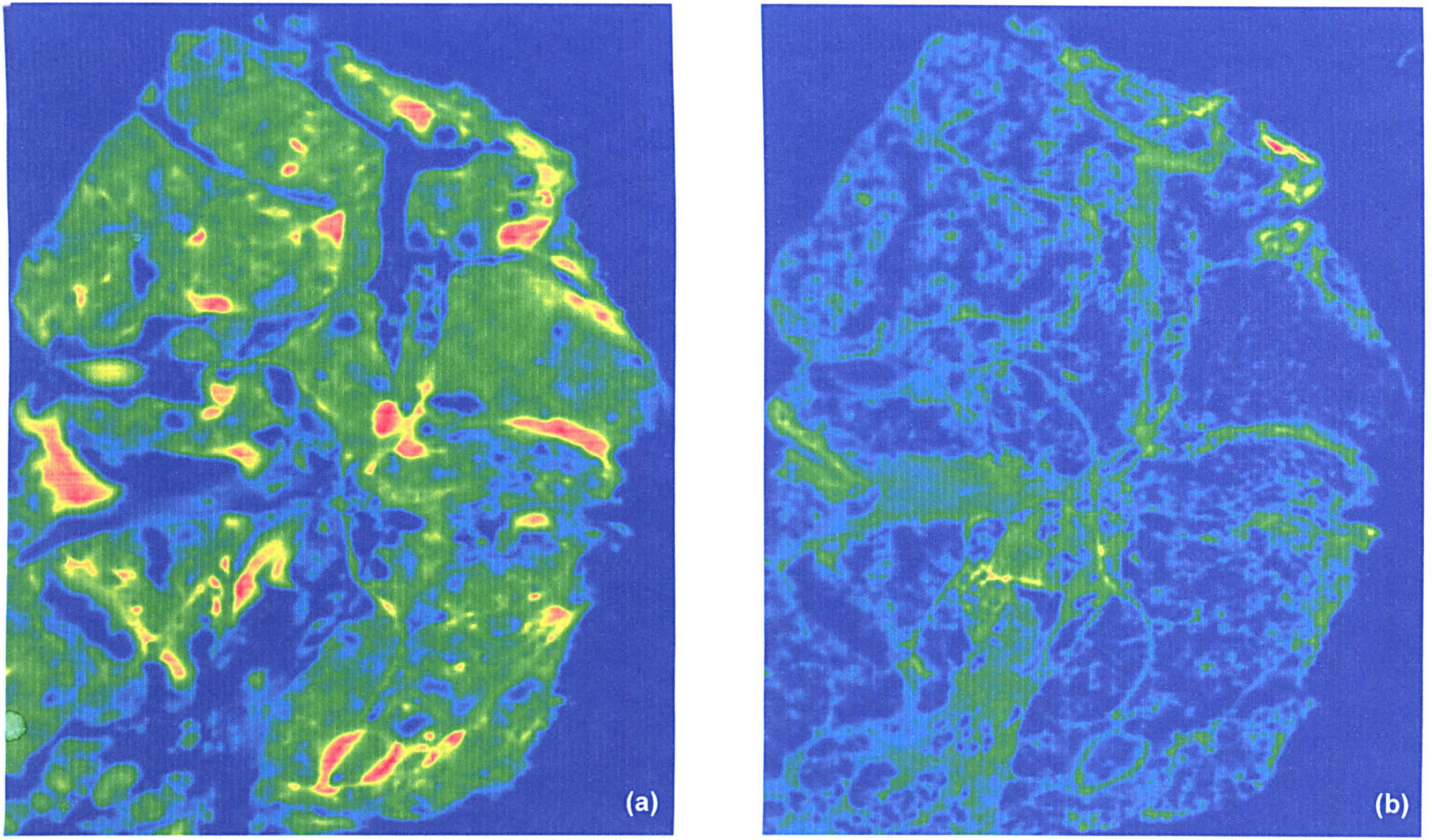

Figure 5: *Functional group maps of malignant lymph node. (a) Lipid functional group map calculated from the peak height intensity of the absorption band located at 1735 cm^{-1}. (b) Protein functional group map calculated from the peak height intensity of the absorption band located at 1655 cm^{-1}. The colour scale ranges from red indicating spectra with a high intensity for that band, to blue which display a weak intensity.*

node, but the main types of tissue can still be visualised via contrast in light intensity of the tissue regions (Figure 4a). An IR micro-spectral map was collected from the entire lymph node. By use of a step size and aperture of 25 μm, a total of 66,402 individual IR spectra were collected from a spatial area of 6900 x 7650 μm. The total absorbance image constructed for this map is further shown in Figure 4b.

Two functional groups maps were calculated for this map and are shown in Figures 5a and 5b respectively. The first of these, shown in Figure 5a, was calculated using the relative intensity of the absorption band located at 1735 cm^{-1}, which is characteristic of the C=O stretching vibration of ester groups in lipids. When scrutinising this map, it can be seen that strong contributions from lipids are clearly apparent within the infiltrated and pure fatty tissue regions. This would make histological sense since these adipose tissues are heavily contributed to by lipids. In contrast, the second functional group map, shown in Figure 5b, was calculated using the relative intensity of the Amide I band located at 1655 cm^{-1}, which is characteristic of proteins. This alternatively highlights regions on the lymph node whereby remnant cancerous and fibrocollagenous scar tissue is located. Unfortunately the monitoring of nucleic acid distribution in tissue sections can be difficult since the intensity of the DNA absorption band located at 1717 cm^{-1} is very weak and often absent in most tissues. In this case, only the phosphate (PO_2^-) symmetric and antisymmetric stretching absorption bands located at 1080 and 1240 cm^{-1} can be used to analyse the nucleic acid distribution. However, both of these absorption bands will contain contributions from the phosphate groups of phospholipids. In addition, the amide III absorption bands of collagen can also provide a substantial contribution, especially in fibrocollagenous tissues. Thus

functional group maps calculated for these bands often provide little useful information and are unwise for monitoring nucleic acids.

Band intensity ratio maps can in some cases provide more detailed information about the biochemical composition of tissues since these utilise two individual chrompohores. For example, tissue specimens that do not include regions of fatty tissue can be more accurately probed for changes in their relative lipid to protein ratio by selecting two chromophores for analysis. A good method to estimate these changes is to calculate the ratio of the CH_2 and CH_3 stretching vibrations since both lipids and proteins give rise to these absorption bands. Within proteins, the amount of CH_3 and CH_2 groups in their side chains are nearly equal, whereas lipids typically contain 14 – 18 CH_2 groups and only one CH_3 group. Thus calculation of the ratio between these stretching vibrations can provide a more accurate analysis to monitor protein-rich and lipid-rich regions. As mentioned above, the monitoring of nucleic acid distribution is more complex since a number of overlapping bands from chromphores characteristic of proteins and lipids can also be present. An alternative method of estimating nucleic acid changes is to monitor the ratio of the absorbance bands at 1240 and 1204 cm^{-1}. Collagen, an important structural protein apparent in most connective tissues, displays absorptions at both these frequencies. In contrast, nucleic acids only exhibit an absorption band at 1240 cm^{-1}. A high ratio will therefore indicate a high nucleic acid contribution, whereas a low band intensity ratio indicates a more significant contribution from collagen. However, this type of analysis is again limited to tissues that do not include significant contributions from fatty tissues since these also provide an absorption at 1240 cm^{-1} from the phosphate groups of phospholipids.

As shown in this example, functional group mapping can readily identify the major tissue types present in a sample. However, such representations are limited for accurate spectroscopic diagnosis since they only allow the distribution of one or two chromophores to be identified. However, another branch of chemometrics termed multivariate analysis can enable the manipulation and investigation of data that contains multiple variables, such as an IR spectrum collected from mammalian tissue. These types of analysis have therefore become increasingly used in the analysis of tissues and cells because of this advantage. Underlying patterns hidden within extensively large and complex datasets can be identified that were previously undetectable using univariate or bivariate analyses such as functional group mapping. Early experiments that utilised such methods for the analysis of IR spectra collected from animal cells were reported by Nauman [55]. In this work, a form of unsupervised analysis, termed Hierarchical Cluster Analysis (HCA), was used to successfully classify IR spectra collected from different strains of bacteria. These types of multivariate analysis do not require any previous knowledge of the sample and alternatively search for similarities within the data to characterise them. Up to the present time, a multitude of unsupervised methods have been utilised for spectroscopic data analysis upon human tissues. These techniques have included Principal Component Analysis (PCA) [34,38,43,57], Hierarchical Clustering Analysis (HCA) [46,55,58,59,60,61,62], K-Means and Fuzzy C-means Clustering (KM, FCM) [63 – 66], and Simulated Annealing Fuzzy C-Means Clustering (SAFCM) [67]. These studies indicated that each multivariate technique could to a degree, be applied to disease diagnosis using spectroscopic data.

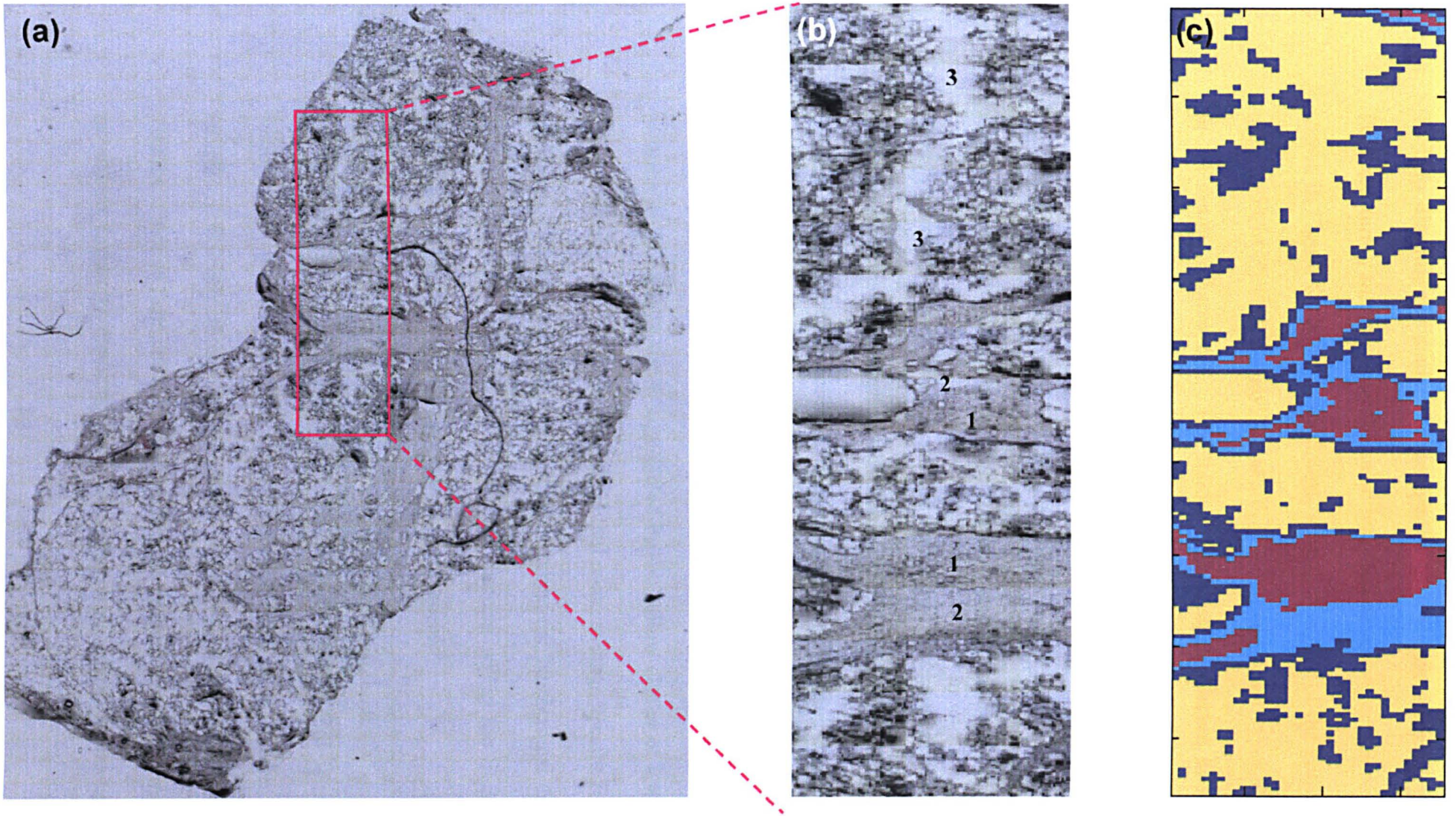

Figure 6: *a) White light image of entire lymph node tissue section. b) Magnified region displaying fibrocollagenous scar tissue that encapsulates small clusters of malignant cortex cells. The IR imaged area (875 x 4300μm) was mapped using a step size and aperture of 25μm for a total 6020 individual IR spectra. Tissue types found within the mapped area include cancerous cortex (1), collagenous scar (2), and fatty (3) tissues. (c) False colour image constructed from a 4 cluster FCM analysis upon the reduced spectral dataset. Each colour in the image represents a separate cluster of spectra that were partitioned in the analysis.*

To help illustrate the distinct advantages of multivariate analyses, an example of FCM clustering analysis upon the same lymph node tissue section detailed above is shown in Figure 6. In this experiment, a smaller region of the collected map was used for FCM clustering analysis, but still incorporated all the types of tissue present in the sample (Figure 6b). The false colour image presented in Figure 6c was constructed from a 4 cluster analysis of the reduced dataset. Each colour in the image represents a separate cluster of spectra that was partitioned by the analysis. By directly comparing the constructed cluster image to the white light image captured from the same region, it is clear to see the analysis has been able to correctly classify the spectra into groups that match histological diagnosis. The red cluster of spectra characterises the cancerous cells located in the central region of the remnant nodal tissue. In contrast, fibrocollagenous scar tissue that surrounds these cancerous cells is highlighted by the cyan cluster of spectra. The remaining remnant nodal tissue scattered across the section is described by the blue cluster of spectra, with the invading fatty tissue highlighted by the yellow cluster of spectra. Another distinct advantage of such an analysis is that mean average spectra for each cluster can easily be calculated and used to help interpret the biochemical differences that are occurring between them. Thus spectral features within the IR spectra can be identified that could be used for spectroscopic diagnosis.

In summary, recent advances in IR micro-spectrometers that incorporate array detectors has enabled the rapid acquisition of spectral datasets from large clinical samples previously thought impossible. The complexity of the data obtained requires sensitive forms of statistical analysis to unravel the underlying patterns that are present within the datasets. However, information relating to the biochemistry of

different pathological states can be assessed and could be used to develop an automated spectroscopic method for disease diagnosis.

1.4 References

[1] L. G. Luna, *Manual of Histologic Staining methods of the Armed Forces Institute of Pathology,* 1968, McGraw-Hill, New York.

[2] M. R. Emmet-Buck et al., *Am. J. Pathol.,* 2000, **156**, 1109.

[3] J. W. Van Sandick, J. J. Van Lanshot, B. W. Kuiken, G. N. Tytgat, G. J. Offenhaus and H. Obertop, *Gut,* 1998, **48**, 216.

[4] D. M. Melville et al., *Hum. Pathol.,* 1998, **20**, 1008.

[5] P. Jensen et al., *Dis. Colon Rectum,* 1995, **38**, 195.

[6] B. J. Reid et al., *Hum. Pathol.*, 1998, **19**, 166.

[7] K. E. Blackwell, T. C. Calcaterra and Y-S. Fu, *Ann. Otol. Rhinol. Laryngol.,* 1995, **104**, 596.

[8] M. Diem et al., *Analyst*, 2004, **129**, 880.

[9] H. Mantsch and M. Jackson, *J. Mol. Struct.,* 1995, **347**, 187.

[10] J. M. Chalmers and P. R. Griffiths, *Handbook of Vibrational Spectroscopy,* 2002, Wiley, New York.

[11] L. P. Choo, M. Jackson, W. C. Halliday and H. H. Mantsch, *Biochim. Biophys. Acta,* 1993, **1182**, 333.

[12] P. Lasch, W. Wasche, W. J. McCarthy, G. Muller and D. Naumann, *P. Soc. Photo-Opt. Inst.,* 1998, **3257**, 187.

[13] H. Fabian, M. Jackson, L. Murphy, P. H. Watson, I. Fichtner and H. H. Mantsch, *Biospectrosc.,* 1995, **1**, 37.

[14] R. A. Meyers, *Encyclopaedia of Analytical Chemistry*, 2000, **1**, Wiley, Chichester.

[15] N. Jamin, P. Dumas, J. Moncuit, W. H. Fridman, J. L. Teilland, G. L. Carr and G. P. Williams, *Cell. Mol. Biol.*, 1998, **44**, 9.

[16] L. Chiriboga, P. Xie, H. Yee, V. Vigorita, D. Zaron, D. Zukim and M. Diem, *Biospectrosc.*, 1999, **4**, 47.

[17] J. N. Jamin, P. Dumas, J. Moncuit, W. H. Fridman, J. L. Teillaurd, G. L. Carr and G. P. Williams, *Proc. Natl. Acad. Sci.*, 1998, **95**, 4837.

[18] H. Y. N. Holaman, M. C. Martin, E. A. Blakely, K. Bjornstad and W. R. McKinney, *Biopolymers*, 2000, **57**, 329.

[19] M. Jackson and H. H. Mantsch, *Infrared Spectroscopy: Ex Vivo Tissue Analysis, Encyclopedia of Analytical Chemistry*, 2000, **1**, Wiley, Chichester.

[20] M. Jackson and H. H. Mantsch, *Spectrochim. Acta Rev.*, 1993, **15**, 53.

[21] H. L. Casal and H. H. Mantsch, *Biochim. Biophys. Acta*, 1984, **779**, 381.

[22] M. Jackson and H. H. Mantsch, *Spectrochim. Acta Rev.*, 1993, **15**, 53.

[23] W. K. Surewicz and H. H. Mantsch, *Infrared Absorption Methods for Examining Protein Secondary Structure, Determination of Protein Structure in Solution by Spectroscopic Methods*, 1994, **8**, Wiley, New York.

[24] H. Susi and D. M. Byler, *Biochem. Biophy. Res Co.*, 1985, **115**, 391.

[25] P. Colarusso, L. H. Kidder, I. W. Levin, J. C. Fraser, J. F. Arens and E. N. Lewis, *Appl. Spectrosc.*, 1998, **52**, 106A.

[26] M. Jackson, L.-P. Choo, P. H. Watson, W. C. Halliday and H. H. Mantsch, *Biochim. Biophys. Acta*, 1995, **1270**, 1.

[27] K. Z. Liu, M. Jackson, M. G. Sowa, H. Ju, I. M. C. Dixon and H. H. Mantsch, *Biochim. Biophys. Acta,*1996, **1315**, 73.

[28] P. C. Painter, M. M. Coleman and J. L. Koenig, *The Theory of Vibrational Spectroscopy and its Applications to Polymeric Materials*, 1982, John Wiley & Sons, New York.

[29] P. T. T. Wong, R. K. Wong and M. F. K. Fung, *Appl. Spectrosc., 1993,* **47(1)**, 1058.

[30] P. T. T. Wong, R. K. Wong, T. A. Caputo, T. A. Godwin and B. Rigas, *Proc. Natl. Acad. Sci.,* 1991, **88**, 10988.

[31] M. A. Cohenford and B. Rigas, *Proc. Nat. Acad. Sci.,* 1998, **95**, 15327.

[32] M. A. Cohenford, T. A. Godwin, F. Cahn, P. Bhandare, T. A. Caputo and B. Rigas, *Gynecol. Oncol.,* 1997, **66**, 59.

[33] M. A. Cohenford, P. S. Bhandore, B. Rigas and K. Krishman, *Mikrochim. Acta.,* 1997, **14**, 433.

[34] B. R. Wood, M. A. Quinn, F. R. Burden and D. McNaughton, *Biospectrosc.,* 1996, **2**, 143.

[35] B. R. Wood, M. A. Quinn, B. Tait, T. Hislop and M. Romeo, *Biospectrosc.,* 1998, **4**, 75.

[36] B. R. Wood, M. Q. Quinn and D. McNaughton, *Spectroscopy of Biological Molecules*, 1997, Kluwer Academic Publishers, Dordrecht.

[37] B. R. Wood, M. Q. Quinn, B. Tait, M. Romeo and H. H. Mantsch, *Biospectrosc.,* 1998, **4**, 75.

[38] M. Romeo, B. R. Wood and D. McNaughton, *Vibr. Spectrosc.,* 2002, **28**, 167.

[39] L. Chiriboga, P. Xie, V. Vigorita, D. Zarou, D. Zadim and M. Diem, *Biospectrosc.,* 1997, **4**, 55.

[40] M. Diem, M. Boydston-White and L. Chiriboga, *Appl. Spectrosc.*, 1999, **53**, 148.

[41] S. Boydston-White, T. Gopen, S. Houser, J. Bargonetti and M. Diem, *Biospectros.*, 1998, **5**, 219.

[42] M. A. Harthcock and S. C. Atkin, *Appl. Spectrosc.*, 1998, **42**, 3.

[43] P. Lasch and D. Naumann, *Cell. Mol. Biol.*, 1998, **44(1)**, 189.

[44] P. Lasch, W. Haensch, E. N. Lewis, L. H. Kidder and D, Naumann, *Appl. Spectrosc.*, 2002, **56**, 1.

[45] P. Lasch, W. Haensch, D. Naumann and M. Diem, *Biochim. Biophys. Acta,* 2004, **1668**, 176.

[46] B. R. Wood, L. Chiriboga, H. Yee, M. A. Quinn, D. McNaughton and M. Diem, *Gynecol. Oncol.*, 2004, **93**, 59.

[47] P. J. Treado, I. W. Levin and E. N. Lewis, *Appl. Spectrosc.*, 1994, **48**, 607.

[48] E. N. Lewis and I. W. Levin, *Appl. Spectrosc.*, 1995, **49**, 672.

[49] C. L. Bennett, M. Carter, D. Fields and J. Hemandez, *Imaging Fourier Transform Spectrometer. In Proceedings of the Imaging Spectrometry of the Terrestrial Environment*, 1993, Orlando, FL, SPIE 191.

[50] E. N. Lewis, P. J. Treado, R. C. Reeder et al, *Anal. Chem.*, 1995, **67**, 3377.

[51] A. Rogalski, *Infrared Phys. Techn.*, 2002, **43**, 187.

[52] E. L. Dereniak and G. D. Boreman, *Infrared Detectors and Systems*, 1996, John Wiley and Sons, New York.

[53] J. M. Kwiatkoski and J. A. Reffner, *Nature*, 1987, **328**, 837.

[54] R. Bhargava and I. W. Levin, *Fourier transform mid-infrared spectroscopic imaging : microspectroscopy with multichannel detectors, Spectrochemical*

analysis using infrared multichannel detectors, 2005, Blackwell Publishing, Oxford.

[55] H. C. Van-Der-Mei, D. Naumann and H. J. Busscher, *Arch. Oral Biol.*, 1993, **38**, 1013.

[56] M. J. Romeo, M. A. Quinn, R.R. Burden and D. McNaughton, *Biospectrosc.*, 2002, **64**, 362.

[57] N. Stone, C. Kendall, N. Shepherd, P. Crow and H. Barr, *J. Raman Spectrosc.*, 2002, **33**, 564.

[58] C. P. Schultz and H. H. Mantsch, *Cell. Mol. Biol.*, 1998, **44(1)**, 201.

[59] M. Jackson, B. Ramjiawan, M. Hewko and H. H. Mantsch, *Cell Mol. Biol.*, 1998, **44(1)**, 89.

[60] M. Diem, L. Chiriboga and H. Yee, *Biopolymers*, 2000, **57(5)**, 282.

[61] C. P. Schultz, K.-Z. Liu, J. B. Johnson and H. H. Mantsch, *Leukemia Res.*, 1996, **20(8)**, 649.

[62] M. J. Romeo and M. Diem, *Vib. Spectrosc.*, 2005, **38**, 115.

[63] J. R. Mansfield, M. G. Sowa, G. B. Scarth, R. L. Somorjai and H. H. Mantsch, *Anal. Chem.*, 1997, **69**, 3370.

[64] X. Y. Wang and J. M. Garibaldi, *Proceedings of 2nd International Conference in Computational Intelligence in Medicine and Healthcare*, 2005.

[65] L. M. McIntosh, J. R. Mansfield, N. A. Crowson and H. H. Mantsch, *Biospectrosc.*, 1999, **5**, 265.

[66] L. Zhang, G. W. Small, A. S. Haka, L. H. Kidder and E. N. Lewis, *Appl. Spectrosc.*, 2003, **57(1)**, 14.

[67] X. Y. Wang and J. M. Garibaldi, *Eur. J. Inform.*, 2005, **29**, 61.

[68] R. Bhargava, *FTIR Spectroscopic Imaging: An Integrative Paradigm for Biomedical Diagnostics, Presented at the 2nd International Conference of Advanced Vibrational Spectroscopy (ICAVS)*, 2005, Nottingham University, UK.

Chapter Two

Lymph Node Cancer

2.1 Introduction

At present, breast cancer is the most common malignancy found among women, with high death rates reported in the United Kingdom (13,000 p.a.) and the United States of America (40,000 p.a.) [1]. The ability to accurately identify the malignancy is crucial for prognosis and preparation of an effective treatment. The current preoperative imaging methodologies that are used, such as x-ray mammography and ultrasound, rely upon the identification of density changes within tissue. Although such techniques can identify areas of tumour growth in breast, they cannot be used to reliably diagnose whether the tumours are benign or cancerous in nature. The determination of whether a suspicious lesion is malignant necessitates an invasive procedure to obtain a tissue biopsy [2]. However, 70-90% of breast biopsies are later found to be benign after pathological analysis. An additional tool for diagnosis is the assessment of lymph nodes in the ipsilateral axilla. The presence of metastasis is an indicator for local disease recurrence and thus a method for identifying patients who are at high risk of developing disease that could spread throughout the body. The well established procedure to assess lymph node metastases is axillary lymph node dissection (ALND). This involves the surgical removal of all lymph nodes that exist under the arm. However, this is a rather substantial surgical procedure that can lead to several serious side effects, including shoulder dysfunction and lymphoedema [3].

The introduction of mammography screening programmes, together with a greater public awareness of breast cancer, have meant that the majority of patients do not have axillary lymph node metastases at presentation and would not therefore gain benefit from ALND. However it is vitally important to complete accurate staging of the malignancy as to negate the possible risk of the disease spreading to other organs. Intra-operative diagnosis has become increasingly important with the recent introduction of sentinel lymph node biopsy [4]. The sentinel node can be described as any lymph node that has a direct lymphatic connection to the tumour, and would be the first invaded by cancer spreading from the breast (Figure 1). Surgical studies have clearly shown that if cancer cannot be found in the sentinel lymph node, the chance of disease being found further down the chain of lymph nodes that drain the breast is negligible [4]. Thus accurate analysis of the sentinel lymph node can alleviate the necessity to remove all nodes present.

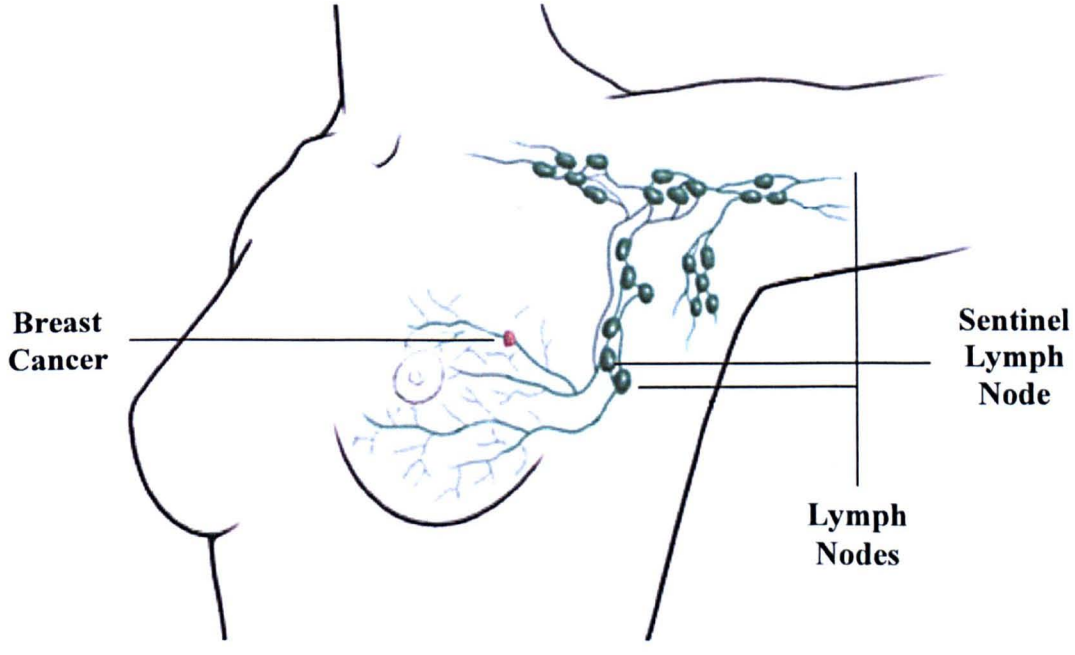

Figure 1: *Typical location of lymph nodes that drain lymph from the breast.[5]*

The Gold Standard of histopathology employed for diagnosis involves the use of formalin (10% solution of formaldehyde in water) fixation, wax embedment and microscope analysis of multiple tissue sections [6]. This procedure can take several days to complete. Alternative techniques have been employed to facilitate faster intra-operative diagnosis of sentinel nodes, including imprint cytology [7,8], and frozen section assessment [9,10]. The processing of results for these techniques is accelerated in comparison to conventional histology involving an analysis time of approximately 30-60 minutes. However, both approaches report wide variation in their sensitivity to detect cancerous lesions, detection levels as low as 44% and as high as 93% when compared against conventional histology [7 – 13]. This would indicate that such methodologies do not solve the lymph node screening problem. In addition, these techniques are both heavily reliant upon the availability of an experienced pathologist. In the UK, this can be a problem in smaller hospitals, where there is a dependence upon a general pathologist to examine these types of samples and can result in lower accuracies than those reported from specialist clinics. However, there is a general lack of consistency between different pathologists that puts the reliability of such intra-operative tools into question.

The problems in lymph node screening highlighted above have resulted in a variety of different spectroscopic methods being investigated for diagnosis. Elastic scattering spectroscopy (ESS) has been used to analyse lymph nodes [14, 15]. This approach is sensitive to the sizes, indices of refraction, and structures of subcellular components (i.e. nucleus, nucleolus and mitochondria) that can change with the progression of malignancy [16]. This method uses short pulses (~ 1μs) of white light (320-920nm) from a xenon lamp via a flexible optical fibre, thus allowing direct

topical access to a tissue sample. The scattered light from the upper layers of the tissue is propagated to a spectrometer and the spectra produced scrutinised. These studies have reported some success in the identification of metastases (cancer) in axillary lymph nodes by adopting the multivariate technique, Linear Discriminant Analysis (LDA) [17]. However, the assessment of lymph nodes using this approach was limited to those that contained heterogeneous patterns of metastatic infiltration [18, 19]. Limitations stem from the collection of individual spectra that may not in all cases be assigned the correct diagnosis. Spectra were collected from a small number of spots across a bivalved lymph node, and histological diagnosis specific to each site was not recorded. Instead, spectra were assigned a diagnosis that represented the overall condition of the node. It should also be noted that the analysis of ESS spectra is an empirical process, and it is not known what features in the spectra reflect histological characteristics in the lymph node.

Recent advances in instrumentation mean that the acquirement of Raman spectra in a clinical setting is possible. This technique has proven applications in the assessment of complex biological systems, with the ability to characterise molecules within biological systems dependent upon the vibrational spectra. For example, Raman spectroscopy has been successfully employed in the identification of silicone inclusions within axillary lymph nodes, excised from patients that have ruptured silicone breast prostheses [20]. This was accomplished by scrutinising the strong Raman band produced by silicone, also commonly used as a laboratory calibration tool. Raman spectroscopy has also demonstrated an ability to distinguish between different tissue types in a number of different organs, including the oesophagus, prostate, bladder and breast [21, 22]. Histopathologic assessment of diseased tissues

is based upon the identification of architectural changes within the cell nuclei, cytoplasm or membrane caused by the progression of malignancy. The ability of Raman spectroscopy to identify small biochemical changes in tissue may allow the potential detection of malignant change before histological features are present. Studies to date have relied upon multivariate analyses to scrutinise Raman spectra produced from biological tissues, such as principle component analysis (PCA), linear discriminant analysis (LDA), and least-squares fitting algorithms. Multivariate techniques have been employed due to the complexity of biological systems, where it is more likely the combined change in the overall biochemical constituents of a cell to be diagnostic, rather than one single biochemical. These types of analyses (see section 4.5) allow the extraction of a number of independently varying components that describe the variations within datasets collected, and can be used to differentiate between tissue types. A recent pilot study using Raman spectroscopic mapping of axillary lymph nodes described the ability of PCA to assess the relative presence of lipids and carotenoids within nodal tissue sections [23]. The spectra were grouped according to histological features, such as histiocytosis, germinal centres, capsule, fatty infiltrate and metastatic carcinoma cells. However, these studies emphasised the need to collect spectra from a large number of patients with varied diagnoses to fully assess the range of pathology present in a node, which will be essential in the development of a robust diagnostic model.

Another factor not addressed by Raman studies, but is also of great importance, is the ongoing debate as to whether histopathology is either aiding or hindering the creation of diagnostic models. It is quite likely that spectroscopic measurements are identifying chemical changes within cells before architectural histological changes

are visually present. However, such claims are held back since current histopathology techniques are used to assign spectra their diagnosis. This could contribute to why such studies have shown varied and imperfect accuracy.

More recently, FTIR spectroscopic mapping has been utilised to assess inguinal lymph nodes [24,25]. This study utilised similar instrumentation as employed in our work, whereby large IR spectral images were collected from lymph node tissue sections by use of a linear array detector system. The large hyperspectral images produced were then scrutinised by unsupervised Hierarchical Cluster Analysis (HCA) to construct pseudo colour maps that were hoped to mimick morphological and histological architecture. However, the reflective substrates utilised in this study appeared to introduce dispersion or reflectance artefacts into the collected IR spectra. A significant shift of the bands to lower wavenumbers was noticed. The amide I and amide II modes were also distorted and displayed an unusual intensity ratio not normally observed for tissue spectra. This artefact occurred predominantly at the edges of tissue and at regions of abnormality (colon adenocarcinoma), which produce glandular metastases with many voids. Such effects dominated the statistical analysis of the IR spectra as the magnitude of the dispersion artefact was greater than the subtle spectral changes that are required for tissue characterisation. Nevertheless, when the dataset was reduced to only include intensities recorded between 1580 – 950 cm^{-1}, a significant improvement upon tissue distinction was made. Unfortunately spectral characteristics that accompanied cluster membership were not detailed.

2.2 Histology of Lymph Nodes

2.2.1 Lymph Node Function

Lymph nodes are small structured organs found in clusters or chains at sites where lymphatic vessels converge and drain an anatomical region, for example in the neck, groins, axillae or para-aortic area. They are responsible for two major functions within the body. Phagocytic cells or macrophages found within the nodes act as non-specific filters of micro-organisms and particulate matter, thus preventing their presence in the general circulation. In addition, they provide an elegant mechanism that facilitates an immune response to an invading exo-genious species. Lymphocytes are allowed to interact with new antigens and antigen presenting cells (APC's) at an interface between the lymph and blood. By recognising passing antigens, lymphocytes within lymph nodes initiate the proliferation of activated cells and therefore amplify the immune response of the body by forming clones of lymphocytes.

2.2.2 Basic Structure

The lymph node is a bean-shaped organ, typically only a few millimetres in length, but may dramatically enlarge when functional demands are increased. They are protected by a fibrocollagenous capsule from which fibrous trabeculae extend into the medulla of the node forming a supportive framework. Afferent lymphatic vessels penetrate the convex surface of the gland and drain lymph into the node, while at the

hilum, a single efferent lymphatic vessel transports lymph to larger collecting peripheral vessels. These larger vessels repeat the filtering process by further transporting lymph to nodes located further along the chain before it is allowed to re-enter the blood stream. Lymph nodes are made up of three main functional inner compartments, as shown in figure 2.

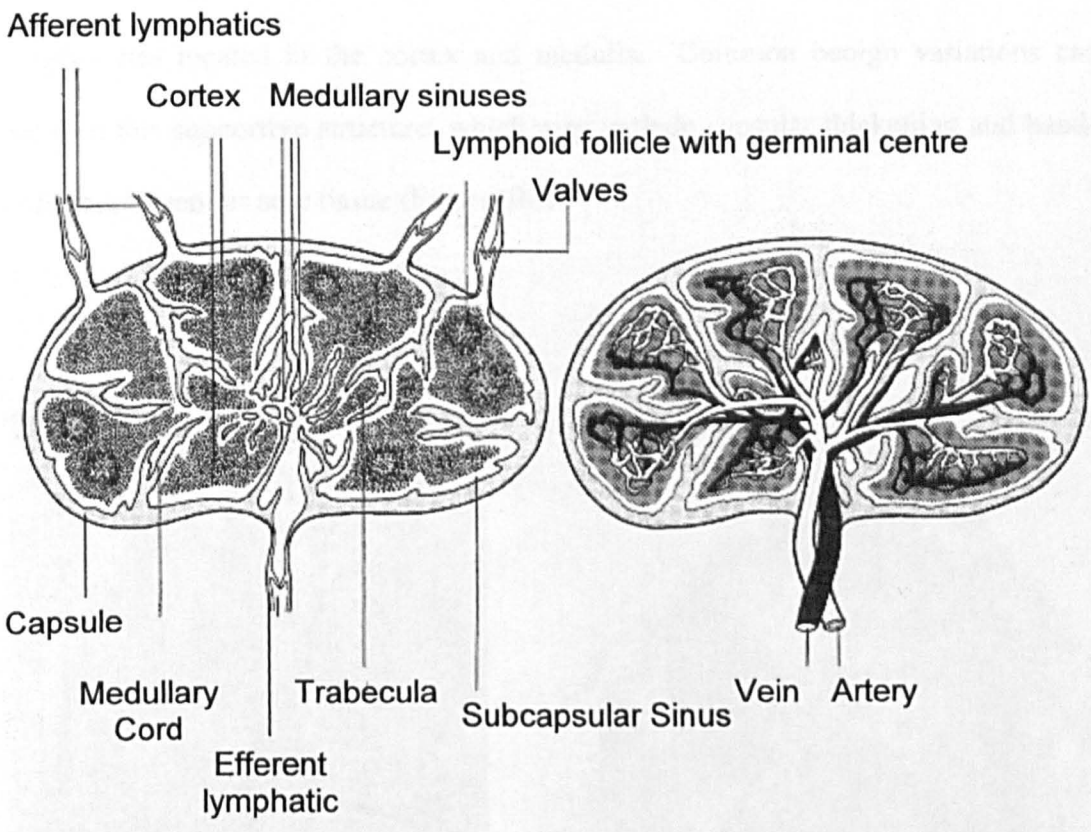

***Figure 2**: A schematic of lymph node anatomy [26].*

These include:

- an intricate network of endothelial-lined lymphatic sinuses that are continuous extensions of the afferent and efferent lymphatic vessels;
- a parenchymal compartment composed of a medulla, paracortex and superficial cortex

- an extended network of small blood vessels that include specialised post-capillary venules that allow circulating lymphocytes to enter the node;

The overall structural integrity of the node is maintained by a framework of dense reticulin fibres, a non-banded form of Type III collagen composed of delicate fibrils around 20nm in diameter, which are linked to the inward reaching trabeculae. These fibres are laid down by fibroblasts (Figure 3a) and act as supporting mesh for lymphocytes located in the cortex and medulla. Common benign variations can occur in this supportive structure, which may include capsular thickening and bands of fibrocollagenous scar tissue (Figure 3b).

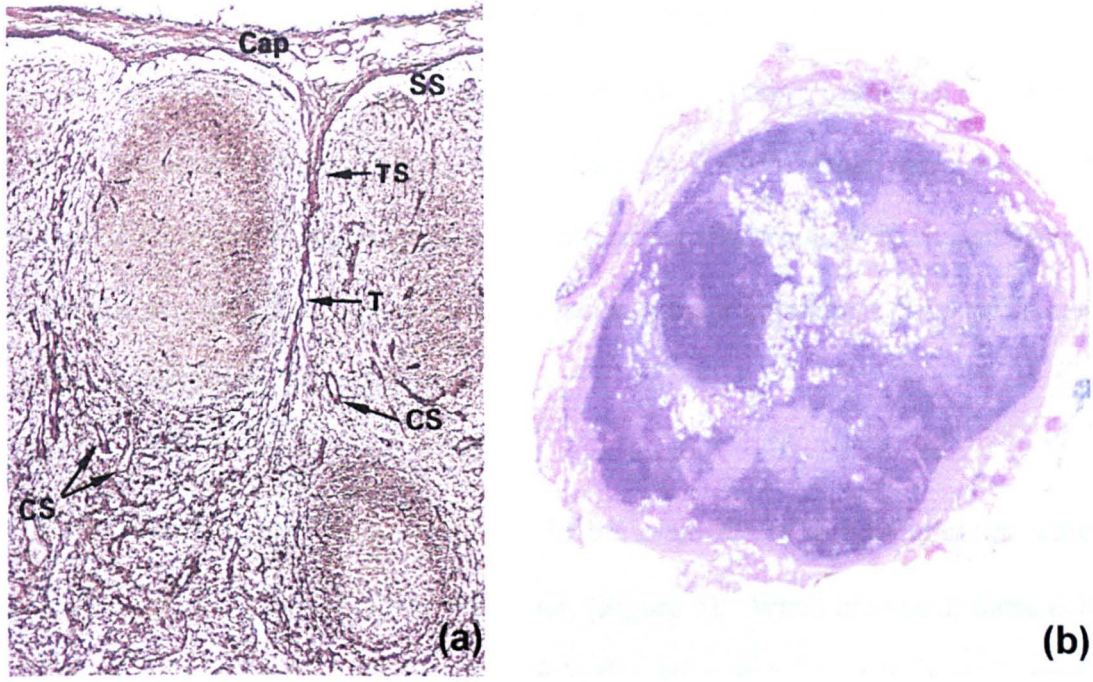

Figure 3: *a) High power magnification image highlighting reticulin mesh that has been stained black in colour. The capsule (Cap), subcapsular sinus (SS), trabecular sinus (TS), trabecula (T) and the cortex sinus (CS) can be visualised. b) Lymph node displaying a thickened capsule and large areas of collagenous scar tissue that has been stained pale pink [26].*

2.2.3 Functional compartments of the lymph node

Antigens, accessory cells and lymphocytes enter the lymph node via the afferent lymphatic system, which pierce the outer capsule of the node and drain into the subcapsular sinus. From here cells can percolate along the cortical sinuses and hence permeate into the superficial cortex or paracortex. However, the majority of lymphocytes enter the node via the blood system. The post-capillary venules are lined with a special type of endothelium bearing lymphocyte homing receptors facilitating their passage into the lymph node. The superficial cortex is dominated by B lymphocytes that form spherical aggregations known as primary follicles. These contain mainly naïve B cells and a small number of memory cells. However, when the follicles are reacting to an antigen presence, only a small number of naïve B cells are present around the periphery and are substituted by a congregation of activated B cells in the centre. Reacting lymphoid follicles such as these are more commonly known as secondary follicles with germinal centres (Figure 4a). Activated B cells proliferate at a very fast rate, quickly producing a large population of identical cells that recognise the same antigen.

The paracortex in contrast is populated in the majority by T lymphocytes, which continuously move in and out of the region (Figure 5). When activated, these cells enlarge to form lymphoblasts that proliferate and create expanded clones. These activated cells are distributed via the circulation and pass to peripheral sites where most of their activity transpires. The final and important production of antibody-secreting plasma cells is thought to occur when both T and B lymphocytes interact

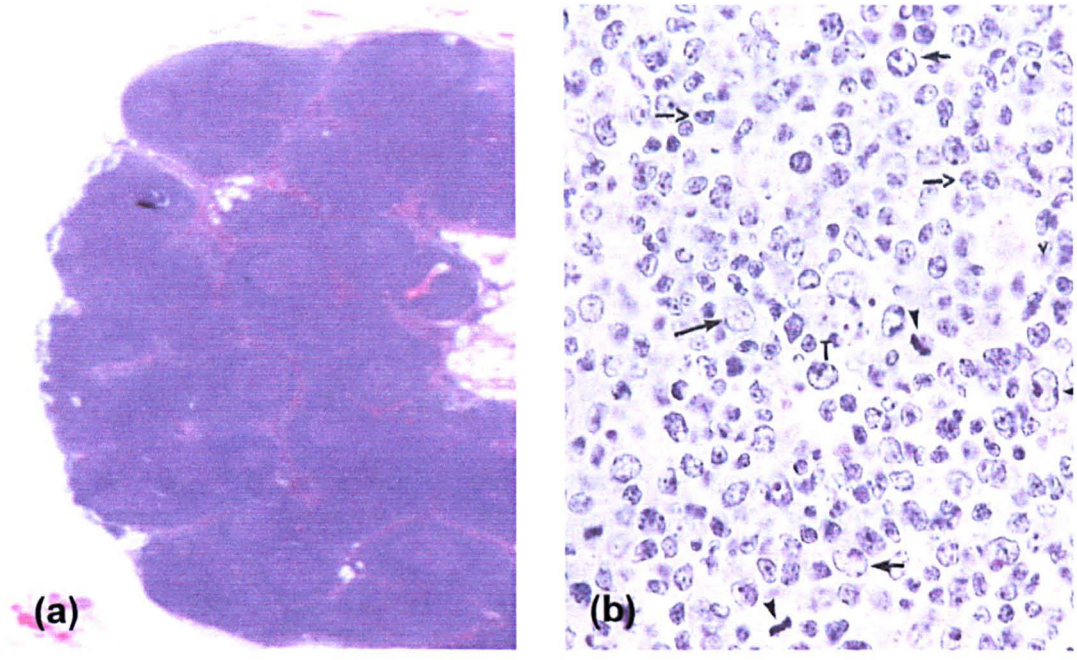

Figure 4: *a) Lymph node with reactive germinal centres. b) High power view of proliferating B lymphocytes (arrows) in a reactive germinal centre; some show mitotic figures (arrowheads) [26].*

within the paracortex. Plasma cells can then migrate directly into the medulla and swiftly pass into the medullary cords, where they can expediently secrete antibodies into the efferent lymph.

The paracortex in contrast is populated in the majority by T lymphocytes, which continuously move in and out of the region (Figure 5). When activated, these cells enlarge to form lymphoblasts that proliferate and create expanded clones. These activated cells are distributed via the circulation and pass to peripheral sites where most of their activity transpires. The final and important production of antibody-secreting plasma cells is thought to occur when both T and B lymphocytes interact within the paracortex. Plasma cells can then migrate directly into the medulla and swiftly pass into the medullary cords, where they can expediently secrete antibodies into the efferent lymph.

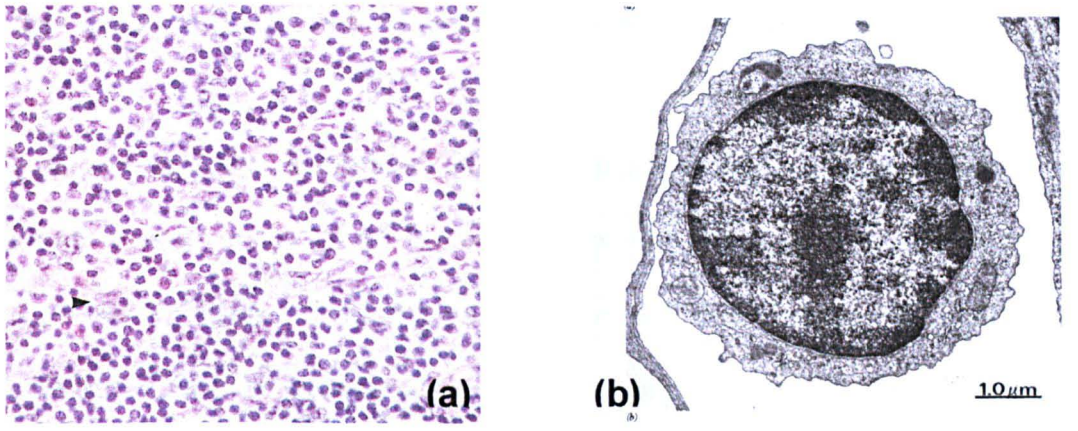

Figure 5: *a) T lymphocytes of the lymph node paracortex with occasional interdigitating cells (arrowhead). b) Electron microscopy image of typical T lymphocyte [26].*

Histological features of acute reactive lymphadenitis, in the absence of lymph node metastasis, are commonly found among breast cancer patients. These reactive changes could result from recent breast biopsies or surgery performed prior to lymph node excision. However, they may also occur due to an immune response against the primary tumour in the breast. Frequently observed changes include large germinal centres that contain multiple mitotic figures and sinus histiocytosis. This latter change occurs when the medullary sinuses hypertrophy and fill with tissue-fixed macrophages called histiocytes (Figure 6).

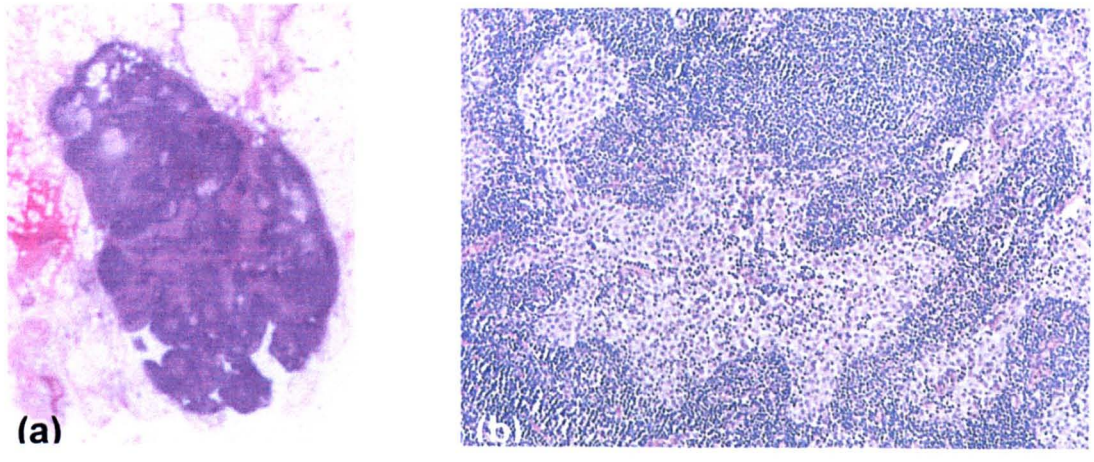

Figure 6: *a) Lymph node with marked sinus histiocytosis. b) High power view of sinus histiocytosis [26].*

2.2.4 Axillary lymph node metastases in breast carcinoma

In the overwhelming majority of cases, architectural features characteristic of the primary tumour in the breast are mirrored in the metastatic invasion of axillary lymph nodes (Figure 7). This is so often the case, that discovery of lymph node metastasis with a contrasting histology to the primary tumour, may indicate the presence of a secondary primary tumour with different invading characteristics. Response within the lymph node to metastatic invasion can involve the enlarging of secondary follicles with reactive germinal centres, sinus histiocytosis and granulation. However, most patients would have previously had a breast biopsy and could therefore display acute lymphadenitis as previously described. A more common and tumour specific change is the formation of collagenous fibrosis around the invading metastatic cells. This reactive change, more commonly termed desmoplasia, can lead to a marked thickening of collagen bands that destroy the

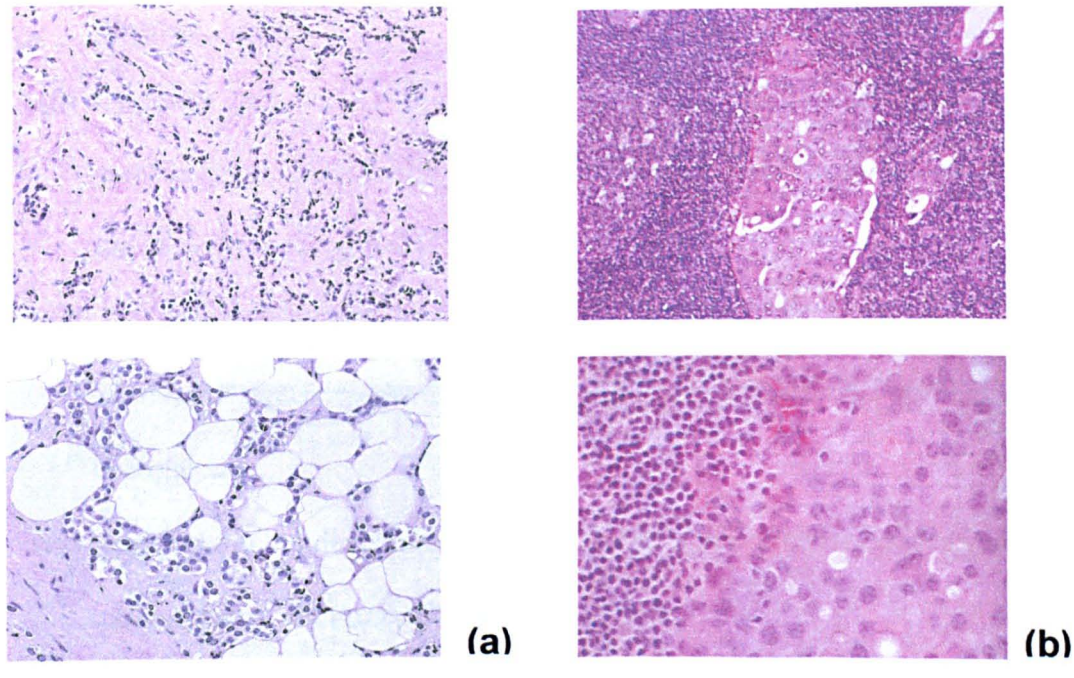

Figure 7: *a) Invasion of an axillary lymph node by lobular carcinoma (above) exhibiting similar histological features to the primary tumour (below). b) Metastatic ductal carcinoma in an axillary lymph node (above) with a similar glandular pattern in the primary tumour (below) [26].*

parenchyma. Early signs of invasive lobular carcinoma appear as single cells or tiny clusters that display a random pattern of invasion within the node. These types of metastatic cells are often hard to discriminate as on occasion they do not display atypical features. In these cases a desmoplastic reaction is absent and large metastases exhibit a signet ring differentiation. However, many breast carcinoma metastases infiltrate the lymph node via the lymphatics. In this pattern, initial metastases are found in the subcapsular sinus, which in time slowly invades deeper into the sinuses before penetrating the parenchyma.

2.3 Results

In this work we have several objectives:

(i) Assess the feasibility of using vibrational spectroscopy for accurate disease diagnosis in lymph nodes.

(ii) Compare and contrast the ability of unsupervised multivariate analysis techniques to discriminate different lymph node tissue types, whether they are diseased or healthy in nature.

(iii) Find the spectral characteristics that are descriptive for each tissue type and search for features which could be utilised for future supervised pattern recognition

(iv) Highlight novel developments we have made for improved classification of tissue spectra via Fuzzy Clustering.

In order to demonstrate this I will:

(i) Compare and contrast multivariate analysis results that were obtained from the examination of an IR spectroscopic dataset collected from one particularly interesting lymph node tissue section.

(ii) Display multivariate IR imaging results from a multitude of different lymph nodes

(iii) Describe experiments undertaken that coalesce IR spectra collected from different lymph nodes for a combined tissue classification

(iv) Chart the novel developments made during this study for improved clustering analysis. These will be described via experiments that were undertaken upon collected tissue spectra datasets.

2.3.1 Evaluation of an Axillary Lymph Node Tissue Section using IR Multivariate Imaging

In this section a multitude of different unsupervised multivariate imaging techniques have been applied to infrared micro-spectral data collected from a positive lymph node tissue section. These techniques include PCA, MCR, FCM Clustering, and a newly developed PCA-FCM Clustering hybrid. Results from the multivariate imaging techniques are assessed via image quality and comparison to conventional histopathology.

2.3.1.1 Histological architecture of Lymph Node

The H&E stained parallel tissue section used for infrared analysis is shown in Figure 8 and allows the main structure of the node to be identified. An IR image was collected from a particularly interesting site on the lymph node where several different types of tissue existed, but more importantly displayed areas of both cancerous invasion and healthy nodal tissue. Figure 8b shows this examined region at higher magnification and allows the easy identification of the surrounding capsule, cortex and invading breast cancer. In the centre of the cortex, with a lighter pigmentation, is a stimulated proliferating secondary follicle or germinal centre. Reticular cells that extend into the sinuses can also be seen and characteristically form a delicate network between the capsule and trabeculae. A small pocket of fatty tissue that normally surrounds the lymph node was also found at the top left corner of the imaged area. This has unfortunately been missed in our H&E stained image.

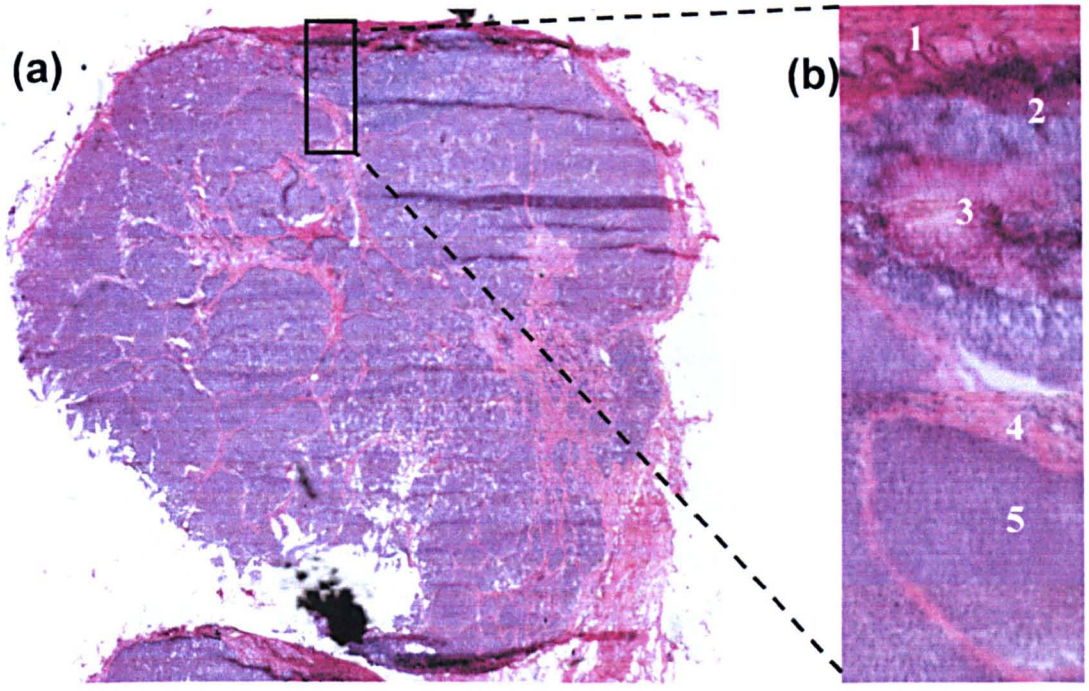

Figure 8: *a) Photomicrograph of the H&E stained parallel lymph node tissue section used for IR analysis. b) IR imaged area at high magnification showing different tissue types. (1) Capsule, (2) cortex, (3) secondary follicle, (4) reticular cells and (5) invading breast cancer tissue. The area (306 x 956 μm) was mapped using a pixel size of 6.25 μm collecting a total of 13,910 individual IR spectra.*

2.3.1.2 PCA Results

The collected spectral IR dataset was first subjected to PCA and primary results are shown in Figure 9. The three dimensional scatter plot shown in Figure 9 displays all spectra in the dataset projected onto the first 3 PCs. Although there is some separation of spectra along these new orthogonal axes, there is no clear clustering of spectra into separate groups in this new uncorrelated multi-dimensional space. The cumulative percentage variance plot in Figure 10 indicates that approximately 97% of the total variance in the dataset is now comprised within the first 5 principal components. Therefore the overwhelming majority of information regarding the patterns within the dataset will be described by these first 5 principal components.

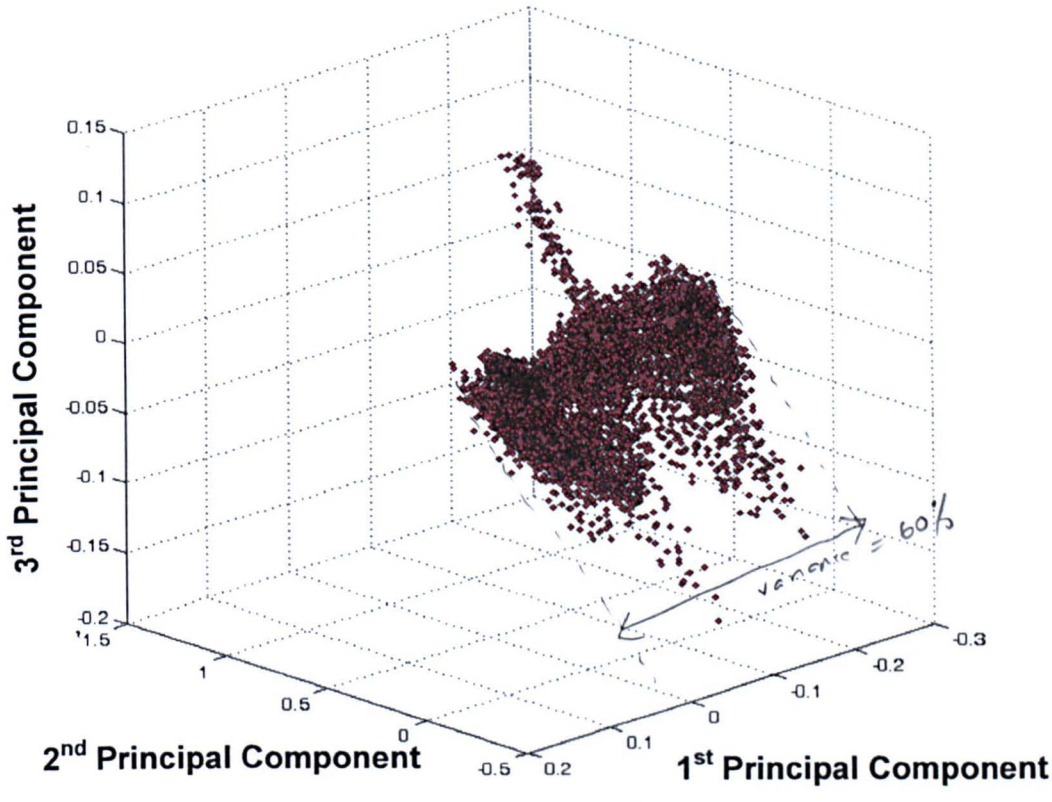

Figure 9: *A three-dimensional scatter plot of the tissue section spectra projected onto the first 3 PCs.*

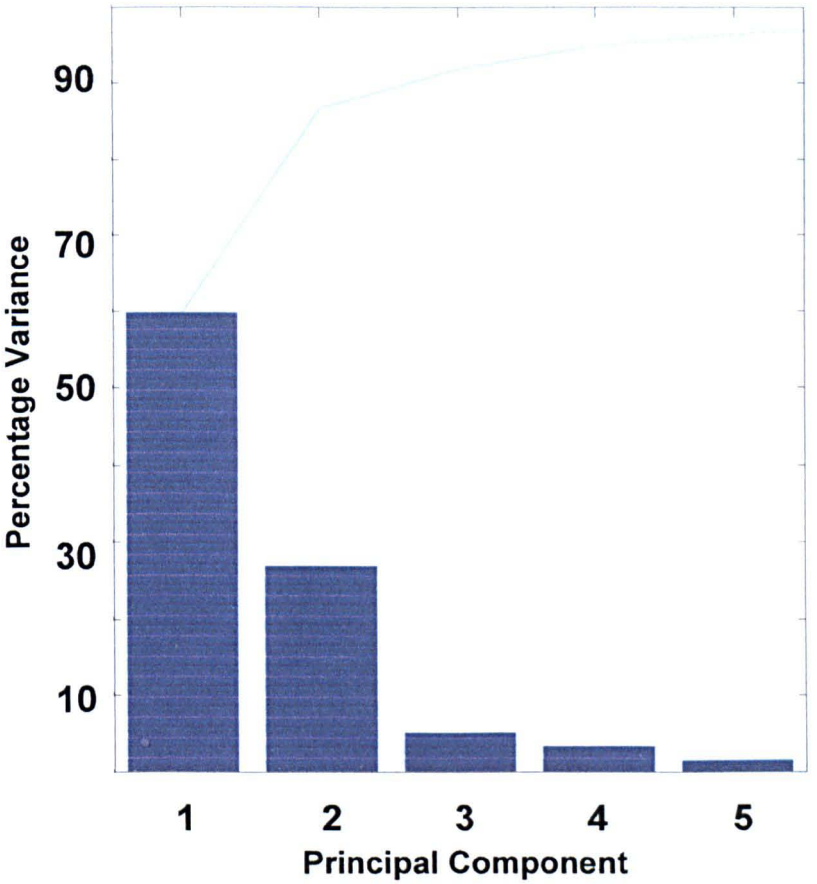

Figure 10: *Combined individual and cumulative percentage variance plot for the first 5 principal components*

False colour weighted images were then created for each of these first 5 PCs, and are displayed in Figure 11. The first PC image shown in Figure 11b clearly differentiates between the fatty tissue located at the top left hand corner of the image and the remaining other tissue types. The second PC image in Figure 11c demonstrates a greater discrimination between tissue types. Both the germinal centre and cancerous tissue show a strong correlation to this component with a red pigmentation. In contrast, the capsule and the central region of reticulum, both fibrocollagenous tissues, display a more negative correlation and are highlighted by a cyan pigmentation. The remaining reticulum and normal cortex tissue that surrounds the germinal centre are marked by a yellow colour. Scrutinising the third PC image,

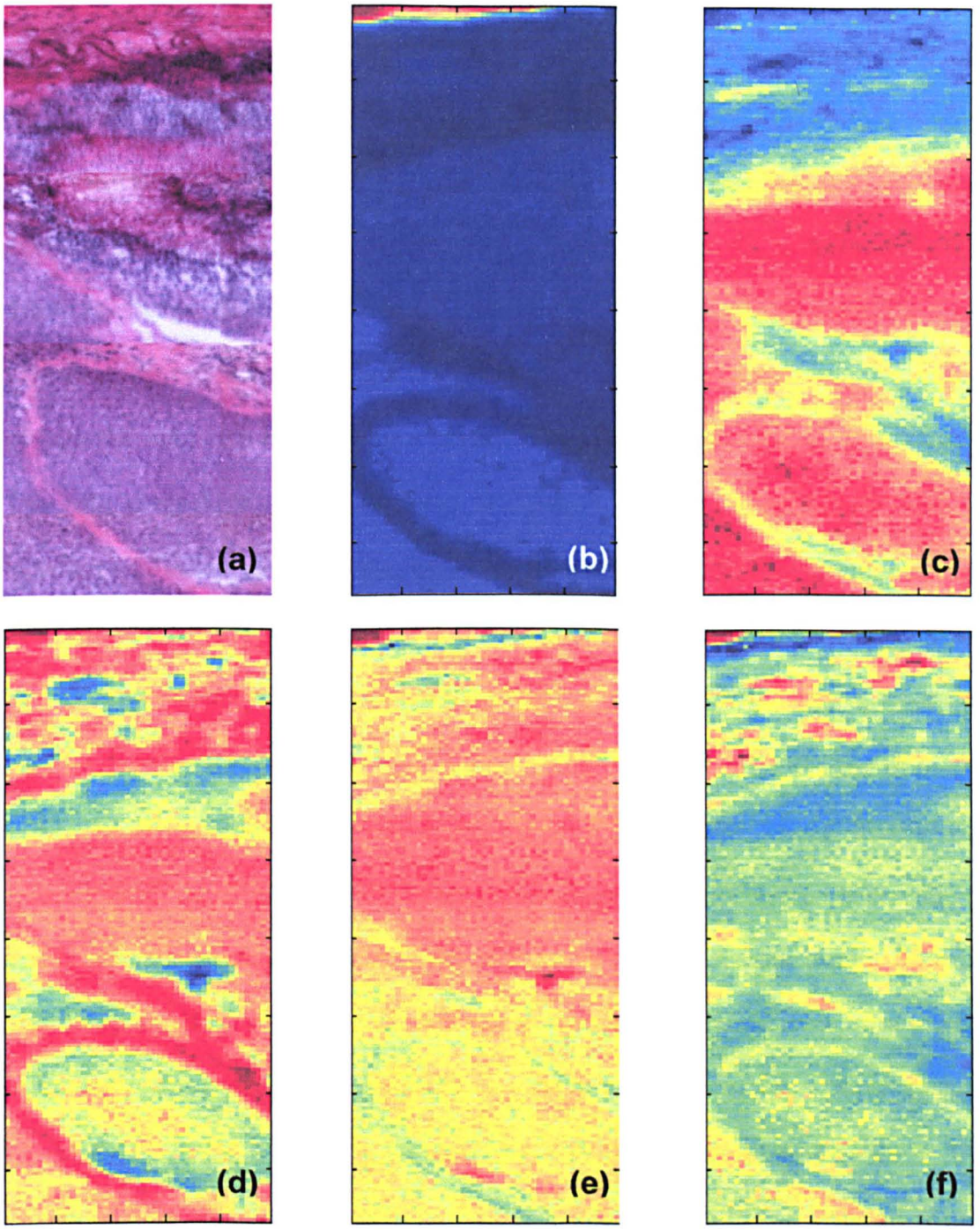

Figure 11: *IR imaging of a lymph node tissue section by PCA. (a) H&E stained image of tissue section. (b) – (f) False colour weighted images for principal components 1 - 5 respectively. Colour scale ranges from red indicating spectra that are very similar to that PC and blue which are greatly dissimilar.*

shown in Figure 11d, the two types of fibrocollagenous tissues are strongly correlated to this component displaying a red pigmentation. However, the germinal centre also displays a similar correlation to this component. The image does however provide a small amount of contrast for the normal cortex tissue that surrounds the germinal centre, highlighted by a cyan colour. It is apparent in the fourth PC image, displayed in figure 11e, that some contrast is being made between the cancerous (yellow colour) and remaining tissue types. However, within regions of both the reticulum and capsule a similar correlation to this component is found. The fifth component image displayed in figure 11f does not reveal any further beneficial information about the tissue section, and provides no useful contrast between the tissues types present. Although these false colour weighted images have enabled some differentiation between tissue pathologies, PC1 is the only component that can describe a single tissue type. Consequently the future effectiveness of using a Linear Discriminant Analysis based upon PCA results would clearly be compromised for clear and distinct tissue pathology. But most importantly, there was no component that could solely describe the invading cancerous tissue. To rule out the possibility that spectral differences between the cancerous tissue and remaining tissue types were statistically very small, component images were created for the first 25 PCs, accounting for well above 99.99% of the total variance in the dataset. These unfortunately gave no further helpful discrimination between tissue pathology.

2.3.1.3 MCR Results

Before the dataset was subjected to MCR analysis, a number of indicator functions developed by Malinowski [23] were used to help ascertain the optimal number of

factors that best describe the data. These calculations indicated that either a 3 or 5 component system would best describe the patterns found within the dataset. Taking into account the previous PCA analysis, the recommended amount of factors appeared reasonable when considering such a high percentage (97%) of the original variance was composed by the first 5 PC's alone. Thus both a 3 and 5 component MCR analysis was then subjected upon the dataset. The false colour weighted images constructed from these analyses are shown in Figure 12.

The imaging results constructed from a 3 component MCR analysis are shown in Figures 12a – c respectively. By use of a colour ranking, pixels on the constructed image now reflect the intensity or correlation of each spectrum to that component. Examining the first component image in Figure 12b, this clearly differentiates the cortex tissue, whether healthy or cancerous in nature (red colouration). The second component image, shown in Figure 12c, alternatively marks a region at the top left hand corner of the tissue section where fatty tissue exists. This component also provides a small amount of contrast between the cancerous and healthy cortex tissue (light blue colouration). Finally the third component image shown in Figure 12d clearly discriminates the capsule and reticular tissues. Although this analysis has provided some discrimination between the three main types of tissue, individual components that describe the further subsets of theses tissues (reticulum, secondary follicle and cancerous cortex) were not achieved.

Imaging results constructed from the subsequent 5 component MCR analysis are displayed in Figures 12e – i respectively. The first component image constructed from the analysis, shown in Figure 12e, now importantly provides more distinct contrast between the cancerous cortex (yellow colouration) and remaining tissue

types. In contrast, the second component image displayed in Figure 12f now provides a component that more exclusively describes the capsule tissue. The third component image shown in Figure 12g again highlights the more collagenous tissue types that include the capsule and reticulum. A small amount of contrast is also provided enabling the visualisation of the secondary follicle (yellow colouration). Alternatively, the fourth component image shown in Figure 12h provides discrimination between the cortex and remaining tissue types. The fifth and final component image displayed in Figure 12i again displays the fatty tissue pocket at the top left of the examined area. Although this 5 component analysis has provided an additional component that more discretely highlights the cancerous cortex region, individual components that provide sole discrimination of the secondary follicle and reticular cells is not apparent. To rule out the possibility that an increased factor number would extract individual components characteristic for all the tissue types, MCR analyses with up to 15 factors were undertaken but provided no additional beneficial information about the tissue section.

2.3.1.4 FCM Clustering Results

The FCM clustering results from the collected spectral dataset are displayed in Figure 13 as false colour images, where a given colour in each image describes spectra that were grouped together in one cluster. It can be seen that as the amount of clusters has been subjectively increased from 2 – 5 (Figure 13b – e), the amount of tissue types that can be discriminated is increased. When comparing these clustering results to the H&E stained parallel section in Figure 13a, the FCM image created for 5 clusters displays a good resemblance given that this is from an adjacent tissue section and small morphological changes are likely. Each colour within the image

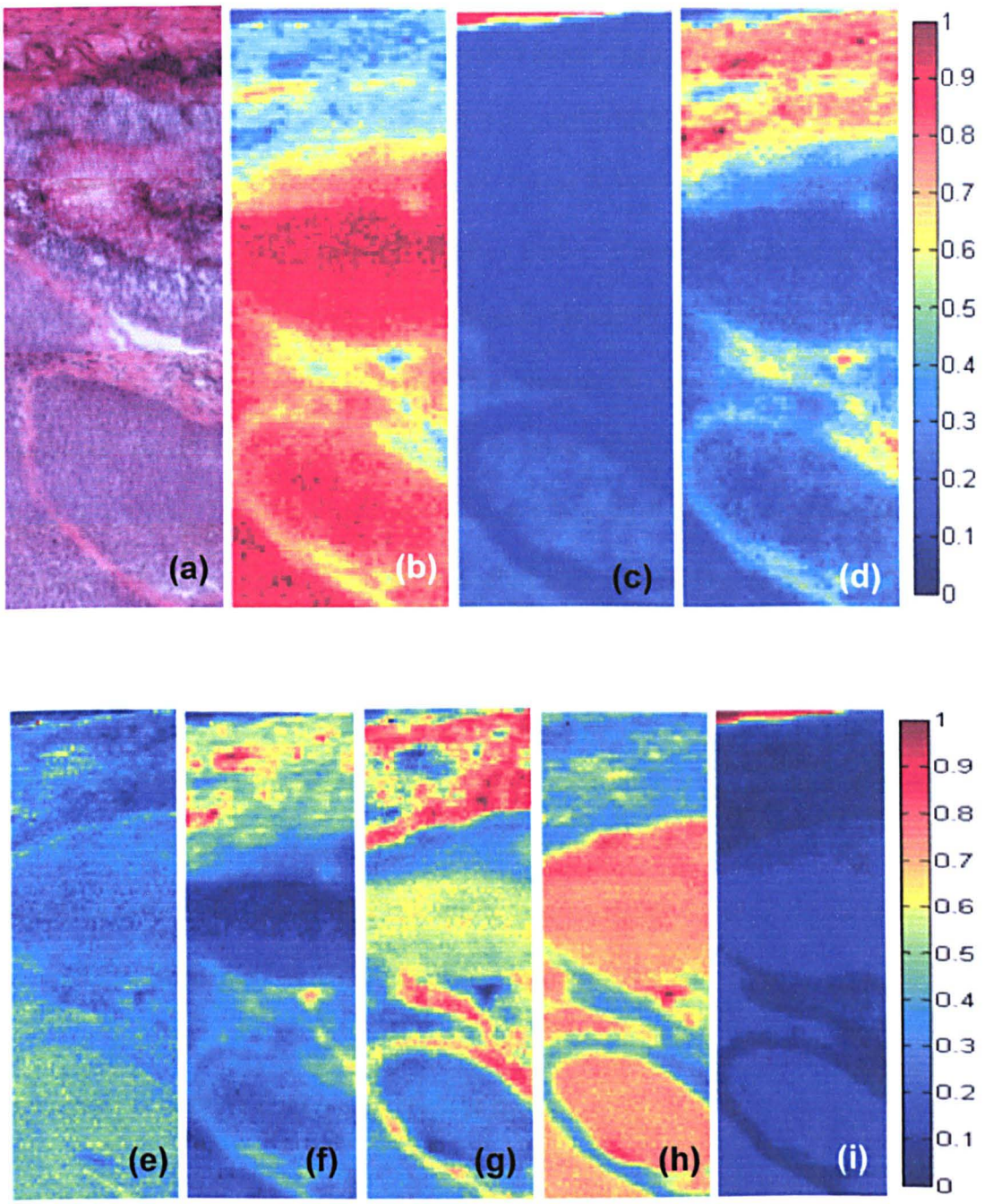

Figure 12: *IR imaging of a lymph node tissue section via MCR Analysis. (a) H&E stained image of the tissue section. (b) – (d) False colour weighted images constructed from a 3 component MCR analysis of the dataset. (e) – (i) False colour weighted images constructed from a 5 component MCR analysis of the dataset. Note that the colour scale ranges from red indicating spectra that are very similar to that component, and blue that is greatly dissimilar.*

can now be assigned to a specific tissue type. Orange pixels describe the capsule, green the reticulum, maroon the healthy cortex surrounding the germinal centre in dark blue, and finally the invading cancerous breast tissue is described by a light blue colour. The only misclassification is of spectra that originate from fatty tissue located at the top left corner of the image. These have been incorrectly grouped into the same cluster as the invading cancerous tissue. Correct clustering of fatty tissue spectra into a single group was not achievable via our FCM analysis. This is a direct consequence of their position in multi-dimensional PC space, and will be examined in greater detail in the discussion. As the amount of clusters is further increased from 6 – 8 (Fig 13f – h), these main tissue types are then further subdivided. The capsule and reticulum begin to show shared clusters and the formation of a lining that surrounds these tissues. This is an understandable result as are very similar in biochemistry, both being fibrocollagenous types of tissue. The invading cancerous tissue also begins to display a second cluster that may describe tissue with a different degree of malignancy, not recognised via conventional histology. When cluster numbers were again increased (>8), no further beneficial tissue discrimination could be made, with images becoming needlessly complex and hard to interpret.

2.3.1.5 PCA – FCM Clustering Hybrid Results

Finally, the collected spectra were subjected to combined PCA – FCM clustering analysis, where the dataset is initially compressed via PCA to its first 10 PC's and then clustered via FCM methodology. Although this algorithm consecutively performs two different multivariate analyses, the total computation time is

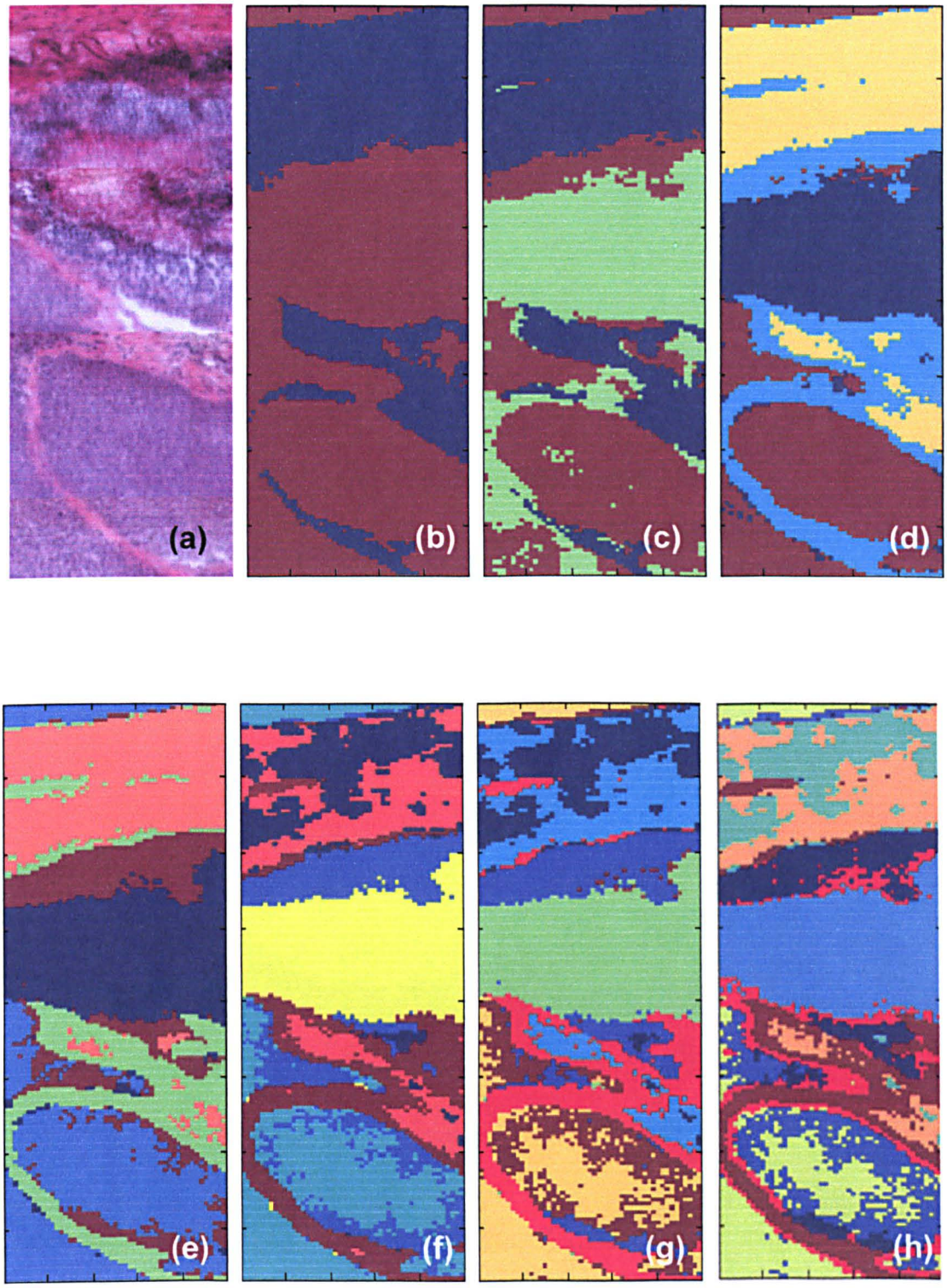

Figure 13: *IR imaging of a lymph node tissue section via FCM Clustering. (a) H&E stained image of the tissue section. (b) – (h) False colour images created using FCM Clustering Analysis results. Note cluster numbers were subjectively increased from 2 – 8. Pixels with the same colour in each image are spectra that were grouped together into the same cluster.*

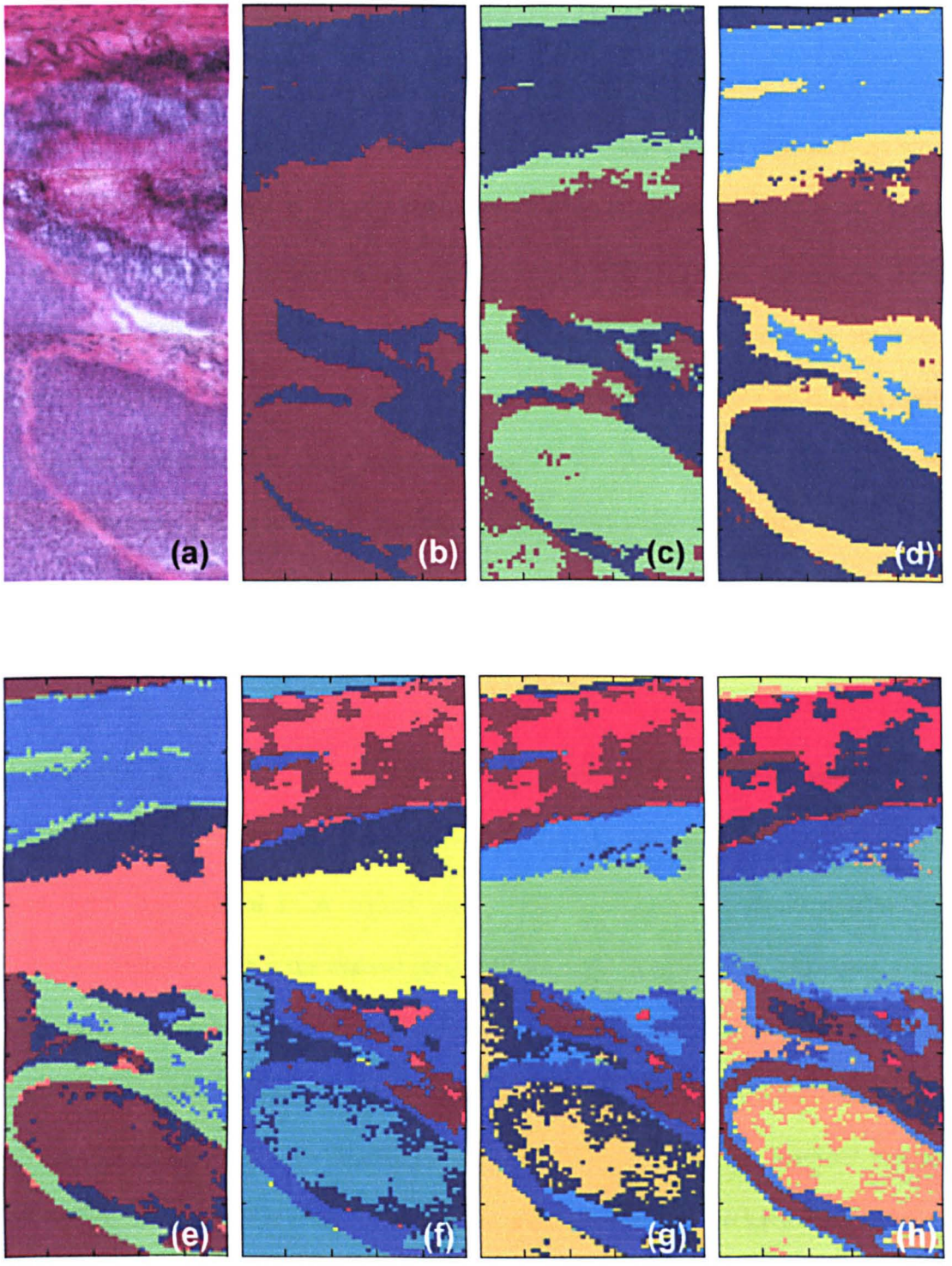

Figure 14: *IR imaging of a lymph node tissue section via PCA-FCM Clustering. (a) H&E stained image of the tissue section. (b) – (h) False colour images created using PCA-FCM Clustering Analysis results. Note cluster numbers were subjectively increased from 2 – 8. Pixels with the same colour in each image are spectra that were grouped together into the same cluster.*

significantly faster than traditional FCM analysis. This is a consequence of the dataset now only being described by 10 dimensions rather than a number defined by the amount of data points in the collected spectra. Results from the analysis were again visually displayed as false colour images and are shown in Figure 14. When comparing these PCA-FCM images directly with those created via conventional FCM clustering, no significant or worrying loss of image quality can be observed. Only a very few pixels in each image have been classified differently. This quite clearly demonstrates that data compression used in a correct statistical fashion can be an effective tool for reduced computation requirements and analysis times.

2.3.1.6 Spectral Characteristics of Tissue Types

All results from the 5 PCA-FCM cluster analyses, which displayed a good resemblance to the H&E stained parallel tissue section are shown in Figures 15 – 17. The three dimensional PCA scatter plot shown in Figure 15b again displays all spectra contained within the dataset projected onto the first 3 PC's. However, these have now been coloured according to the cluster they belong to. This suggests that PCA analysis alone would have difficulty discriminating the different tissue types present, as spectra are very closely packed together in PC space. Another distinct advantage of FCM clustering is that mean average spectra for each cluster in an analysis can easily be calculated and used to help interpret the biochemical differences that are occurring between them. The mean spectra calculated for the 5 cluster analysis are displayed in Figure 16. At first glance, spectra from the different tissue types appear to be very similar, with the most discernable changes occurring within the 1800 – 700 cm^{-1} region. When examining this region in greater detail,

Figure 17, two main spectral profiles are revealed. The first profile best describes the capsule and reticular cell mean spectra. These tissues exhibit strong overlapping collagen bands in the 1180 – 1380 cm^{-1} region, with peaks occurring at 1205, 1232, 1280 and 1335 cm^{-1} respectively. This series of peaks are characteristic of the complex vibrations produced by amide III bending and wagging modes in proteins [29-31]. An additional collagen peak is found at 1448cm^{-1}, and is distinctly more intense in these tissue types. A marked reduction in the symmetric vibration characteristic of phosphodiester groups in nucleic acids located at 1085 cm^{-1} is clearly distinguishable and likely to reflect the reduced nucleic acid concentration in these cell types. Another definable feature of these tissues is the position of the amide I band that occurs later than other tissue types at 1664 cm^{-1}. The capsule and reticulum spectra are only discernable via small peak intensity variations across the spectrum and a change in their amide II / amide I ratio.

The second spectral profile alternatively describes the cortex, germinal centre and cancerous tissue. These spectra display a pronounced symmetric phosphodiester vibration at 1085 cm^{-1}, and a more distinguishable antisymmetric vibration at 1240 cm^{-1}. Previous studies have indicated that the relative intensity of theses bands can be descriptive to a cells divisional activity [32-34]. Our results agree with this finding, whereby cancerous and secondary follicle cells that proliferate at a fast rate are observed to have high intensity for this band. These spectral changes indicate an overall increase in the nucleic acid concentration of these tissue types. Contributions from collagen to these spectra are reduced allowing shoulder peaks at 1468 and 1408

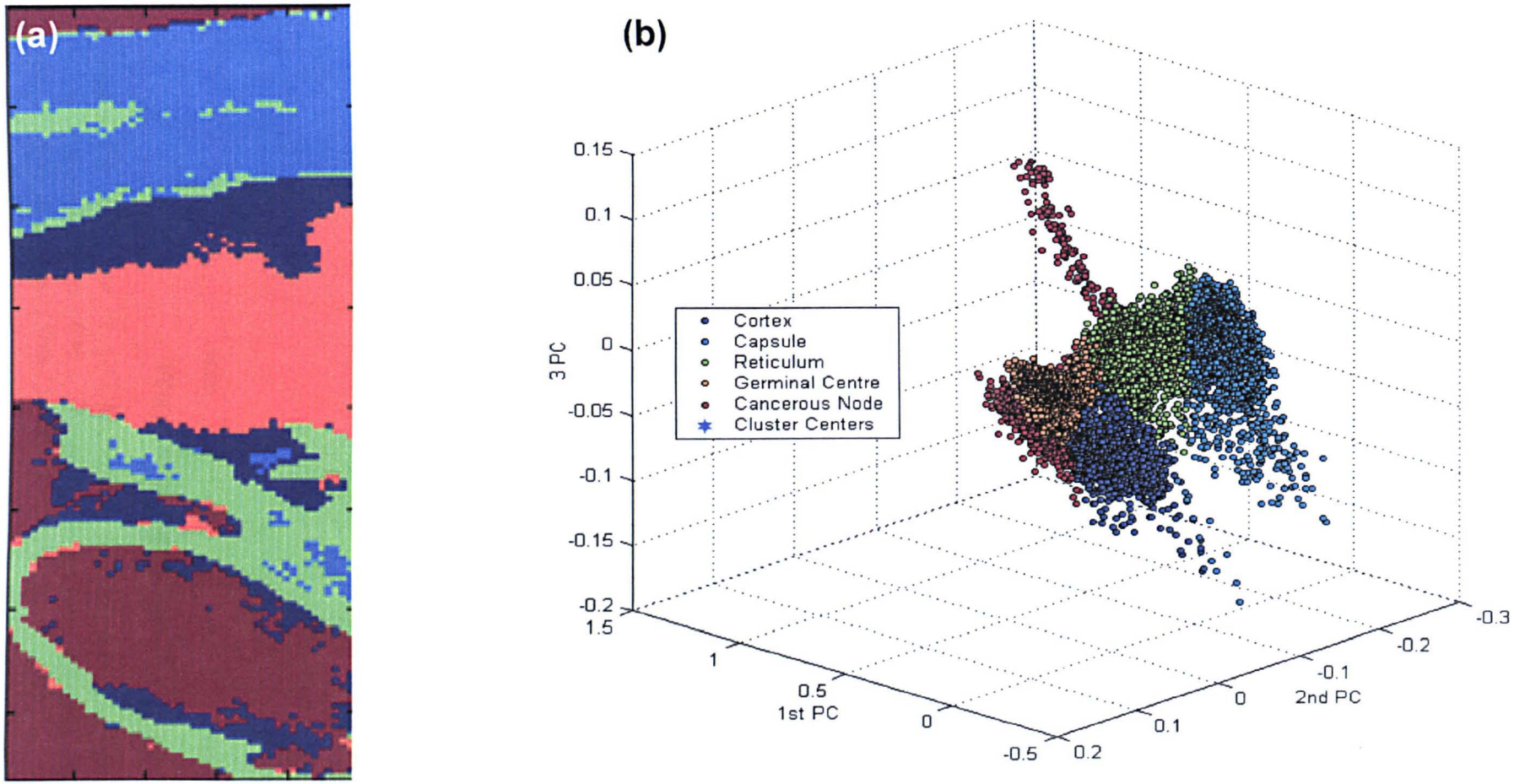

Figure 15: *5 Cluster PCA-FCM Analysis Results (a) False colour image. (b) Three-dimensional scatter plot of tissue section spectra projected onto the first 3 PCs. Note spectra are coloured according to cluster membership.*

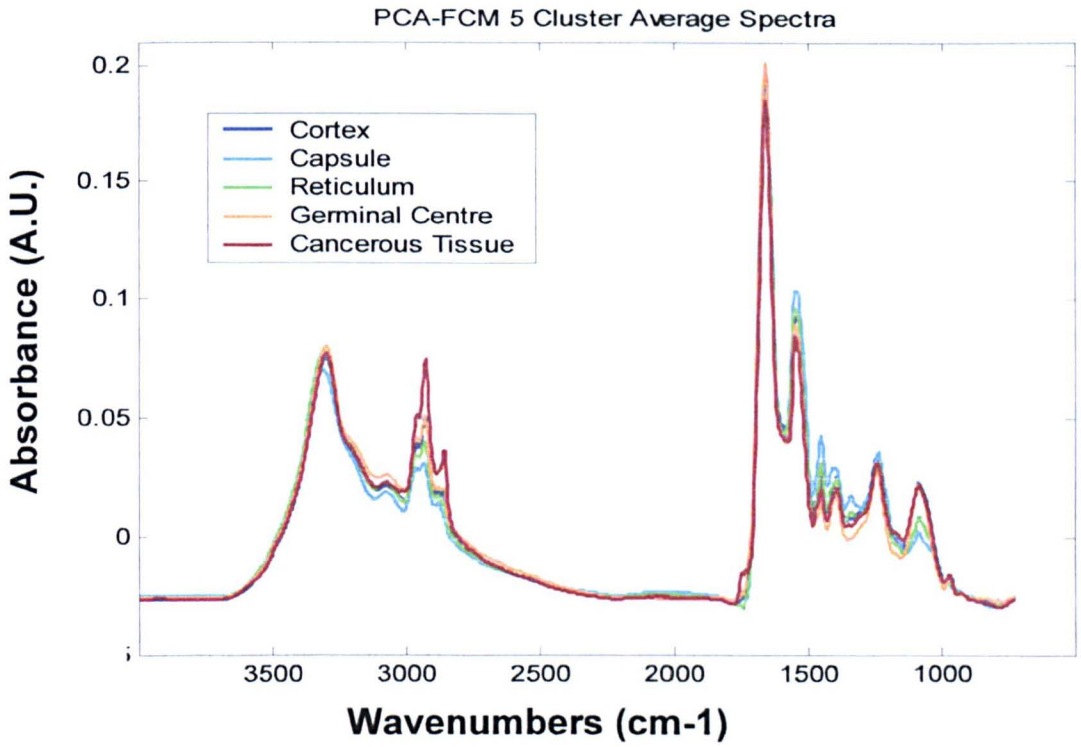

Figure 16: *5 Cluster PCA-FCM Analysis Results. Mean average spectra for each cluster in the analysis.*

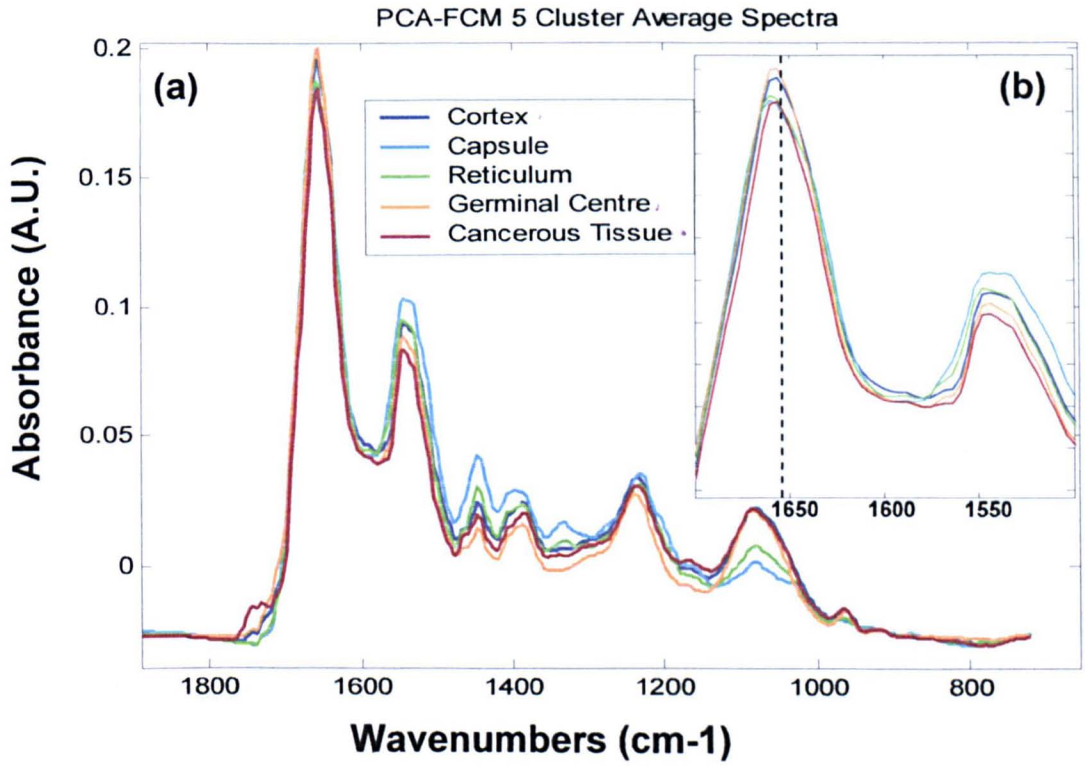

Figure 17: *5 Cluster PCA-FCM Analysis Results. (a) Spectral window displaying mean spectra between 1800 – 720cm^{-1}. (b) Spectral window displaying the amide I and II spectral region (1700-1500cm^{-1}).*

cm^{-1} to be revealed, likely attributable to CH_2 scissoring vibrations in lipids and methyl deformations in lipids. Again these three tissue types have only small peak intensity changes across the majority of the spectrum, with their most discernable differences occurring in the Amide I – Amide II region. Spectra representative of the cancerous tissue showed a significant reduction in the amide II/amide I intensity ratio when compared to the healthy cortex tissue. This observation is in agreement with previous studies examining cervical tissues [35,36], where a reduction in this ratio was again identified in diseased tissue. Interestingly, our results further show that spectra originating from the secondary follicle have an even greater reduction in their amide II/ amide I ratio. It should finally be noted that the mean spectrum for the cancerous tissue also exhibits a small lipid peak at 1744 cm^{-1}. This peak is attributed to the ν(CO) band of the ester group within lipids, and could be an artefact introduced by the misclassification of fatty tissue spectra into this cluster. When taking into account the large protein content of animal cells, it is not surprising that tissue differentiation has been dominated by changes occurring within this spectral region. The band position and intensity of these peaks can indicate the relative protein concentration and their secondary structure, being the summation of several underlying and overlapping bands.

2.3.1.7 Multivariate Analysis Discussion

MCR analysis was clearly the least computationally expensive (Table 1), providing contrast for a majority of the tissues types present. PCA also took a relatively small amount of time to complete its analysis. However it showed poor tissue differentiation, with fatty tissue accounting for a very large amount of the total

variance contained within the dataset (approximately 60%). This dominance of the variance is caused by a dramatic difference in fatty tissue spectral characteristics.

Techniques	Computation Times (mins)
PCA	1
MCR	0.5
FCM	28
PCA-FCM	2.5

Table 1: *Computation time comparison between PCA, MCR, FCM and PCA-FCM analysis techniques using the same computational hardware.*

These large spectral differences are likely to have caused the analysis to be less sensitive to the small spectral differences that occur between the remaining tissue types. Although taking a greater amount of time to complete, the FCM analysis displayed a marked improvement in tissue discrimination. All tissue spectra could be clustered into their histological groups apart from spectra originating from the fatty tissue. The reason for this incorrect clustering can be explained by the examination of the spectra in multi-dimensional PC space. In Figure 18, the original three-dimensional PC plot shown in Figure 15b has been rotated to best describe the differences between the outlier fatty tissue (encircled) and remaining tissue spectra. Previously work has shown that when using the Euclidean distance to define fuzzy cluster membership values, inefficient clustering can occur when the shape of the data points in multi-dimensional space is not ideal (spherical) [37-40]. In our dataset the first PC is descriptive of the lipid content in the fatty tissue. The small amount of spectra collected from this region on the tissue section display a very large natural variation in the intensity of these lipid peaks. This has caused the fatty tissue spectra to be sparsely distributed along this PC axes and therefore render the FCM clustering

less efficient. Unfortunately this has led to the persistent mis-clustering of the fatty tissue spectra into the same cluster as the cancerous node spectra (dark-red). A possible solution to this problem could be to consider these spectra as a separate cluster before multivariate analyses were carried out. This could be achieved by the creation of a filtering test that seeks out fatty tissue spectra, looking for large lipid peak intensities characteristic of this tissue.

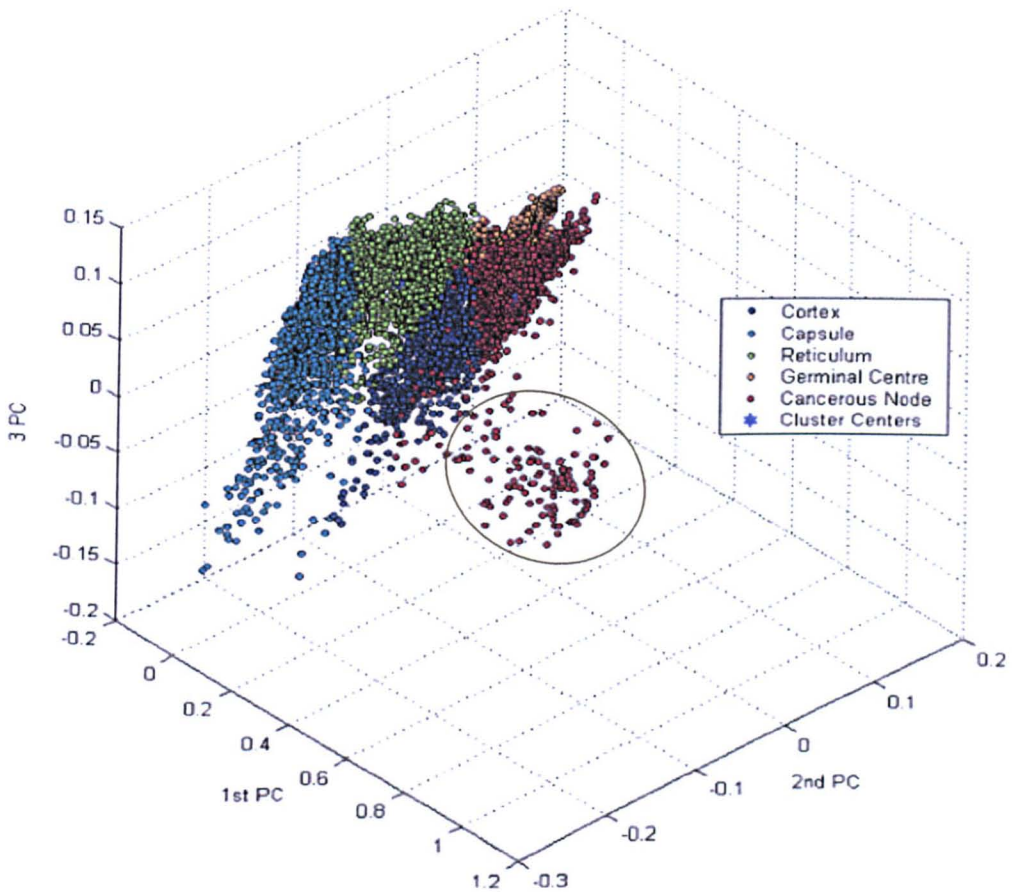

Figure 18: *Rotated three-dimensional scatter plot of tissue section spectra projected onto the first 3 PCs. Outlier fatty tissue spectra are encircled.*

Finally, the combined PCA – FCM analysis showed both the enhanced tissue discrimination achieved via FCM analysis, but also a greatly improved computation

speed without a significant loss of information from the original dataset. Unlike Hierarchical Clustering Analysis (HCA), where the memory requirements and computation times are excessively large [24,32,33,41], PCA – FCM clustering is an exciting technology that can allow high quality analysis in dramatically reduced times.

2.3.2 The Characterisation of a Catalogue of Axillary Lymph Node Tissue Sections by use of Infrared Multivariate Imaging

During this study a multitude of different axillary lymph node tissue sections with contrasting histological architectures were examined via IR mapping. The micro-spectral datasets produced were then further scrutinised by three different types of unsupervised multivariate imaging techniques. These include PCA, MCR and a newly developed PCA-FCM clustering hybrid. In this section we describe the results produced by direct comparison to conventional histology, and thus assess their ability to distinguish the contrasting tissue types that exist within the sections analysed.

2.3.2.1 Axillary Lymph Node LNII7

The first tissue section in our library (named LNII7) was cut from a diseased lymph node that displayed multiple areas of invading cancerous tissue. Both the white light image of the tissue section and a photomicrograph of its parallel H&E stained section are shown in Figures 19a and 19b respectively. An infrared micro-spectral map was

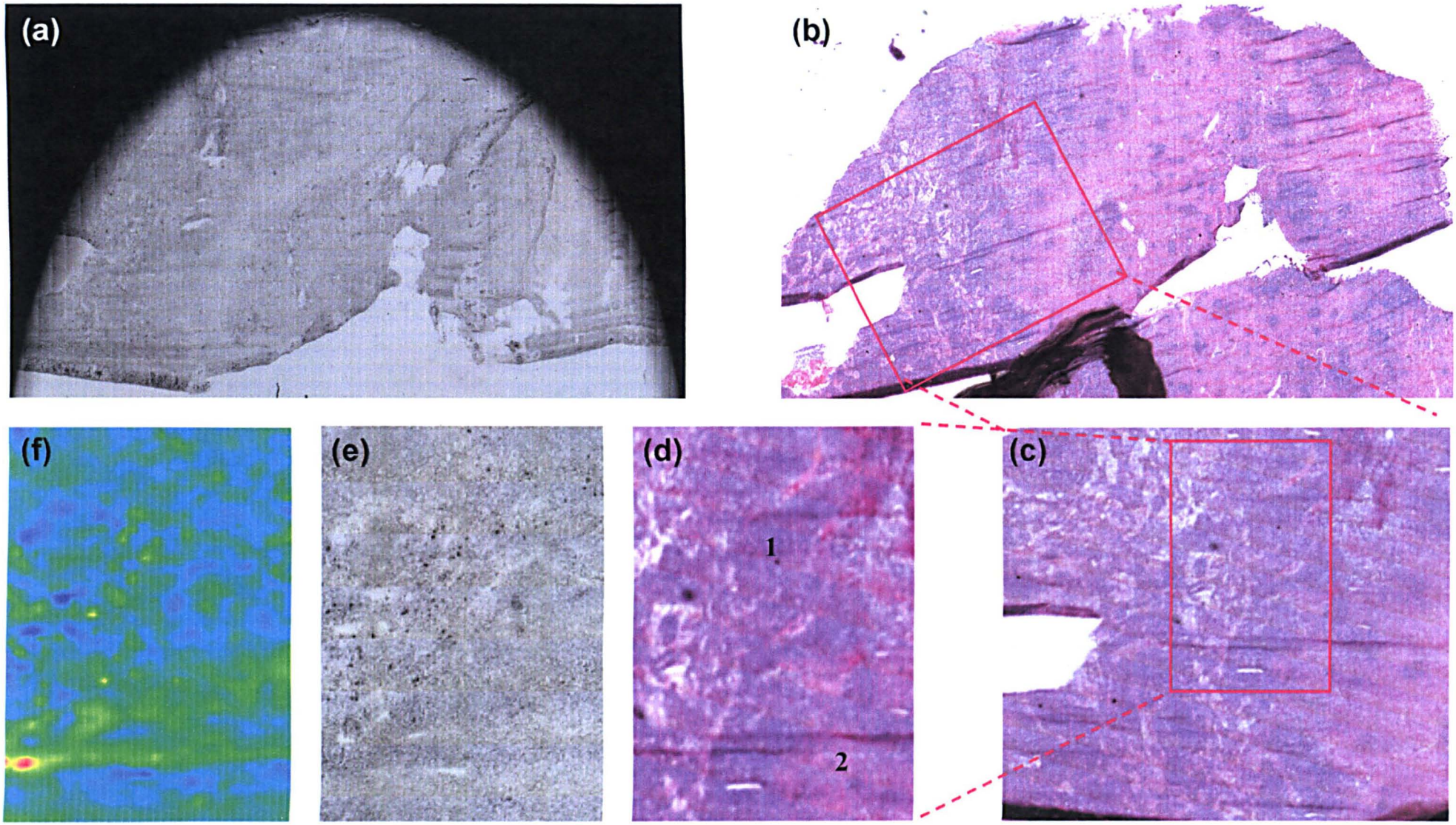

Figure 19: *a) White light image of entire lymph node tissue section. b) Photomicrograph of the H&E stained parallel section. c) Magnified region displaying benign and malign anatomical features. d) IR imaged area (1325 x 2125μm) mapped using a step size and aperture of 25μm for a total 4505 individual IR spectra. Benign (1) & malignant (2) tissues are identifiable via purple and pink colourations respectively. e) White light image of mapped area. f) Total absorbance IR image of mapped area.*

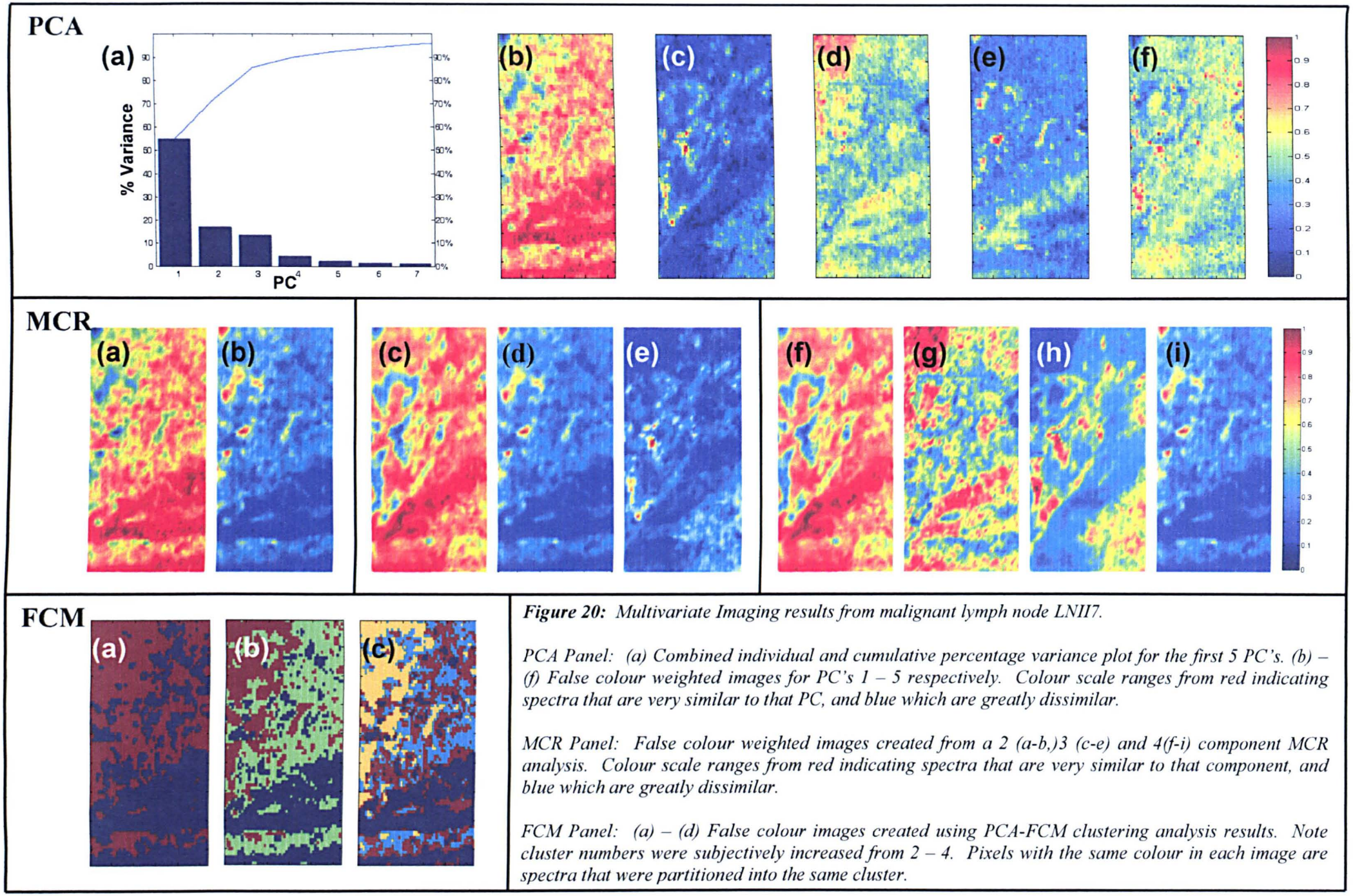

Figure 20: *Multivariate Imaging results from malignant lymph node LNII7.*

PCA Panel: (a) Combined individual and cumulative percentage variance plot for the first 5 PC's. (b) – (f) False colour weighted images for PC's 1 – 5 respectively. Colour scale ranges from red indicating spectra that are very similar to that PC, and blue which are greatly dissimilar.

MCR Panel: False colour weighted images created from a 2 (a-b,)3 (c-e) and 4(f-i) component MCR analysis. Colour scale ranges from red indicating spectra that are very similar to that component, and blue which are greatly dissimilar.

FCM Panel: (a) – (d) False colour images created using PCA-FCM clustering analysis results. Note cluster numbers were subjectively increased from 2 – 4. Pixels with the same colour in each image are spectra that were partitioned into the same cluster.

collected from the entire tissue section using a step size and aperture of 25μm. However, this map contained over 75,000 individual IR spectra. Due to limitations of available computer memory and CPU processing speeds, the time required to process a map of such magnitude was prohibitive. Therefore, a large section within this map, displaying both malignant and benign anatomical features was extracted for multivariate analysis. The area chosen for further analysis is shown in Figures 19d-f, and samples an area of 1325 x 2125 μm for a total of 4505 individual spectra. Examining closely the approximated studied area on the H&E stained parallel section (Figure 19), it can be seen that the area scrutinised displays a region whereby healthy cortex tissue (purple pigmentation) is being infiltrated on various fronts by cancerous tissue (pink pigmentation). Pockets that contain intermingled cancerous and healthy tissues can be located in both the top left and bottom right of the photomicrograph. The multivariate imaging results produced for this dataset are shown in Figure 20. Each method applied has been allocated an individual panel and only displays imaging results that produce meaningful information about the tissue section and the technique that was used.

Examining the PCA panel, it can be seen in figure (a) that over 95% of the total variance contained within the dataset is comprised within the first 7 PC's. When examining the first PC image displayed in figure (b), it is apparent that this component is highlighting areas upon the tissue section where cells are tightly packed or very dense (intense red colour). The second PC image shown in figure (c) more interestingly appears to highlight regions upon the tissue section where cancerous and normal tissue are intermingled and likely to be undergoing malignant change (red and cyan colour). In contrast, the third PC image displayed in figure (d)

more clearly highlights the cancerous regions of the tissue section (red and yellow colour). Both the remaining and subsequent PC images that were constructed gave no further beneficial tissue discrimination.

The MCR panel displays the resulting images constructed from a 2, 3 and 4 component analysis of the same dataset. When comparing these imaging results against the H&E stained section, the 4 component system gives the best characterisation of the tissue section (images f – i). The first component in the analysis (image f), displays areas upon the tissue section where the cells are tightly packed. The second component (image g) appears to be characteristic of the cancerous tissue and the third (image h) of areas where normal and cancerous tissues are intermingled. The final component (image i) reflects areas upon the section where the cells are not as tightly packed and therefore less dense.

The final panel displays images created via PCA-FCM clustering. Images (a) to (c) were constructed by subjectively increasing the amount of clusters found by the analysis from 2 – 4 respectively. When comparing these directly against the H&E stained section, the image constructed from a 3 cluster analysis seems to best mimic the histological architecture of the tissue section. The blue cluster of spectra appears to be characteristic of the invading cancerous tissue. In contrast, the green cluster of spectra can be attributed to the healthy cortex tissue. The final red cluster located in the top left and several small pockets around the section is descriptive of areas where normal tissue is intermingled with cancerous, and likely to be undergoing malignant change. The subsequent 4 cluster analysis shown in image (c) partitions the spectra surrounding these intermingled areas into a further cluster and could again be

descriptive of a further subtype of tissue that is at later or earlier stage of malignant change.

An additional infrared micro-spectral map was collected from the same tissue section but at higher spatial resolution. The region examined displayed a more distinct boundary between healthy and cancerous tissue and is shown in Figures 21a-f. On this occasion the detector array was set to examine the sample with a 6.25 μm pixel size. This map contained 5764 individual spectra and sampled an area of 275 x 818.75 μm. The multivariate imaging results produced from this dataset are shown in Figure 22, and again display 3 panels for each individual multivariate method.

Examining the PCA panel, it can be seen in figure (a) that over 95% of the total variance contained within the dataset is comprised within the first 10 PC's, the overwhelming majority accounted by the first 5. When studying the first PC image shown in figure (b), it would appear that this component highlights the remnant healthy nodal tissue that is located at the top of the imaged area (red and yellow pigmentation). In contrast, the cancerous invading tissue is marked by a cyan and deep blue colouration. The second component image shown in figure (c) again appears characteristic of the healthy tissue region (red pigmentation), but additionally highlights areas within the cancerous tissue that have a similar strong correlation. This component may therefore be diagnostic of healthy tissue that is undergoing malignant change, but this conclusion is only speculative. The remaining and subsequent PC images do not reveal any further beneficial tissue discrimination.

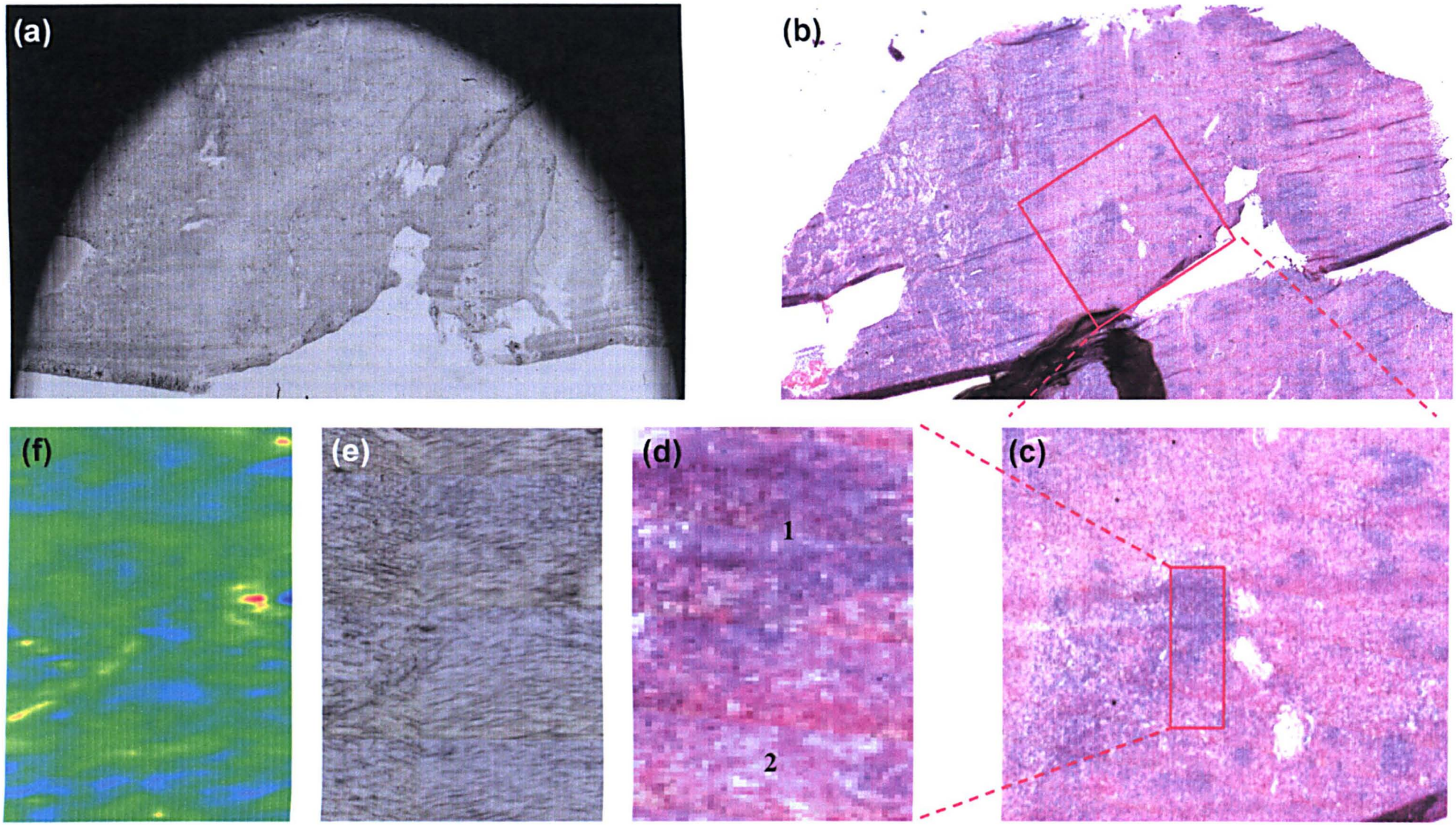

Figure 21: *a) White light image of the entire lymph node tissue section. b) Photomicrograph of the H&E stained parallel section. c) Magnified region displaying benign and malign anatomical features. d) IR imaged area (275 x 818.75 μm) mapped using a pixel size of 6.25 μm for a total of 5764 individual IR spectra. Benign (1) & malignant (2) tissues are identifiable via purple and pink colourations respectively. e) White light image of mapped area. f) Total absorbance IR image of mapped area.*

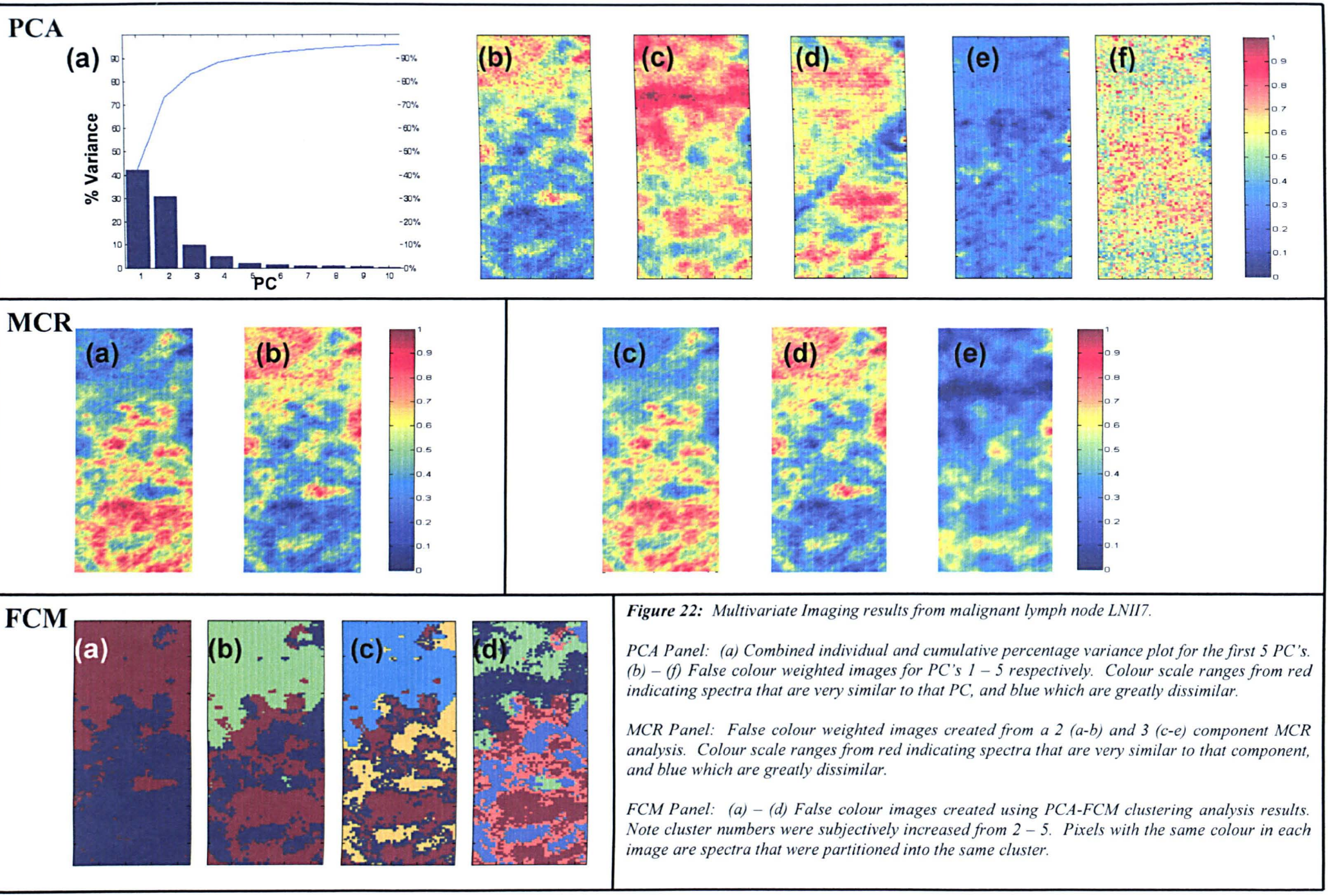

Figure 22: *Multivariate Imaging results from malignant lymph node LNII7.*

PCA Panel: (a) Combined individual and cumulative percentage variance plot for the first 5 PC's. (b) – (f) False colour weighted images for PC's 1 – 5 respectively. Colour scale ranges from red indicating spectra that are very similar to that PC, and blue which are greatly dissimilar.

MCR Panel: False colour weighted images created from a 2 (a-b) and 3 (c-e) component MCR analysis. Colour scale ranges from red indicating spectra that are very similar to that component, and blue which are greatly dissimilar.

FCM Panel: (a) – (d) False colour images created using PCA-FCM clustering analysis results. Note cluster numbers were subjectively increased from 2 – 5. Pixels with the same colour in each image are spectra that were partitioned into the same cluster.

The MCR panel displays the resulting images constructed from both a 2 and 3 component analysis of the same dataset. When comparing these imaging results against the H&E stained section, the 3 component system gives the best characterisation of the tissue section (images c – e). The first component in the analysis (image c), displays areas of cancerous invasion, whereas the second component (image d) is descriptive of healthy tissue. The third and final component (image e) appears to be characteristic of tightly packed cancerous cells.

The final panel again displays images created via PCA-FCM clustering. Images (a) to (d) were constructed by subjectively increasing the amount of clusters found by the analysis from 2 – 5 respectively. When comparing these directly against the H&E stained section, the image constructed from a 3 cluster analysis seems to best mimic the histological architecture of the tissue section. The green cluster of spectra appears to be characteristic of the healthy cortex tissue. In contrast, the red cluster of spectra can be attributed to the invading cancerous tissue. The final blue cluster located in several small pockets at the bottom of the image describes areas where tightly packed cancerous cells exist. The subsequent 4 and 5 cluster analyses displayed in images (c) and (d) further partition both the cancerous and healthy tissue spectra into additional subsets that may be descriptive of alternative stages of disease change.

2.3.2.2 Axillary Lymph Node LN57

The second tissue section in our library (named LN57) was cut from a healthy lymph node undergoing benign reactive changes. The reactive change in the lymph nodes

architecture was most likely caused by an immune response to the invading primary tumour in the breast or from previous surgical procedures such as a biopsy. Both the white light image of the tissue section and a photomicrograph of its parallel H&E stained section are shown in Figures 23a and 23b respectively. An infrared micro-spectral map was collected from the entire tissue section using a step size and aperture of 25μm. However, this map again contained above 50,000 spectra and was thus reduced to include the left region of the node that displayed the most prevalent reactive changes. The area chosen for further analysis is shown in Figures 23c-e, and samples an area of 3250 x 2675μm for a total of 13,910 individual IR spectra. The multivariate imaging results produced for this dataset are shown in Figure 24. Each method applied has been allocated an individual panel and only displays imaging results that produce meaningful information about the tissue section and the technique that was used.

Examining the PCA panel, false colour weighted images for the first 5 PC's have been constructed and are shown in figures (a) – (e). In this analysis, over 95% of the original variance contained within the dataset was now accounted by the first PC alone. Studying the constructed image for this component in figure (a), we can see that this PC clearly gives contrast between the fatty and remaining nodal tissue of the lymph node. Both the second and third PC images shown in figures (b) and (c) appear to highlight the capsule tissue of the lymph node. The constructed image for the fourth PC shown in figure (d) highlights a small globule of dense fatty tissue located at the bottom left of the capsule region. The fifth PC image shown in figure (e) and further subsequent PC images reveal no additional information about the lymph node. With over 95% of the variance being accounted by the fatty tissue

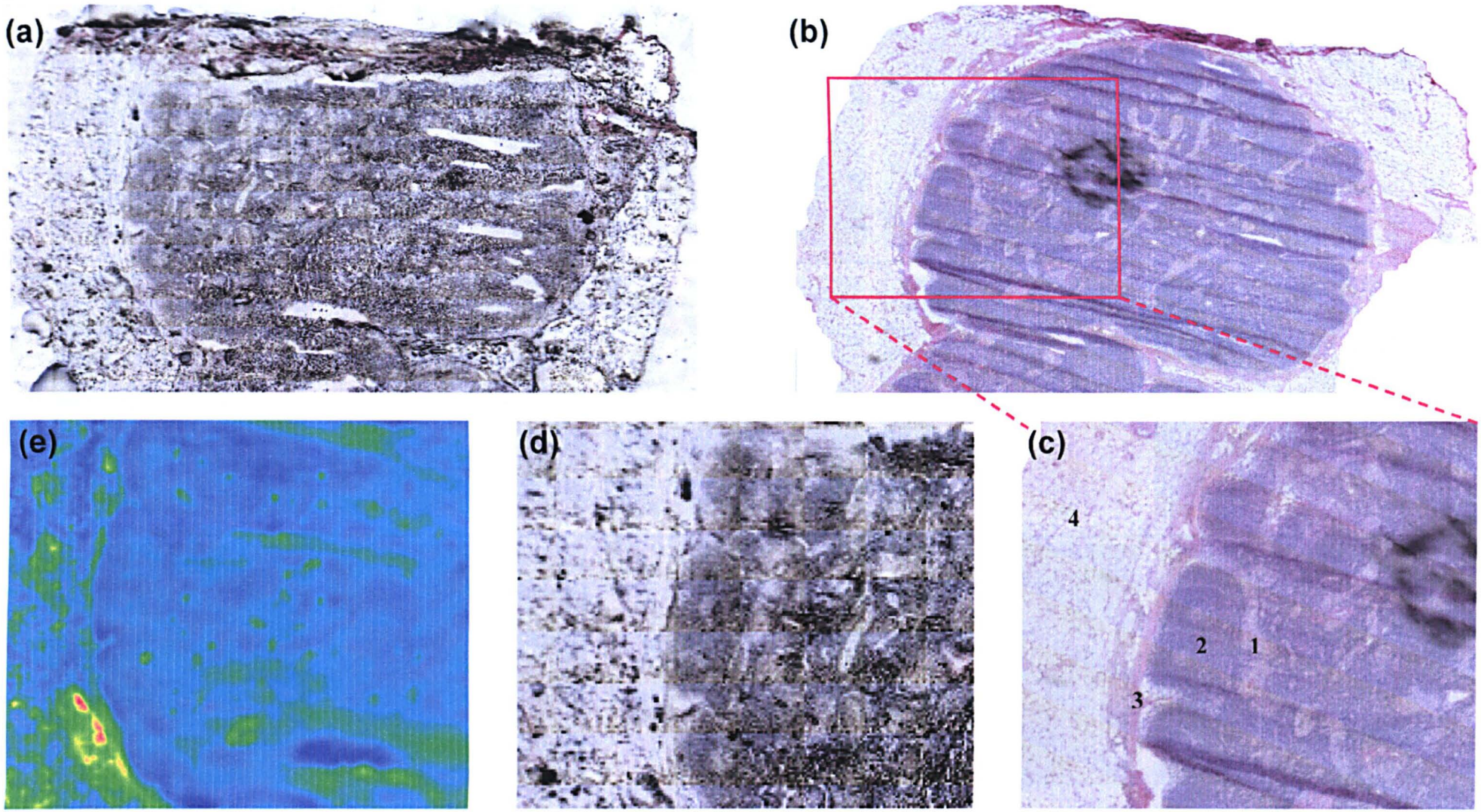

Figure 23: *a) White light image of entire lymph node tissue section. b) Photomicrograph of the H&E stained parallel section. c) IR imaged area (3250 x 2675 µm) at high magnification. The area was mapped using a step size and aperture of 25 µm for a total of 13,910 individual IR spectra. The typical anatomical features of a healthy lymph node undergoing reactive changes can be seen. These include the hypertrophy of the medullary sinuses filling with histocytes (1), enlarged and reactive secondary follicles with large germinal centres (2), the lymph node capsule (3) and surrounding fatty tissue (4). e) White light image of mapped area. f) Total absorbance IR image of mapped area.*

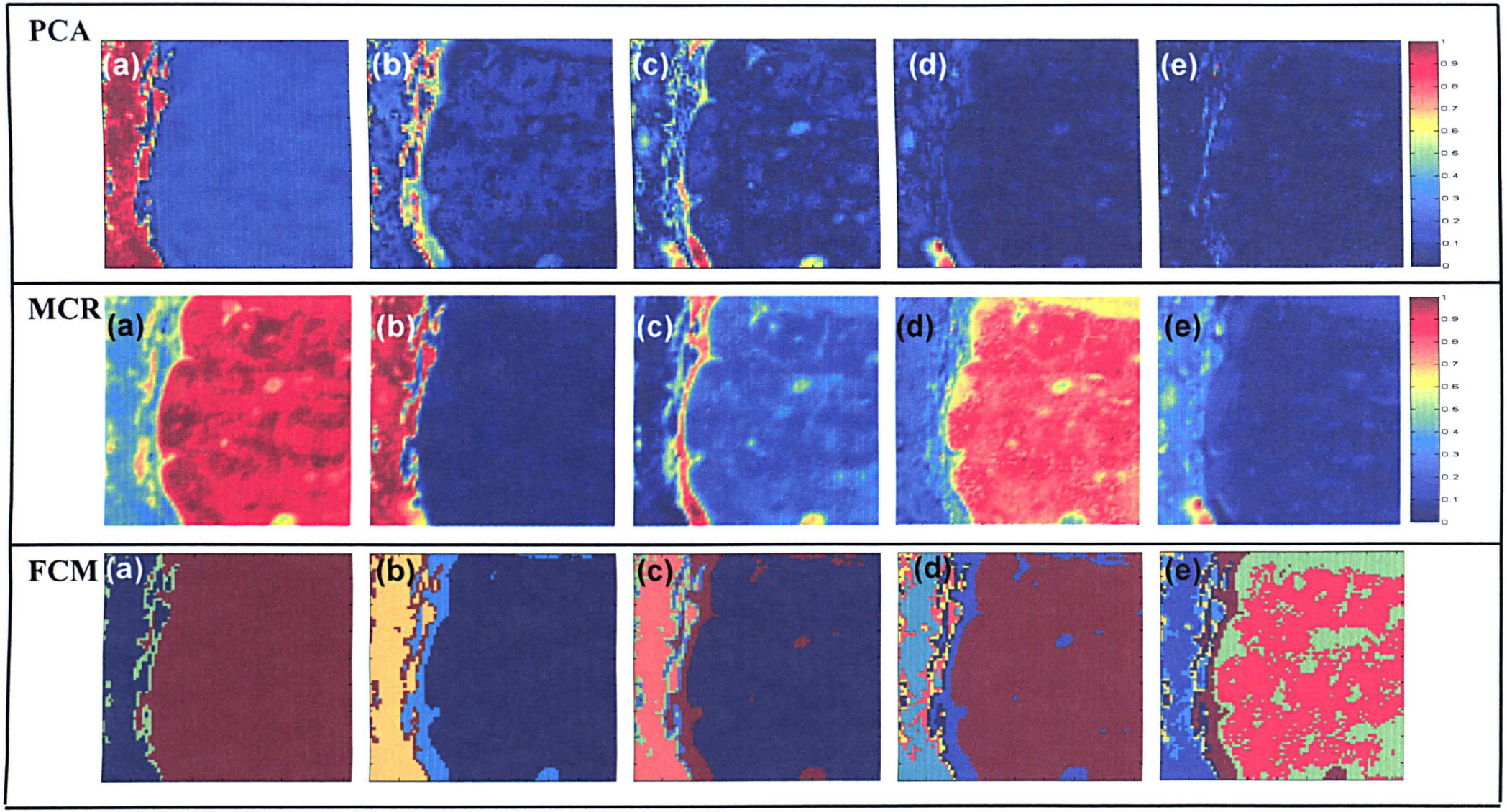

Figure 24: *Multivariate Imaging results from benign lymph node LN57. PCA Panel: (a) – (f) False colour weighted images for PC's 1 – 5 respectively. Colour scale ranges from red indicating spectra that are very similar to that PC, and blue which are greatly dissimilar. Approximately 98% of the total variance contained within the dataset is comprised by the first PC alone.. MCR Panel: (a) – (e) False colour weighted images created from a 5 component MCR analysis. Colour scale ranges from red indicating spectra that are very similar to that component, and blue which are greatly dissimilar. FCM Panel: (a) – (e) False colour images created using PCA-FCM clustering analysis results. Note cluster numbers were subjectively increased from 3 – 7. Pixels with the same colour in each image are spectra that were partitioned into the same cluster.*

alone, the sensitivity required for further tissue discrimination has almost certainly been affected.

The MCR panel displays the resulting images constructed from a 5 component analysis of the same dataset (images a – e). This 5 component system gave the best characterisation of the tissue section when compared to the H&E stained section. The first component in the analysis (image a), is representative of all the normal nodal tissue, whereas the second component (image b) is again descriptive of outlining fatty tissue. Studying the third component (image c), this image clearly highlights the capsule tissue of the lymph node. The fourth component (image d) again highlights the nodal cortex tissue, but now provides a small amount of contrast between the reacting secondary follicles and the medullary sinuses that have hypertrophied. Image (e) constructed from the fifth component marks the globule of fatty tissue located at the bottom left of the tissue section.

The final panel again displays images created via PCA-FCM clustering. Images (a) to (e) were constructed by subjectively increasing the amount of clusters found by the analysis from 3 – 7 respectively. When comparing these directly against the H&E stained section, the image constructed from a 7 cluster analysis seems to best mimic the histological architecture of the tissue section. The dark blue, royal blue, cyan and yellow clusters of spectra appear to characterise the outlining fatty tissue. In contrast, the maroon cluster of spectra can be attributed to the lymph node capsule tissue. The orange cluster on the other hand highlights regions upon the tissue section where highly proliferating secondary follicle are present. The final green cluster of spectra located in several small pockets surrounding the secondary follicles

is likely to represent the medullary sinuses that have hypertrophied. Subsequent 7 - 10 cluster images revealed no further information about the lymph node and proceeded to further partition the fatty tissue spectra into multiple groups. Overall in must be noted that the fatty tissue comprised so much natural variation within its spectral characteristics that the analysis could only differentiate between the secondary follicle and medullary sinuses after the fatty tissue had been partitioned into 4 subsets of spectra.

2.3.2.3 Axillary Lymph Node LNPE

The third tissue section in our library (named LNPE) was cut from another healthy lymph node. This node, however, displayed a typical benign variation that can occur within the reticulum support structure, whereby the capsule is thickened and bands of fibrocollagenous scar tissue invade into the core of the node. A white light image of the entire tissue section and the region chosen for analysis are shown in figures 25a to 25d respectively. Unfortunately a parallel H&E stained section was not made available for this node, but the main types of tissue can still be visualised via contrast in light intensity of the tissue regions (figure 25c). An infrared micro-spectral map was collected from a cross section of the node that incorporated all tissue types present on tissue section. By use of a step size and aperture of 25 μm, a total of 2522 individual IR spectra were collected from an area of 650 x 242 μm. The multivariate imaging results produced for this dataset are shown in Figure 26. Each method applied has been allocated an individual panel and only displays imaging results that produce meaningful information about the tissue section and the technique that was used.

Examining the PCA panel, it can be seen in figure (a) that over 95% of the total variance contained within the dataset was comprised within the first 3 PC's. When studying the colour weighted image for the first PC in figure (b), we can see that this PC clearly gives contrast between the fatty tissue invading the capsule (blue colouration) and the remaining nodal tissue (red colouration). Studying the second PC image shown in figure (c), this clearly provides contrast between the tissue section itself (red pigmentation) and the region at the top left hand corner of the mapped area where no tissue exists. All subsequent PC images do not provide any further reliable contrast between the tissue types present.

The MCR panel displays the resulting images constructed from both a 2 and 3 component analysis of the same dataset (images a – e). By comparison to the known histological tissue types on the section, the 3 component system gave the best characterisation of the tissue section. The first component in the analysis (image c), is representative of all the normal nodal tissue, whether being cortex or collagenous scar tissue. The second component (image d), is descriptive of fatty tissue invading the node producing capsular thickening. This component image also shows some contrast between the collagenous scar tissue (yellow colouration) and central cortex tissue (cyan colouration). Studying the third and final component (image e), this clearly highlights the region at the top left of the mapped area where no tissue exists.

The final panel again displays images created via PCA-FCM clustering. Images (a) to (c) were constructed by subjectively increasing the amount of clusters found by the analysis from 2 – 4 respectively. When comparing these directly against the

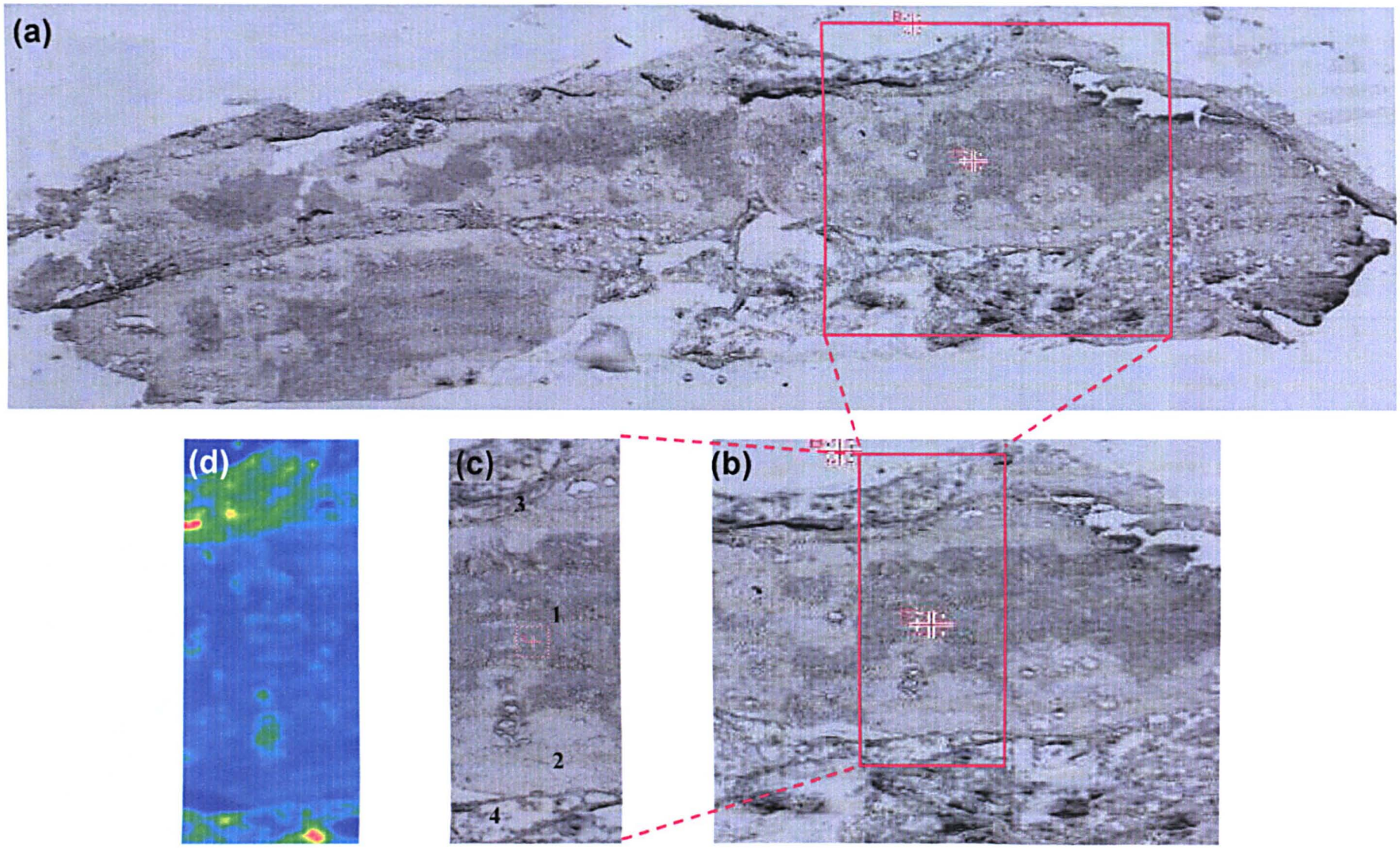

Figure 25: *a) White light image of entire lymph node tissue section. b) Magnified region displaying a typical benign variation that can occur within the reticulum support structure of a lymph node, which includes capsular thickening and bands of fibrocollagenous scar tissue. c) IR imaged area (650 x 2425 μm) mapped using a step size and aperture of 25 μm for a total 2522 individual IR spectra. Tissue types found within the mapped area include cortex (1), collagenous scar (2), capsule (3) and fatty (4) tissues. d) Total absorbance IR image of mapped area.*

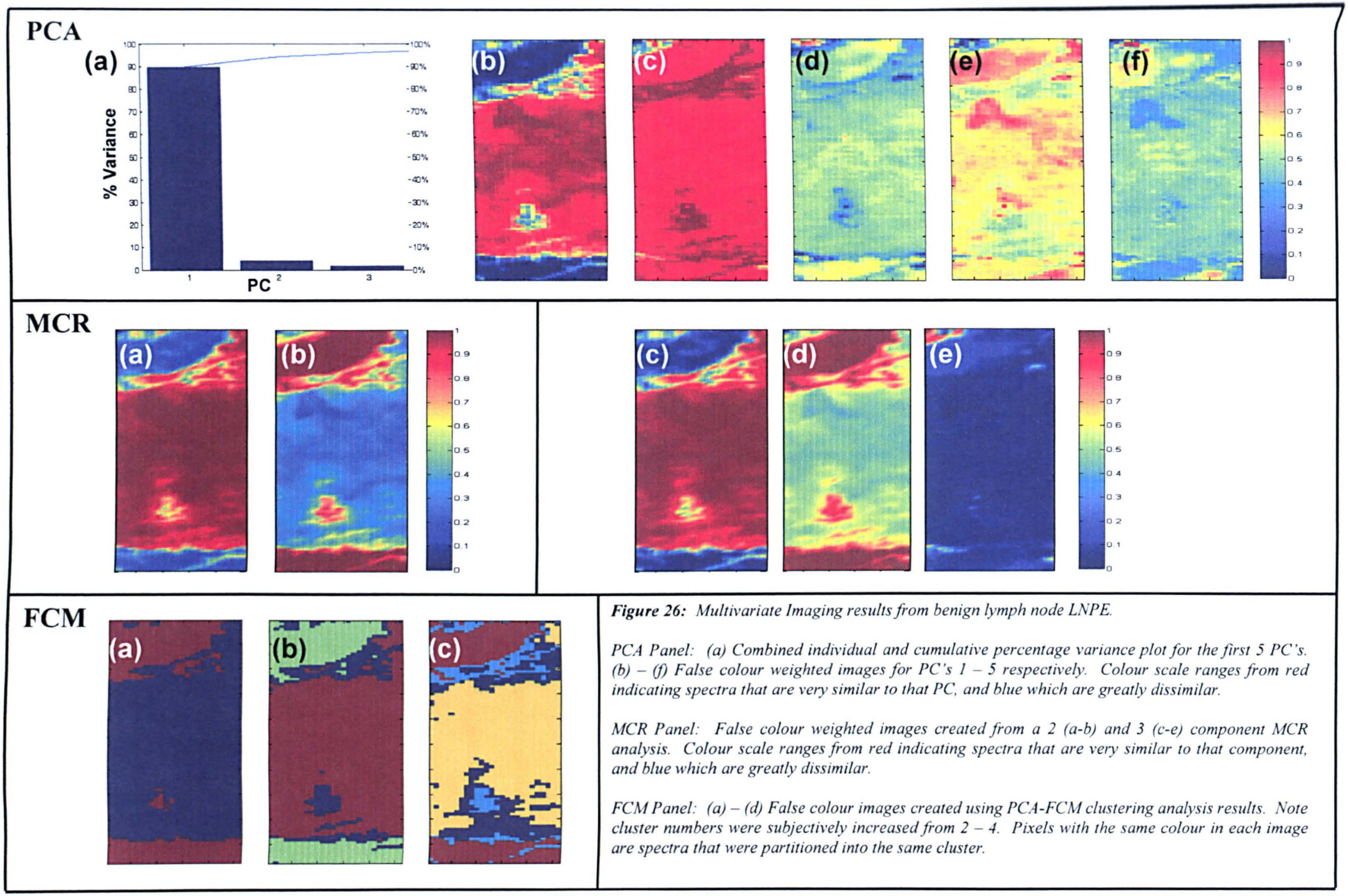

Figure 26: *Multivariate Imaging results from benign lymph node LNPE.*

PCA Panel: (a) Combined individual and cumulative percentage variance plot for the first 5 PC's. (b) – (f) False colour weighted images for PC's 1 – 5 respectively. Colour scale ranges from red indicating spectra that are very similar to that PC, and blue which are greatly dissimilar.

MCR Panel: False colour weighted images created from a 2 (a-b) and 3 (c-e) component MCR analysis. Colour scale ranges from red indicating spectra that are very similar to that component, and blue which are greatly dissimilar.

FCM Panel: (a) – (d) False colour images created using PCA-FCM clustering analysis results. Note cluster numbers were subjectively increased from 2 – 4. Pixels with the same colour in each image are spectra that were partitioned into the same cluster.

known tissue type regions, the image constructed from a 4 cluster analysis seems to best mimic the histological architecture of the tissue section. The yellow cluster of spectra characterise the central cortex tissue. In contrast, the blue cluster of spectra highlights the collagenous scar tissue that is invading into the core of the node. Remnant capsule tissue is highlighted by the cyan cluster of spectra, surrounding and lining the fatty tissue areas. Unfortunately a small globule of collagenous scar tissue has also been partitioned into this cluster (central part of image). However, after scrutinising spectra collected from these co-ordinates, it is apparent they have also taken on strong lipid characteristics similar to the invaded capsule. The final red cluster of spectra highlights fatty tissue that has invaded the capsule of the node.

2.3.2.4 Axillary Lymph Node LN24

The fourth tissue section in our library (named LN24) was cut from a malignant lymph node that had almost been completely infiltrated by fatty and fibrocollagenous scar tissues. However, a few small pockets of remnant cancerous tissue could still be found. A white light image of the entire tissue section and the region chosen for analysis are shown in figures 27a – c respectively. Unfortunately a parallel H&E stained section was not made available for this node, but the main types of tissue can still be visualised via contrast in light intensity of the tissue regions (figure 27b). An infrared micro-spectral map was collected from a cross section of the node that incorporated all tissue types present on the tissue section. By use of a step size and aperture of 25 μm, a total of 6020 individual IR spectra were collected from a spatial area of 875 x 4300 μm. The multivariate imaging results produced for this dataset are shown in Figure 28. Each method applied has been allocated an individual panel

and only displays imaging results that produce meaningful information about the tissue section and the technique that was used.

Examining the PCA panel, it can be seen in figure (a) that over 95% of the total variance contained within the dataset was comprised by the first 2 PC's. When studying the colour weighted image for the first PC in figure (b), it is apparent that this PC clearly provides contrast between the invading fatty (blue colouration) and the remaining nodal tissue (red colouration), whether it is cancerous or fibrocollagenous. The second PC image shown in figure (c) highlights two small pockets of dense fatty tissue (red pigmentation) located above and below the central region containing the remnant nodal tissue. The third PC image shown in figure (d) again highlights the central area containing the remnant cancerous tissue but now provides a small amount of contrast between the malignant cells (red colouration) and the fibrocollagenous scar tissue (blue colouration). All subsequent PC images provide little tissue discrimination that is useful.

The MCR panel displays the resulting images constructed from a 2, 3 and 4 component analysis of the same dataset (images a – i). By comparison to the known histological tissue types found in the sample, the 4 component system gave the best characterisation of the tissue section. The first component in the analysis (image f), discriminates the two small pockets of dense fatty tissue, whereas the second component (image g) describes the central region of remnant nodal tissue. A small amount of contrast between the cancerous cells (dark red) and surrounding fibrocollagenous scar tissue (light red) is provided by this component. In contrast, the third component (image h) clearly marks the areas of fatty tissue invasion. The

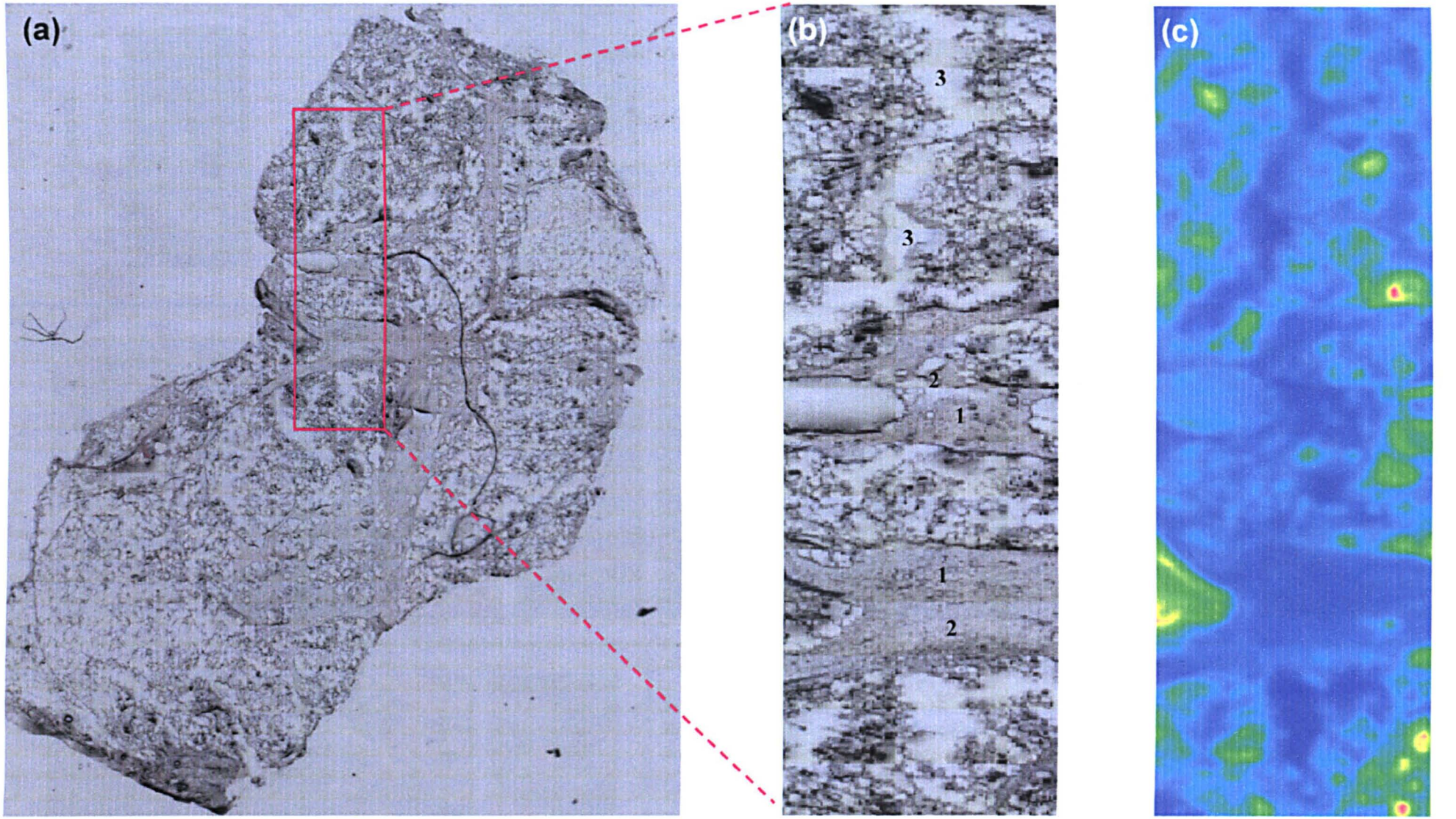

Figure 27: *a) White light image of entire lymph node tissue section. b) Magnified region displaying fibrocollagenous scar tissue that encapsulates small clusters of malignant cortex cells. The IR imaged area (875 x 4300μm) was mapped using a step size and aperture of 25μm for a total 6020 individual IR spectra. Tissue types found within the mapped area include cancerous cortex (1), collagenous scar (2), and fatty (3) tissues. c) Total absorbance IR image of mapped area.*

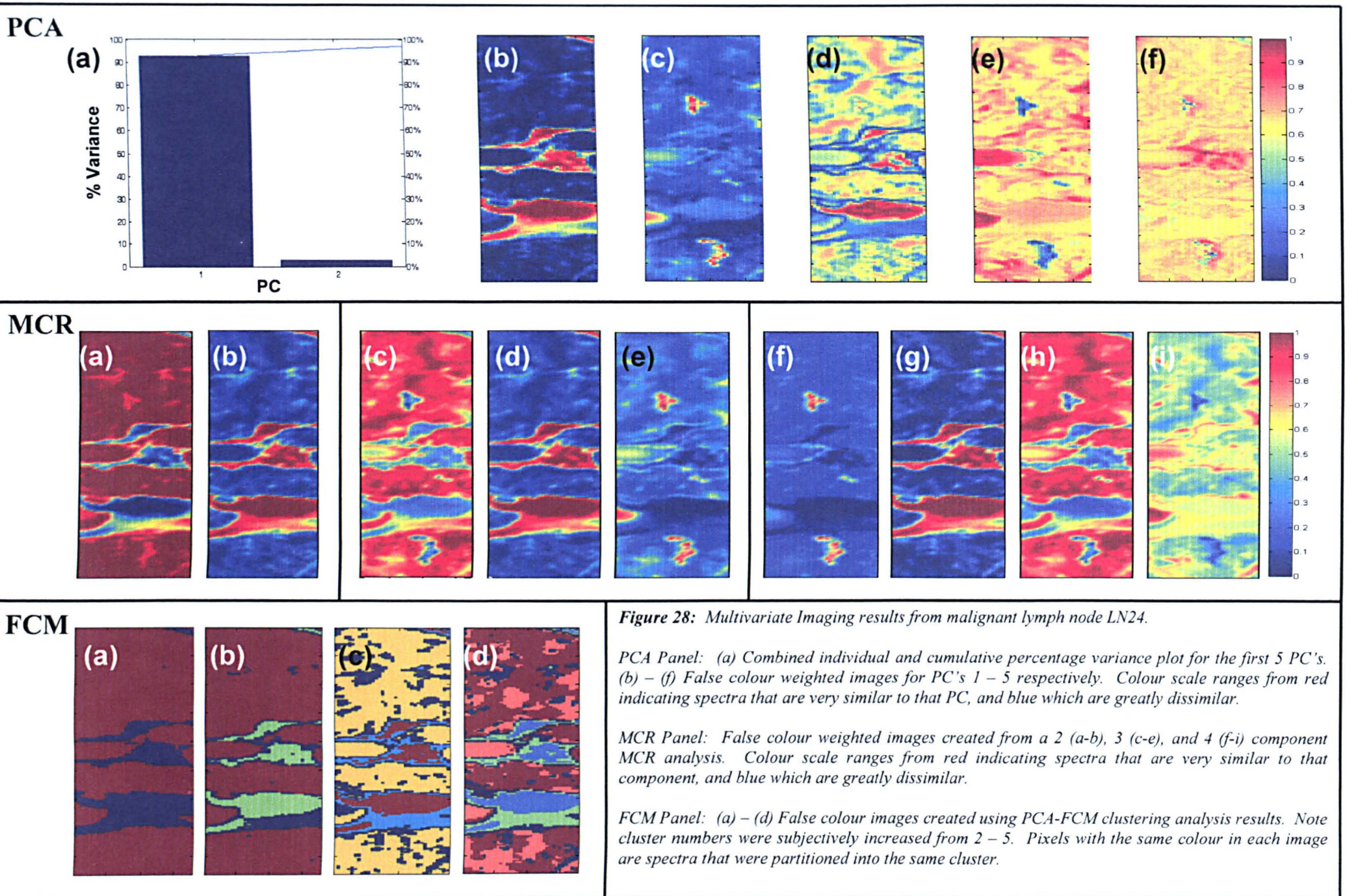

Figure 28: *Multivariate Imaging results from malignant lymph node LN24.*

PCA Panel: (a) Combined individual and cumulative percentage variance plot for the first 5 PC's. (b) – (f) False colour weighted images for PC's 1 – 5 respectively. Colour scale ranges from red indicating spectra that are very similar to that PC, and blue which are greatly dissimilar.

MCR Panel: False colour weighted images created from a 2 (a-b), 3 (c-e), and 4 (f-i) component MCR analysis. Colour scale ranges from red indicating spectra that are very similar to that component, and blue which are greatly dissimilar.

FCM Panel: (a) – (d) False colour images created using PCA-FCM clustering analysis results. Note cluster numbers were subjectively increased from 2 – 5. Pixels with the same colour in each image are spectra that were partitioned into the same cluster.

fourth and final component (image i) appears to highlight areas up on the tissue section where some remnant nodal tissue remains, although intermingled with fatty tissue.

The final panel displays images created via PCA-FCM clustering. Images (a) to (d) were constructed by subjectively increasing the amount of clusters found by the analysis from 2 – 5 respectively. When comparing these directly against the known tissue type regions, the image constructed from a 4 cluster analysis seems to best mimic the histological architecture of the tissue section (image c). The red cluster of spectra characterise the cancerous cells located in the central region of the remnant nodal tissue. Fibrocollagenous scar tissue that surrounds these cancerous cells is highlighted by the cyan cluster of spectra. The remaining remnant nodal tissue scattered across the section are described by the blue cluster of spectra, with the invading fatty tissue marked by the yellow cluster of spectra. When further increasing the amount of clusters found by the analysis, the fatty tissue is partitioned into further subsets of spectra

2.3.2.5 Axillary Lymph Node LNPF

The fifth tissue section in our library (named LNPF) was cut from a benign lymph node that had again been infiltrated by collagenous scar and fatty tissues. A white light image of the entire tissue section and the region chosen for analysis are shown in figures 29a – c respectively. Unfortunately a parallel H&E stained section was not made available for this node, but the main types of tissue can still be visualised via contrast in light intensity of the tissue regions (figure 29b). An infrared micro-

spectral map was collected from a cross section of the node that incorporated all tissue types present on the tissue section. By use of a step size and aperture of 25 μm, a total of 1920 individual IR spectra were collected from a spatial area of 500 x 2400 μm. The multivariate imaging results produced for this dataset are shown in Figure 30. Each method applied has been allocated an individual panel and only displays imaging results that produce meaningful information about the tissue section and the technique that was used.

Examining the PCA panel, it can be seen in figure (a) that over 95% of the total variance contained within the dataset was comprised by the first 4 PC's. When studying the colour weighted images for the first 5 PC's shown in figures (b) – (f), no clear or distinctive tissue differentiation can be made. The analysis again seems dominated by the variance contained within the fatty tissue spectra and regions where no tissue exists. All subsequent PC images provided no additional or insightful information about the area examined.

The MCR panel displays the resulting images constructed from a 2, 3 and 4 component analysis of the same dataset (images a – i). By comparison to the known histological tissue types found in the sample, the 4 component system gave the best characterisation of the tissue section.

The final panel again displays images created via PCA-FCM clustering. Images (a) to (c) were constructed by subjectively increasing the amount of clusters found by the analysis from 2 – 4 respectively. When comparing these directly against the known tissue type regions, the image constructed from a 4 cluster analysis seems to

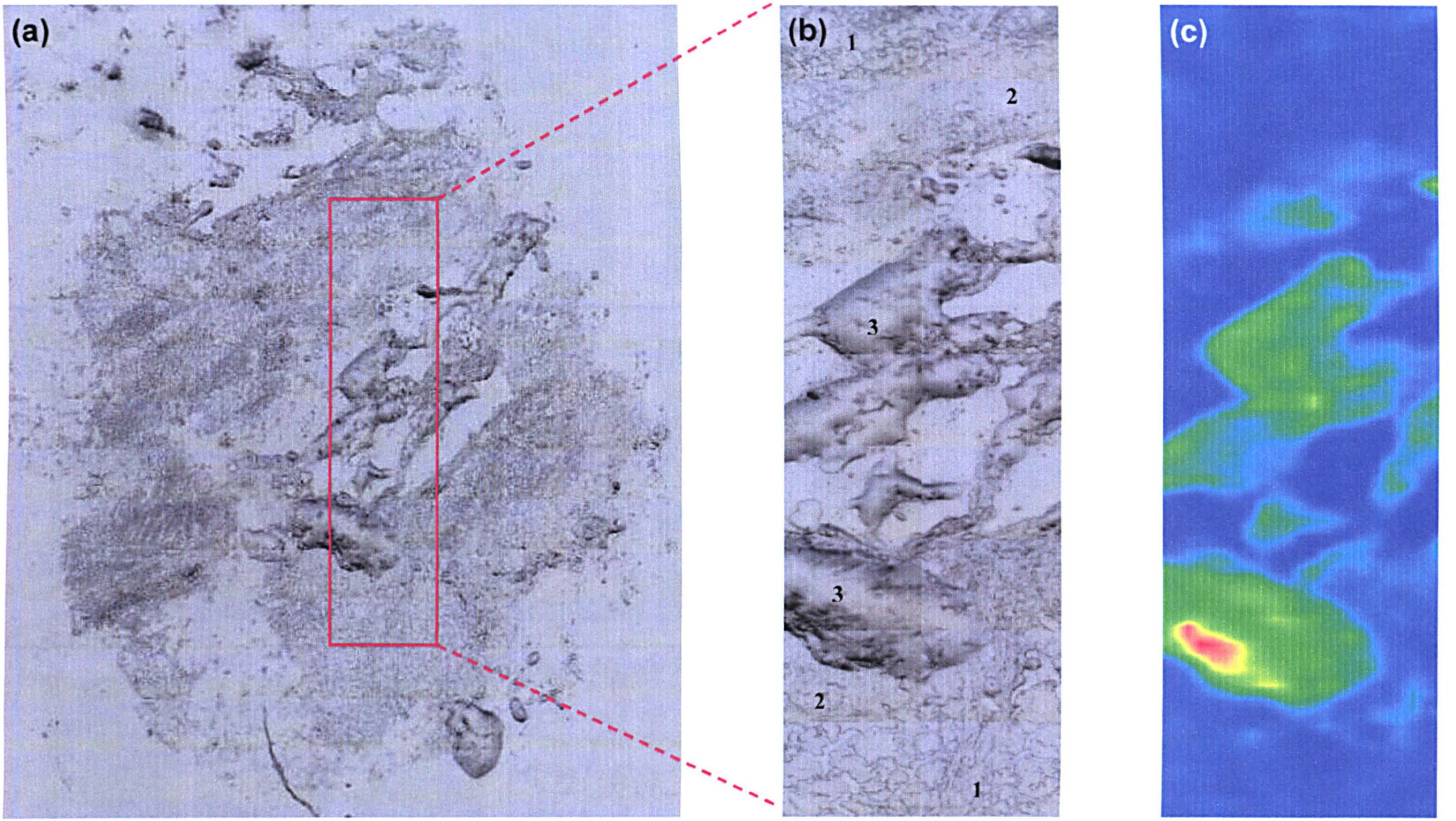

Figure 29: *(a) White light image of entire lymph node tissue section. (b) Magnified region displaying fibrocollagenous scar tissue that surrounds benign cortex tissue. The IR imaged area (1920 x 2400μm) was mapped using a step size and aperture of 25μm for a total 6020 individual IR spectra. Tissue types found within the mapped area include benign cortex (1), collagenous scar (2), and fatty (3) tissues. (c) Total absorbance IR image of mapped area.*

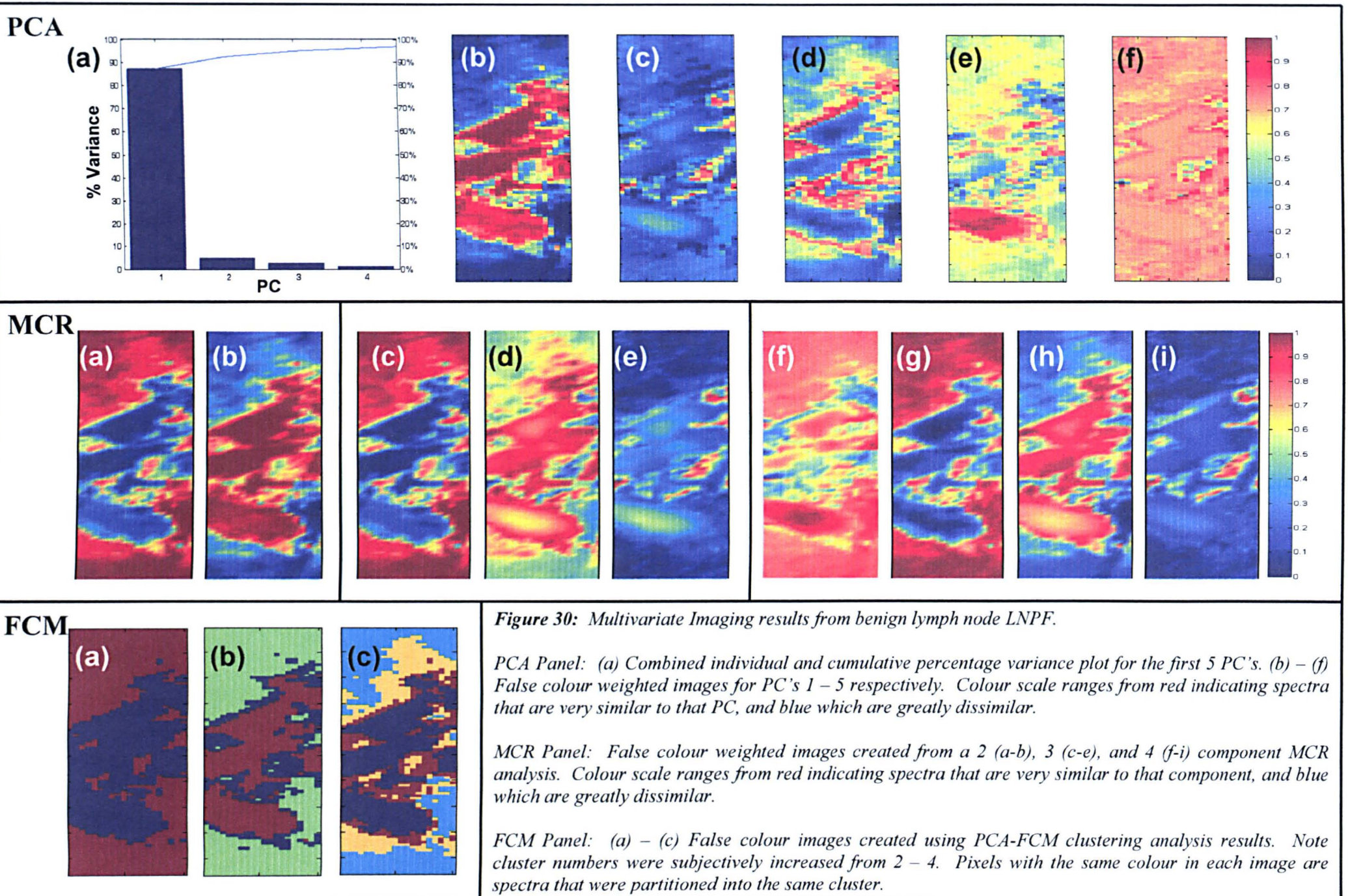

Figure 30: *Multivariate Imaging results from benign lymph node LNPF.*

PCA Panel: (a) Combined individual and cumulative percentage variance plot for the first 5 PC's. (b) – (f) False colour weighted images for PC's 1 – 5 respectively. Colour scale ranges from red indicating spectra that are very similar to that PC, and blue which are greatly dissimilar.

MCR Panel: False colour weighted images created from a 2 (a-b), 3 (c-e), and 4 (f-i) component MCR analysis. Colour scale ranges from red indicating spectra that are very similar to that component, and blue which are greatly dissimilar.

FCM Panel: (a) – (c) False colour images created using PCA-FCM clustering analysis results. Note cluster numbers were subjectively increased from 2 – 4. Pixels with the same colour in each image are spectra that were partitioned into the same cluster.

best mimic the histological architecture of the tissue section (image c). The central blue cluster of spectra clearly marks the area where large globules of fatty tissue have invaded into the lymph node. Tissue surrounding this fatty region where the infiltration is not as complete can be visualised by the red cluster of spectra. Collagenous scar tissue that separates this invaded area from the remaining node is highlighted by the yellow cluster of spectra, allowing the healthy cortex tissue spectra to be visualised with a cyan colouration. When increasing the amount of clusters above this number, the analysis further partitions the fatty tissue into multiple subsets of spectra.

2.3.3 The Combined Tissue Classification of Multiple Lymph Node IR Micro-spectral Datasets via FCM Clustering

The first two sections of this chapter have demonstrated the ability of unsupervised multivariate techniques to discriminate and characterise different tissue types that exist in both healthy and diseased lymph nodes. However, these experiments were carried out upon IR micro-spectral datasets collected from a single lymph node. To fully assess the potential of micro-spectroscopy as a tool for automated pathology, natural variation between patients and tissues must also be evaluated. In this section we therefore describe experiments that coalesce tissue spectra collected from multiple lymph nodes for a combined tissue classification.

In these experiments, tissue spectra were extracted from IR micro-spectral maps collected from three different lymph nodes. Cancerous tissue spectra were extracted from datasets of positive lymph nodes LNII5 and LNII7. Spectra were collected

from distinct regions upon the tissue sections where only malignant tissue existed. Healthy tissue spectra were alternatively collected from benign lymph node LN57. These healthy tissue spectra were collected from a region upon the tissue section that described an area of reacting secondary follicles. Spectra were extracted from this particular region of the lymph node as the secondary follicles comprise highly proliferating cells similar to malignant tissue. This would therefore provide a sterner test for any subsequent multivariate analyses. The defined regions where tissue spectra were collected from each lymph node are shown in figures 31a – c respectively.

The first experiment undertaken in this study combined healthy tissue spectra collected from lymph node LN57 and cancerous tissue spectra from lymph node LNII7. This combined dataset was then scrutinised by PCA-FCM clustering analysis, specifying a 2 cluster result. The results from this cluster analysis are shown in figure 32a, with all spectra projected onto the first two principal component dimensions in multi-dimensional space. Healthy lymph node spectra were correctly partitioned into the same single cluster illustrated by the green data points, whereas the cancerous spectra have been grouped into a separate cluster described by the red data points.

The second experiment alternatively combined the healthy tissue spectra from lymph node LN57 with the cancerous tissue spectra from positive lymph node LNII5. Again this combined dataset was scrutinised by PCA-FCM clustering analysis, specifying a 2 cluster result. The clustering results from this analysis are shown in

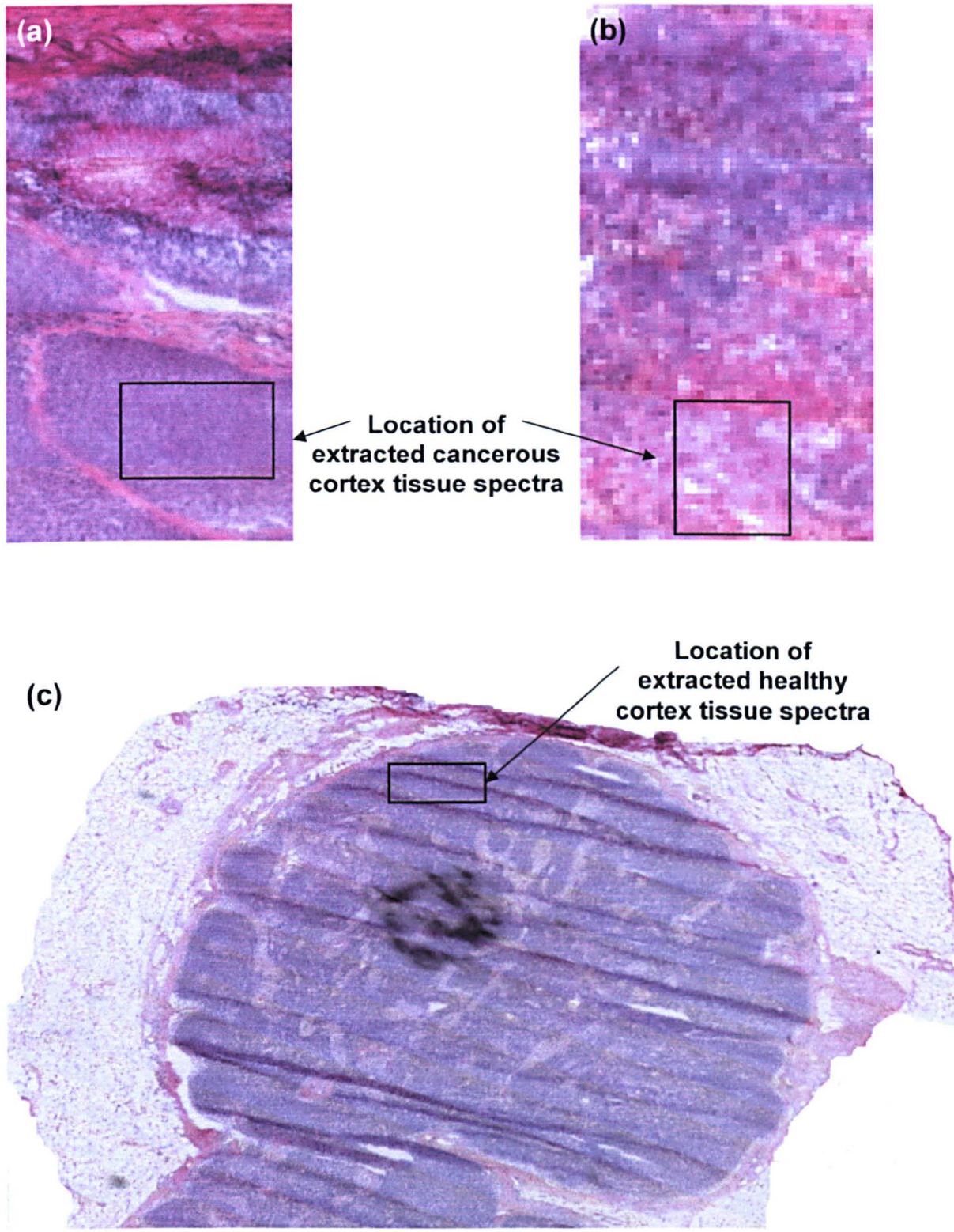

Figure 31: *(a) H&E stained photomicrograph of malignant lymph node LNII5. (b) H&E stained photomicrograph of malignant lymph node LNII7. (c) H&E stained photomicrograph of healthy lymph node LN57. Spectra were extracted from previously recorded IR micro-spectral maps, the locations of which have marked with black boxes.*

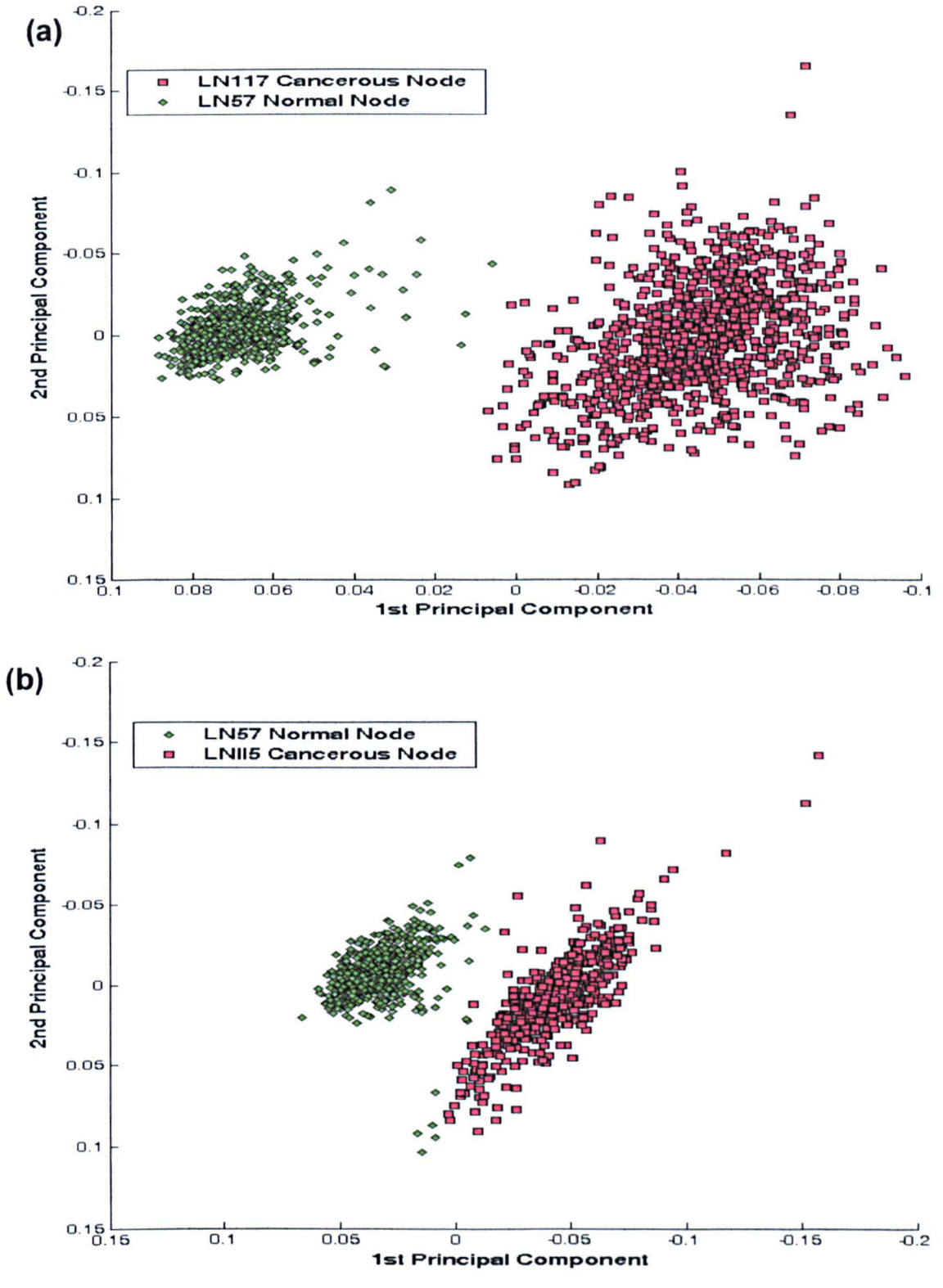

Figure 32: *Clustering results of combined lymph node tissue spectral datasets via the PCA-FCM Algorithm. (a) Healthy cortex tissue spectra from LN57 and malign cortex tissue spectra from LNII7; 2 specified clusters. (b) Healthy cortex tissue spectra from LN57 and malign cortex tissue spectra from LNII5; 2 specified clusters.*

figure 32b, all spectra again projected onto the two first principal component dimensions. PCA-FCM analysis again proved successful, partitioning the healthy (green data points) and cancerous tissue spectra (red data points) into separate and defining clusters. However, a very small number of cancerous tissue spectra were misclassified in this instance.

Although both these analyses proved very successful, the clustering experiments had not yet tried to partition tissue spectra of the same diagnosis from different lymph nodes into the same group. Therefore, in the third experiment we combined tissue spectra from all three datasets and again preceded with a 2 cluster PCA-FCM analysis. The clustering result from this analysis is shown in figure 33a, all spectra again projected onto the two first principal component dimensions. As shown in the diagram, healthy tissue spectra were again correctly clustered into one group (green data points), but more importantly the cancerous tissue spectra from the two separate lymph nodes were now partitioned into one defining malignant group (red data points).

In an attempt to test the sensitivity of the PCA-FCM clustering approach, we repeated the analysis on this dataset but increased the specified number of clusters to 3. The clustering result from this analysis is shown in figure 33b, all spectra again projected onto the first two principal component dimensions. As shown in the diagram, the PCA-FCM clustering analysis has on this occasion further partitioned the cancerous tissue spectra into two clusters that represent malignant lymph nodes LNII5 (red data points) and LNII7 separately (blue data points). This was an interesting result as it clearly showed that cancerous tissue spectra exhibited natural

variation between alternative lymph nodes and was identifiable by clustering techniques. To assess the difference between the spectral characteristics of these two cancerous lymph nodes, average spectra for each cluster were calculated and are shown in figures 34 and 35 respectively.

At first glance, spectra from the three different lymph nodes appear to be very similar, with the most discernable changes occurring within the finger print region. When examining this spectral region in greater detail, as shown in figure 35, all three groups of tissue spectra display an almost identical spectral profile. The absence of broad overlapping collagen peaks between 1180 – 1380 cm^{-1} is noticeable [29-31], allowing shoulder peaks at 1468 and 1408cm^{-1} to be revealed, likely attributable to CH_2 scissoring vibrations in lipids and methyl deformation modes respectively. These spectra also display pronounced antisymmetric and symmetric phosphodiester vibrations at 1240 cm^{-1} and 1085 cm^{-1} respectively [32-34]. If we now consider the average spectra from the three lymph nodes in this region, discernable changes can be identified between the three clusters. The healthy yet highly proliferating tissue spectra from LN57 (green profile) display a strong symmetric phosphodiester band at 1085 cm^{-1}, but provides the weakest peak intensity from the 3 tissue clusters. In contrast, the cancerous tissue spectra collected from LNII5 (red profile) and LNII7 (blue profile) display a marked increase in the intensity of this band. In fact both the intensity and width of this band for spectra originating from LNII7 are significantly increased when compared to the remaining two clusters. These observations are in agreement with previous studies that indicated the relative intensity of this band can be descriptive to a cells divisional activity [32-34,42,43], with diseased tissues displaying a marked intensity increase. Nevertheless, it is apparent from these

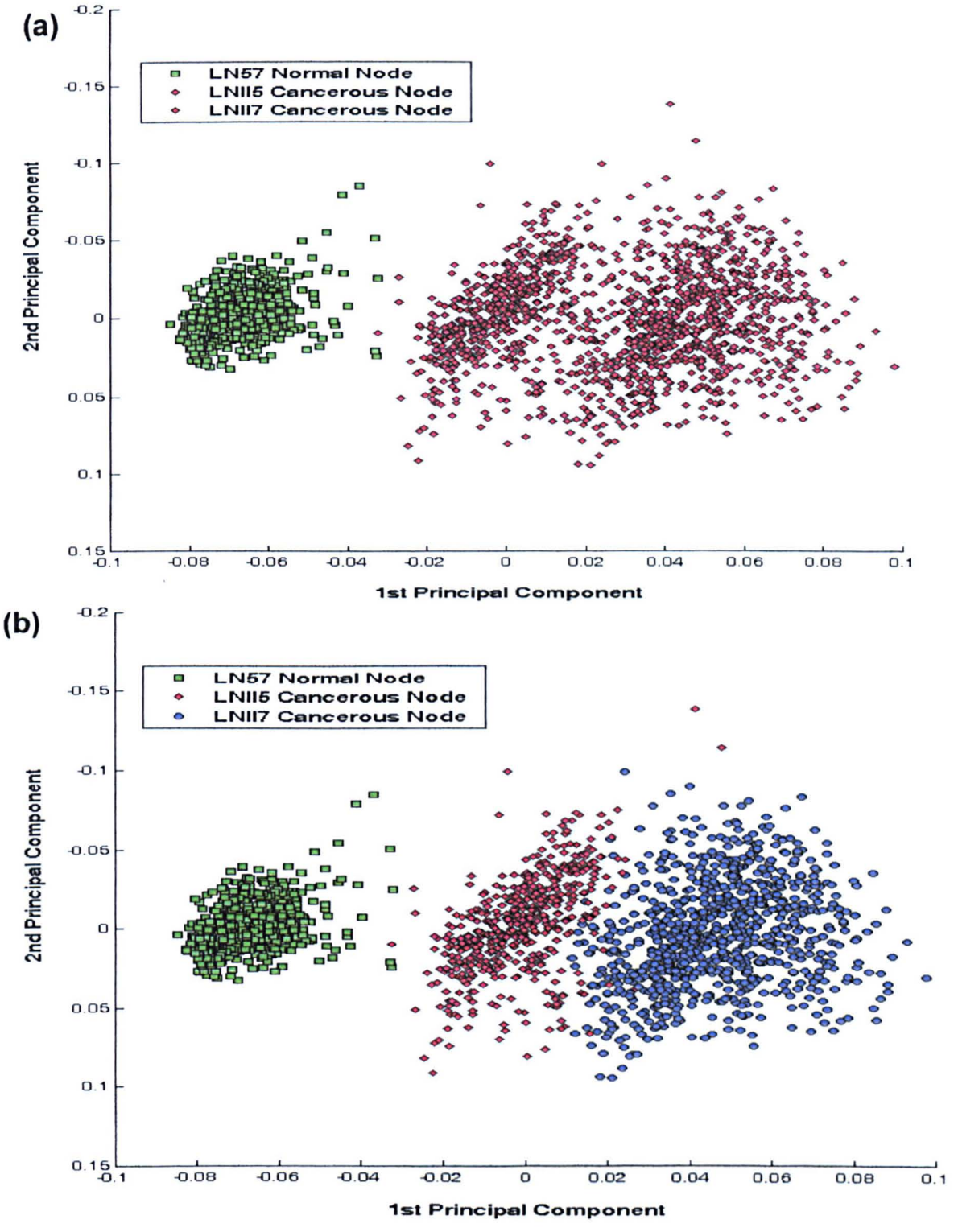

Figure 33: *Clustering results of combined lymph node tissue spectral datasets via the PCA-FCM Algorithm. (a) Healthy cortex tissue spectra from LN57 and malign cortex tissue spectra from LNII5 & LNII7; 2 specified clusters. (b) Healthy cortex tissue spectra from LN57 and malign cortex tissue spectra from LNII7 & LNII6; 3 specified clusters.*

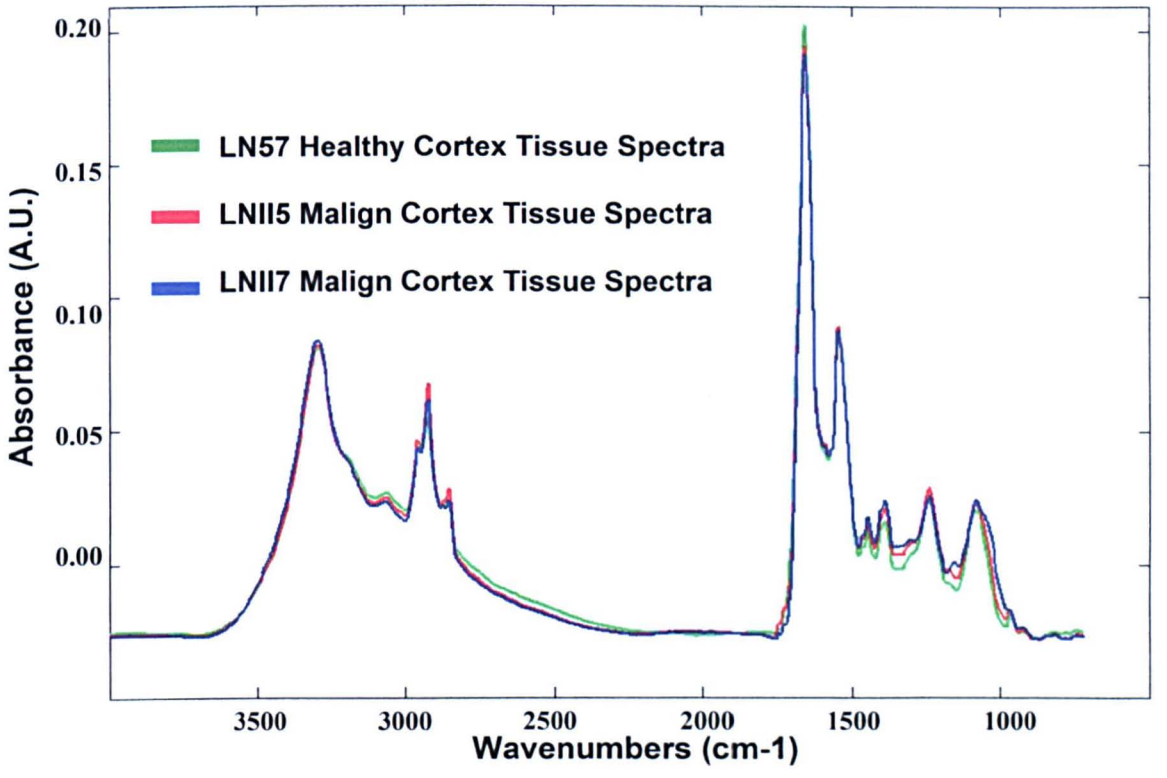

Figure 34: *3 Cluster PCA-FCM Analysis Result. Mean average spectra for each cluster in the analysis.*

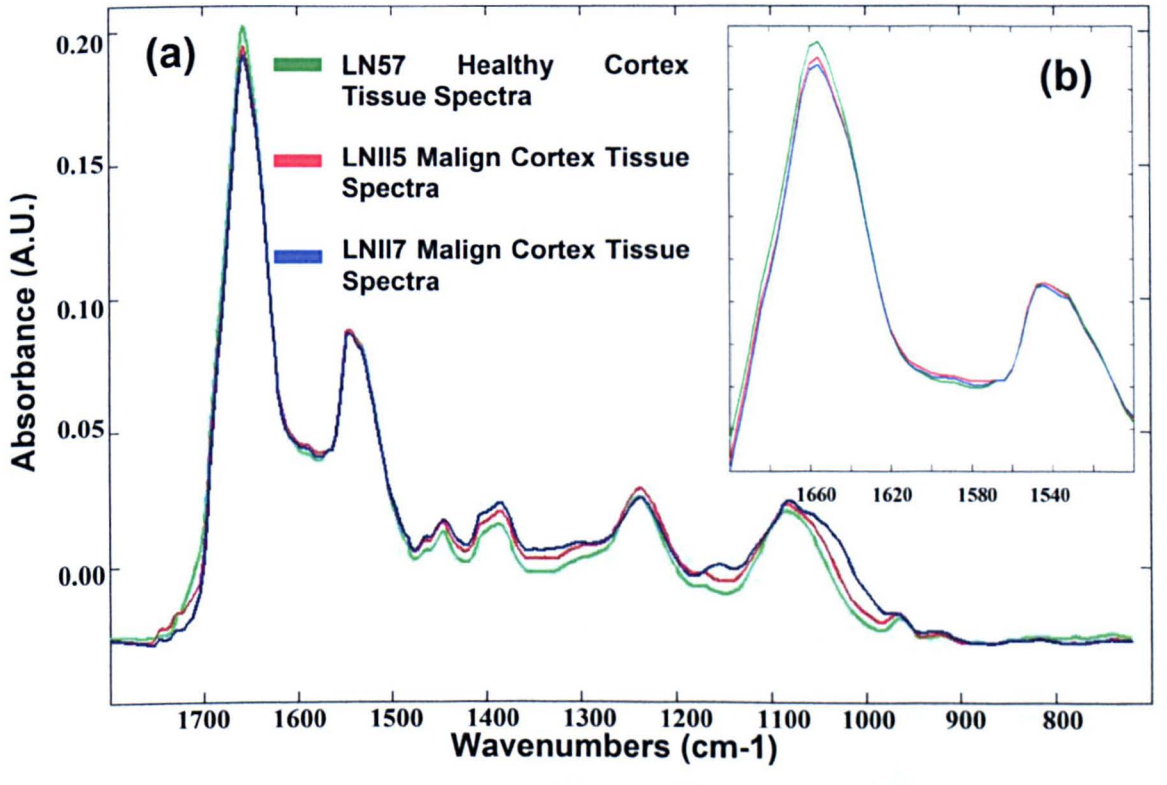

Figure 35: *3 Cluster PCA-FCM Analysis Result. (a) Spectral window displaying mean spectra between 1800 – 720cm^{-1}. (b) Spectral window displaying amide I and II region (1700-1500cm^{-1}).*

spectral changes that there is an overall increase in the nucleic acid concentration of cancerous tissue.

Further spectral changes between the three clusters of tissue spectra can be distinguished within the Amide I – Amide II region (1700 – 1500 cm^{-1}), shown in figure 35b. Spectra originating from the cancerous lymph nodes displayed a significant reduction in the amide II / amide I peak intensity ratio when compared to the healthy tissue spectra. However, a small yet distinctive difference between the two cancerous tissue clusters could be observed, with spectra originating from lymph node LNII7 displaying a greater reduction in this peak ratio. This observed spectral change bares agreement with previous studies examining cervical tissues [35,36], where a reduction in this peak intensity ratio was attributed to disease change.

Taking into account the spectral differences found between the two cancerous lymph nodes LNII5 and LNII7, it is apparent that the greatest amount of natural variation occurring within the malignant cells of these tissues is that of nucleic acid and protein concentrations. An increase in nucleic acid concentration combined with a reduction in amide II / amide I peak intensity ratio are strong indicators of disease change. The tissue spectra collected from lymph node LNII7 displayed the most marked of these spectral changes, and could describe tissue with a more severe degree of malignancy when compared to lymph node LNII5. Any future infrared spectroscopic study of lymph nodes must consider the natural variation of nucleic acid and protein concentrations of malignant cells. Only after this natural variation has been fully assessed may a robust model for automated pathology be created.

2.3.4 Novel Development of Clustering Algorithms for FTIR Spectroscopic Diagnosis of Human Tissues

2.3.4.1 Application and Assessment of Clustering Techniques for Tissue Classification

In an initial study, the performance of three different clustering techniques for tissue spectra classification was assessed. HCA, KM and FCM clustering methods, described in section 4.5, were applied to FTIR spectral datasets collected and provided by John Chalmers et al. [44]. In this work, point spectra were collected from three different oral tissue sections utilising a synchrotron source located at the Daresbury SRS laboratory. Seven small datasets were collected in total and comprised of IR spectra taken from areas upon tissue sections clinically diagnosed as being tumour (abnormal), stroma (healthy connective tissue), early keratinisation and necrotic in nature. The hemotoxylin and eosin stained photomicrograph from one of the tissue sections cut for the analysis is shown in Figure 36a. Both tumour and stromal tissue can be found in this section. These can be identified visually by their dark and light staining respectively. The photomicrograph shown in Figure 36b displays a magnified region of the parallel tissue section cut from the same specimen for spectroscopic analysis. The superimposed dashed white lines on this image describe the boundaries between the two tissue types. In the first experiment, five single point spectra were recorded from each of the three distinct regions, the locations of which are marked by a "+" and numbered 1-5 for the upper tumour region, 6-10 for the central stroma layer and 11-15 for the lower tumour region. The fifteen FTIR transmission spectra from these positions are recorded as dataset 1, and

the corresponding spectra are shown in Figure 37. All datasets in this study were collected in a similar fashion from areas upon the tissue sections that described multiple tissue pathology. Data pre-processing included water vapour subtraction, baseline correction and normalisation.

In order to facilitate the discrimination of the collected spectra into their respective tissue types, Chalmers et al utilised both HCA and PCA analyses to classify spectra according to their spectral characteristics [44]. These calculations were performed using Infometrix Pirouette® (Infometrix, Inc., Woodinville, W.A, USA) software and data limited within the spectral range 1800 – 900cm^{-1}. The results from the multivariate analyses indicated that the partitioning of spectra into their respective

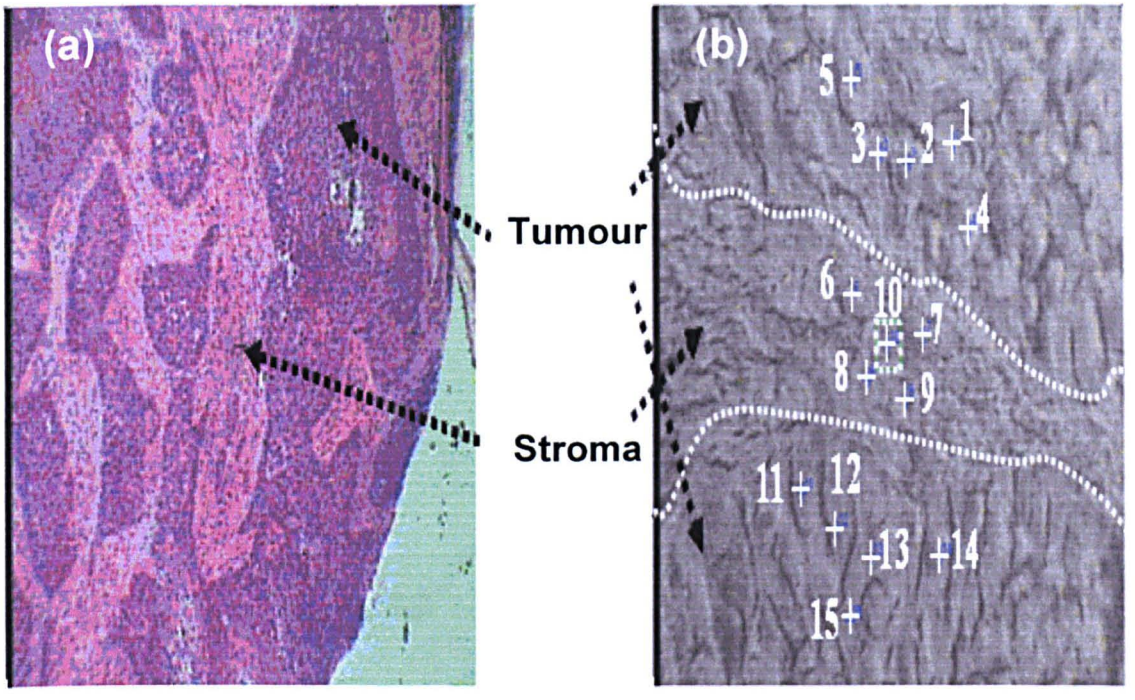

Figure 36: *(a) Photomicrograph of a H&E stained oral tissue section. Dark and light pigmentation describes tumour and stromal tissue respectively. (b) Enlarged 32x photomicrograph of unstained parallel tissue section used for spectroscopic analysis. The upper and lower tumour regions are separated by a stromal layer, the boundaries of which are described by a dashed white line. The numbered cross hairs indicated the co-ordinates at which each spectrum was collected in the analysis.*

tissue types was only achievable via additional pre-treatment to the data. These pre-treatments included mean-centring, variance scaling and the conversion to first-derivative spectra. However, the effectiveness of such data treatments varied according to sample characteristics and the clustering algorithm used. In our experiments we applied the three previously mentioned clustering techniques to these seven FT-IR spectral datasets without any additional pre-treatment.

Agglomerative clustering algorithms can use a variety of different linkage methods. These provide a measure of similarity between clusters based upon the data held within each cluster [41,45]. The main linkage methods include: single link, complete link and minimum-variant algorithms (Ward's algorithm) [46,47]. Other linkages

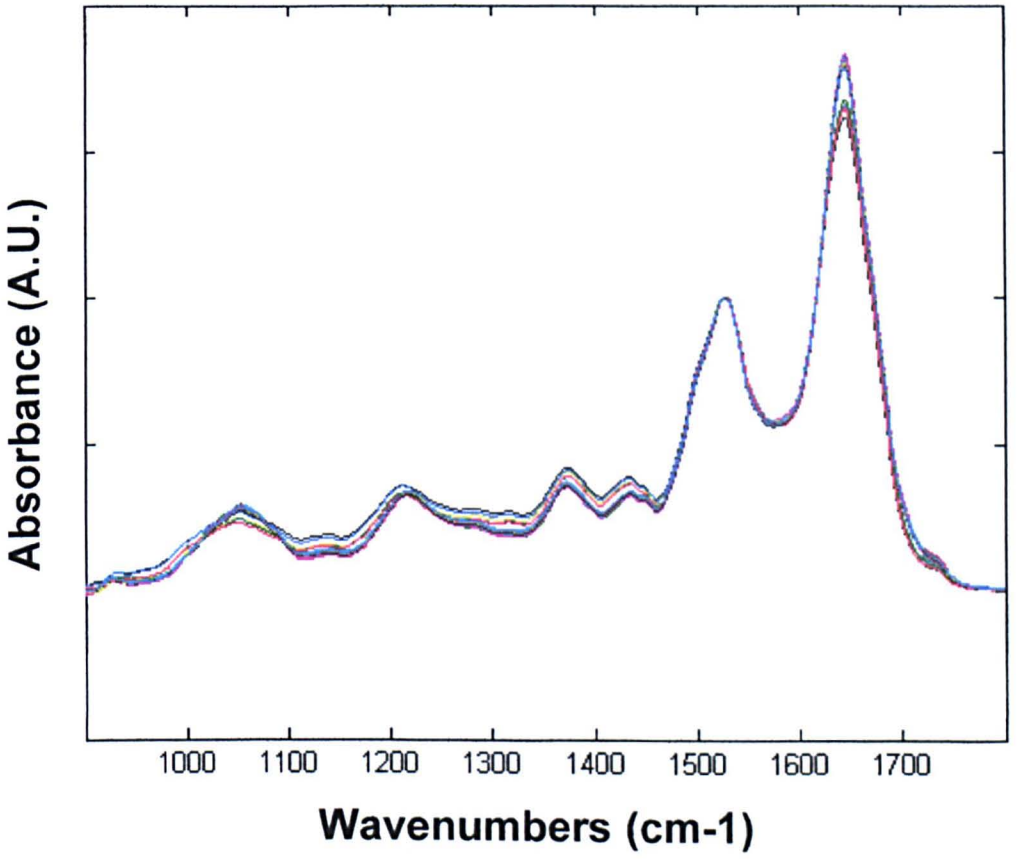

Figure 37: *Spectral window displaying FTIR spectra from Dataset 1 (1800-900 cm^{-1}).*

are derivatives of these main methods. In the single link algorithm, the distance between two clusters is measured by the two closest data points held within the different clusters. By contrast, in the complete link algorithm, the distance between two clusters is measured between the two furthest data points within the different clusters. The minimum-variant algorithm is distinct from the other two methods as it uses a variance analysis approach to measure the distance between two clusters. In general, this method attempts to minimise the sum of squares of any two hypothetical clusters which can be generated at each step. This is based on the Euclidean distance between centroids [41]. Each linkage method can thus provide a different agglomerative clustering result. We therefore individually applied each linkage method for HCA clustering analysis. The final result achieved by KM and FCM clustering techniques can also be sensitive to the initial randomisation step made by the algorithms. This can therefore allow the iterative clustering steps to follow several different routes to completion. To assess the different clustering distributions that may arise from these algorithms, each method was repeated 10 times. The amount of clusters identified by each method was set to match the amount of tissue types identified clinically. The number of spectra collected from each tissue type identified both clinically and via cluster analysis for each dataset is shown in Table 2.

In Table 2 we can see that in most datasets, the number of spectra that belong to each tissue category does not exactly match those identified clinically. These discrepancies arise from the misclassification of tumour spectra into the stroma cluster and vice versa. For example, in dataset 2, using the hierarchical clustering single linkage method, the number of spectra considered as being tumour is 17, while

Datasets names	*Tissue types*	*Clinical study*	*Hierarchical clustering*			*KM*	*FCM*
			Single	*Complete*	*Ward*		
Dataset 1	Tumour	10	10	10	10	10	10
	Stroma	5	5	5	5	5	5
Dataset 2	Tumour	10	17	9	9	9	9
	Stroma	8	1	9	9	9	9
Dataset 3	Tumour	8	4	8	7	3 6 4	4
	Stroma	3	7	3	4	8 5 7	7
Dataset 4	Tumour	12	19	12	12	11 19 13	19 11
	Stroma	7	5	7	7	8 5 6	5 8
	Early keratinisation	12	7	12	12	12 7 12	7 12
Dataset 5	Tumour	18	1	18	18	14 17	14
	Stroma	12	29	12	12	16 13	16
Dataset 6	Tumour	10	10	10	10	10	10
	Stroma	5	5	5	5	5	5
Dataset 7	Tumour	21	28	17	15	17	18
	Stroma	14	13	18	20	18	16
	Necrotic	7	1	7	7	7	8

Table 2: *Oral tissue spectra classification results via HCA, KM and FCM methods of clustering. The number of spectra classified into each cluster can be directly compared to that identified clinically.*

only 1 spectrum is considered as being stroma. We will regard the extra spectra that are partitioned into each cluster as the number of disagreements in classification compared to clinical diagnosis. These comparison results are displayed in Table 3.

After repeating each clustering technique 10 times, it can be seen that the KM and FCM algorithms obtained more than one clustering result in some datasets. As previously mentioned, this is due to a random initialisation step used by both algorithms to locate the initial cluster centres. Examining the results shown in Tables 2 and 3 in greater detail, the KM technique displays a greater number of variations in the clustering result (3 out of 7 datasets) when compared to the FCM clustering method (1 out of 7 datasets). The frequency of each clustering variation for these datasets is further detailed in Table 4.

Datasets names	*Tissue types*	*Hierarchical clustering*			*KM*			*FCM*	
		Single	*Complete*	*Ward*					
Dataset 1	Tumour	0	0	0	0			0	
	Stroma	0	0	0	0			0	
Dataset 2	Tumour	7	0	0	0			0	
	Stroma	0	1	1	1			1	
Dataset 3	Tumour	0	0	0	0	0	0	0	
	Stroma	4	5	3	5	2	4	4	
Dataset 4	Tumour	7	3	3	3	7	3	3	7
	Stroma	5	3	3	4	5	2	4	5
	Early keratinisation	0	0	0	0	0	0	0	0
Dataset 5	Tumour	0	0	0	0		0	0	
	Stroma	17	0	0	4		1	4	
Dataset 6	Tumour	0	0	0	0			0	
	Stroma	0	0	0	0			0	
Dataset 7	Tumour	7	0	0	0			0	
	Stroma	0	4	6	4			2	
	Necrotic	1	0	0	0			1	

Table 3: *Comparison of oral tissue classification via HCA, KM and FCM methods of clustering. These are based upon the number of disagreements made in direct comparison to clinical diagnosis.*

Datasets names	*KM*			*FCM*	
Dataset 3	2/10	3/10	5/10	-	
Dataset 4	3/10	3/10	4/10	9/10	1/10
Dataset 5	5/10		5/10	-	

Table 4: *Clustering variations made by KM and FCM analysis. Results are displayed as the amount of times each result was reached out of ten repetitions.*

	Hierarchical clustering			*KM*	*FCM*
	Single	*Complete*	*Ward*		
Average Number of Disagreements	48	16	16	18.8	19.5

Table 5: *Average number of disagreements made by each clustering technique.*

In order to further investigate the performance of each clustering method, the average number of disagreements for all datasets was calculated, and is shown in Table 5. It can be seen that the hierarchical clustering single linkage method displayed the worst overall performance, whereas the complete linkage and Ward methods displayed the best overall clustering results. However, a major drawback of agglomerative techniques is that these methods are computationally expensive. For such algorithms, the CPU time required can be scaled to the square of the amount of objects in a dataset (proportional to n^2, where n is the number of spectra in a dataset) [32]. These requirements become more and more important with large datasets (n>1000), where a great amount of dependence is placed upon searching for the global minimum in the distance matrix. The KM and FCM clustering techniques gave a similar performance when compared to the Wards agglomerative method. For these clustering methods, the computational effort scales in a first-order approximation linearly with n. Hence, when compared to hierarchical clustering, these techniques will be far less time consuming for extensively large datasets. Moreover, although KM has a slightly better performance than FCM, displaying slightly fewer disagreements on average, it can bee seen from Table 4 that KM exhibits far more variation in its final clustering result than FCM. Therefore, when taking all these factors into consideration, the FCM clustering method appears more suited for rapid analysis of very large and complex spectroscopic datasets that are recorded in this study of human cells and tissues.

2.3.4.2 A comparison of Fuzzy and Non-Fuzzy Clustering Techniques when applied to a large spectral dataset

In this study, FCM (Fuzzy) and K-Means (non-Fuzzy) clustering techniques were used to cluster an IR spectral dataset collected from a large spatial area of an axillary lymph node tissue section. The map or IR image created was composed of 7497 spectra, a significantly higher number than previously examined, and proves a sterner test when assessing the diagnostic ability of each clustering method. In previous work, the datasets used in our analysis (seven sets of oral IR tissue spectra) had a comparatively small number of spectra, n ($n<50$). By contrast, the experiments reported in this study were carried out upon an IR map where $n>1000$. All spectra were recorded using a spectral resolution of 8 cm^{-1}, over the spectral range 4000 – 720 cm^{-1}. Each spectrum therefore comprised 821 data points (4 cm^{-1} data point interval). This can alternatively be visualised as 821 different dimensions in multivariate space. In this circumstance, the clustering methods have a substantially high computational requirement. It is apparent that if we can reduce the dimensionality of the original data without losing a significant amount of useful information, the performance of the clustering algorithms will be computationally more efficient. Principal component analysis was therefore used to reduce the dimensionality of the original data [48,49]. The dataset was projected onto its first 10 principal components and cluster analysis carried out upon the data in these dimensions. These first 10 principal components comprised over 99% of the variance contained in the original data. Therefore the data had been reduced by 811 dimensions with a loss of only 1% of the variance contained in the original data. A comprehensive comparison between techniques was achieved by subjectively setting the amount of clusters found by the analyses from 2 to 9. Each experiment was repeated 10 times to assess any variation in the clustering result. For the KM method, the squared Euclidean distance was used as a distance measure; the initial

cluster centre positions being randomly selected. In FCM, the fuzziness index m was set to a value of 2. For both clustering methods the maximum number of iterations was set to 100. A stop criterion was set to end the analyses when the minimum amount of improvement in the cluster density was found to be smaller than 10^{-5}, previously utilised in a number of spectroscopic studies [32,50,51].

The IR map used in these experiments is the same as that reported in section 2.3.1 and will therefore not be discussed in any great detail. A full description of the lymph nodes histological architecture can be found in section 2.3.1.1. The H&E stained photomicrograph and total absorbance IR image for the spatial area mapped are shown in Figure 38.

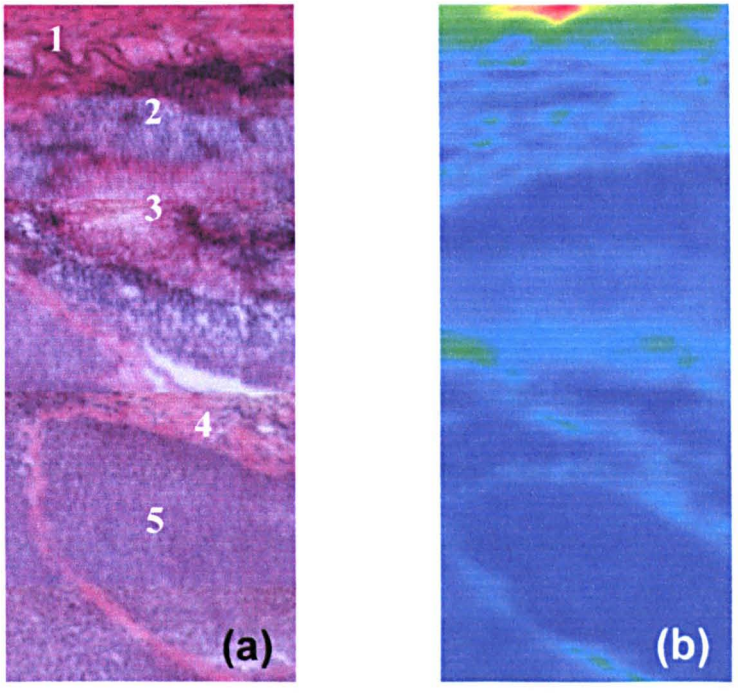

Figure 38: *a) Photomicrograph of the H&E stained parallel lymph node tissue section. (1) Capsule, (2) cortex, (3) secondary follicle, (4) reticular cells and (5) invading breast cancer tissue. b) Total absorbance IR image of same region.*

Since each spectrum contained in the IR map has a unique spatial x, y position, false-colour images could now be generated to describe the cluster analysis results as a function of their spatial position. By assigning each cluster a colour, these colours can then be plotted as pixels at the x, y coordinates from which the spectrum was collected. Therefore, pixels with the same colour in the false image are spectra that were grouped together into the same cluster.

During the initial stage of the experiments, the FCM method produced considerably varied results. Figure 39 displays the 3 different clustering results that were obtained when constraining the algorithm to find 2 clusters. In these examples the FCM clustering method performed particularly badly and gave very unstable clustering results. The two main tissue types found within this map are fibrocollagenous tissue (capsule and reticulum) and the remaining nodal tissue (secondary follicle, invading breast cancer, cortex), which have the most distinct of spectral differences. However, these tissue types were thoroughly mixed in all clustering results, the least being found in the first clustering scenario depicted in Figure 39a.

Based on this observation, we went back and studied the dataset when projected onto its first three principal components, which comprised over 93% of the original variance. The ranges of each component were found to be:

1st Principal Component	[-0.0075, 0.0751]
2nd Principal Component	[-0.0117, 0.0069]
3rd Principal Component	[-0.0096, 0.0047]

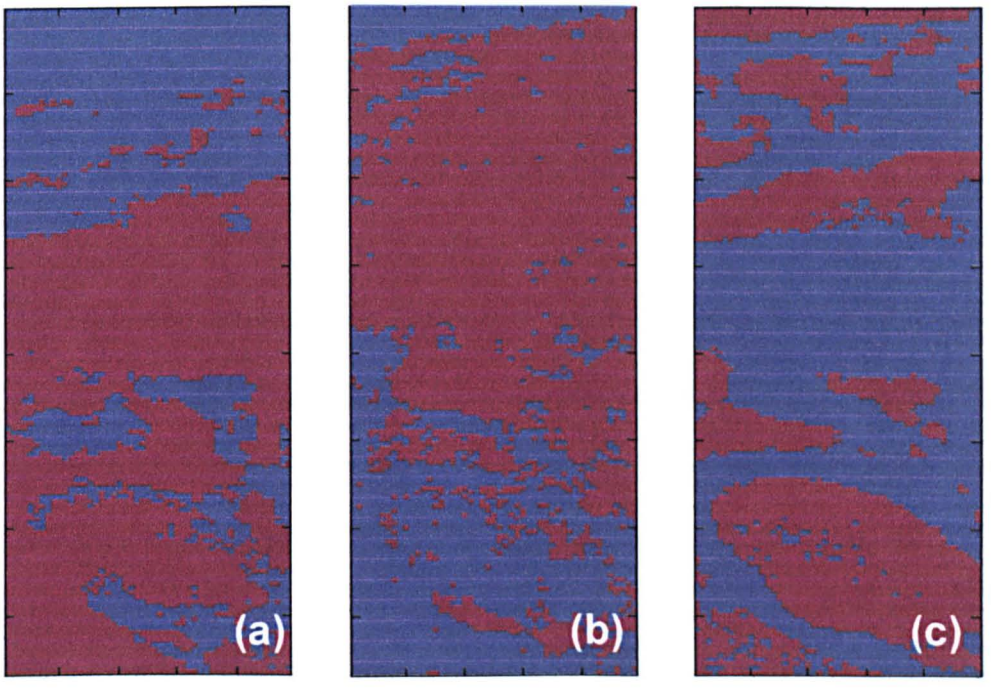

Figure 39: *FCM clustering results when constrained to a 2 cluster analysis. (a) – (c) display the three clustering results achieved via the analysis.*

The order of -0.0075 is -10^{-2} and 0.0751 is 10^{-1} and so on. Thus their corresponding range sizes are 10^{-1}, 10^{-2} and 10^{-2}. The smaller the size of the component range, the more compact the data. Therefore, the distance between the data and their ideal centres would be smaller. In FCM, the objective function *J(U,V)* (see equation 9, section 4.5.4.2) is proportional to the squared Euclidean distance between each datum and the cluster centres. In this case, the squared Euclidean distances would now be even smaller and within the range of 10^{-1} and 10^{-4}. Hence, a small range size may lead to a very small objective function value. When the difference between two objective function values is less than the minimum amount of improvement that can be made in a further iteration, the algorithm stops the clustering process. Therefore, if the minimal amount of improvement was not small enough (i.e. 10^{-5} in this scenario) to allow improvements upon the cluster centre positions, the FCM performance is significantly reduced. Due to this finding, we then utilised a value of

10^{-7} as the minimal amount of improvement for the clustering experiments. It was found that the performance of the FCM method was significantly improved and now achieved stable clustering results for all defined cluster numbers.

The KM clustering method also displayed a marked improvement in stability after this change in methodology, but did however exhibit two clustering structures when constrained to a 2 cluster analysis. The cluster structures for both FCM and KM methods when constrained to a two cluster analysis are shown in Figure 40.

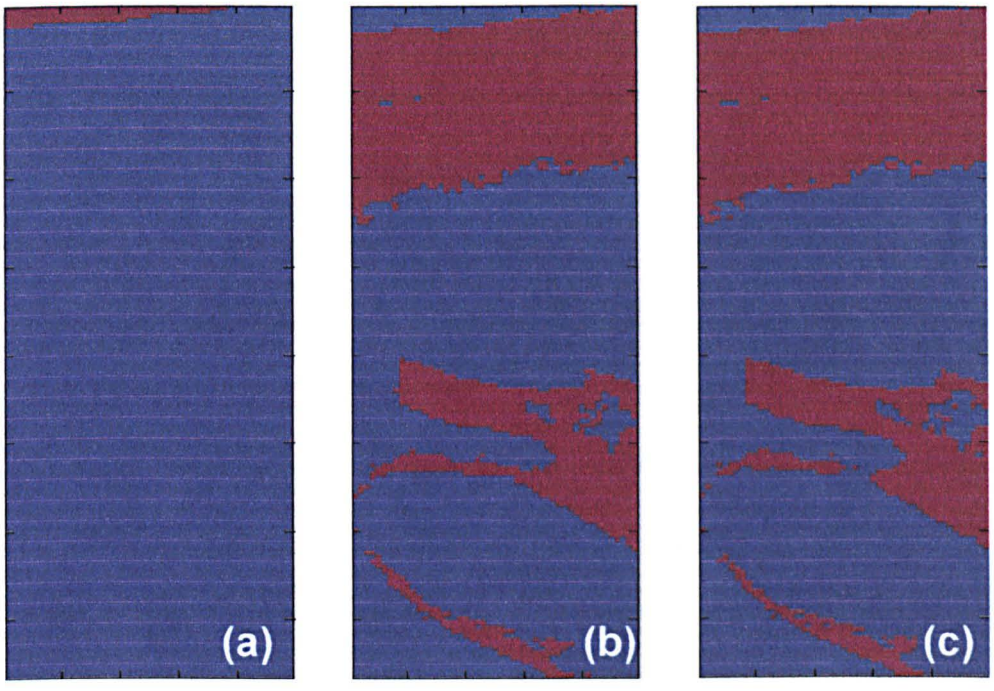

Figure 40: *Clustering results obtained when conducting a two cluster analysis using both KM (a & b) and FCM (c) methods.*

The first clustering structure obtained by the KM method shown in Figure 40a clearly separates the fatty tissue and remaining tissue types. The second clustering structure found by the KM method shown in Figure 40b is very similar to that obtained via the FCM method displayed in Figure 40c. This clustering result

describes two clusters that comprise fibrocollagenous (capsule, reticulum) and nodal tissue (cortex, secondary follicle, invading breast cancer). The variation in clustering results made by the KM method can be attributed to the positioning of the fatty tissue spectra in multidimensional space (see figure 18, section 2.3.1.7). KM clustering is very sensitive to outliers in a dataset, and if one data point is assigned to one cluster rather than another, the results may substantially distort the distribution of the data. The fatty tissue spectra being outliers have therefore distorted the clustering process to achieve this clustering structure.

Figure 41 displays the clustering results achieved via KM and FCM clustering when the number of clusters were subjectively increased from 3 – 5. When the number of clusters was set to 3 in the KM analysis (see figure 41a) the analysis partitioned the spectra into groups that describe fatty tissue (blue), capsule and reticulum (red), and nodal tissue whether healthy or diseased (green). In comparison, the three clusters found by FCM analysis (see figure 41d) describe cancerous and fatty tissues (red), capsule and reticulum (blue), and the secondary follicle (green). In general the FCM method clusters the spectra into the three main tissue types present. These include cancerous tissue, the secondary follicle containing highly proliferating cells, and fibrocollagenous tissue. These clusters are nevertheless partially intermingled; i.e. cancerous tissue being mixed with fatty tissue. However, when directly comparing the two clustering techniques it is apparent that the FCM method was able to differentiate the invading cancerous tissue.

When the number of clusters is increased to 4 and 5 (see figures 41b, c, e, f) the KM and FCM methods further partition the fibrocollagenous tissues into their individual

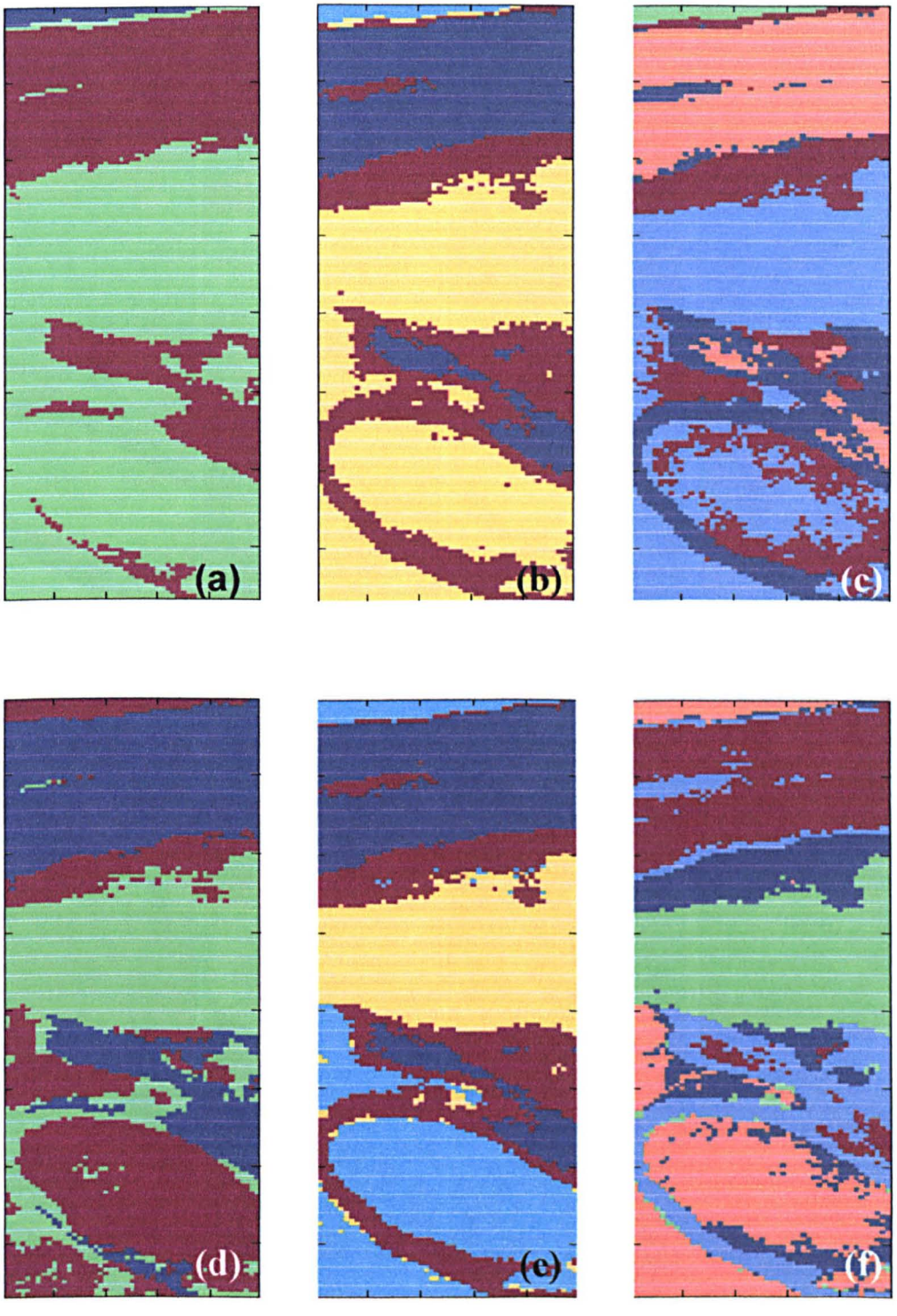

Figure 41: *Clustering results obtained when subjectively increasing the cluster number from 3 – 5. (a) – (c) display clustering structures for 3 – 5 clusters via KM clustering. (d) – (f) display clustering structures for 3 – 5 clusters via FCM clustering.*

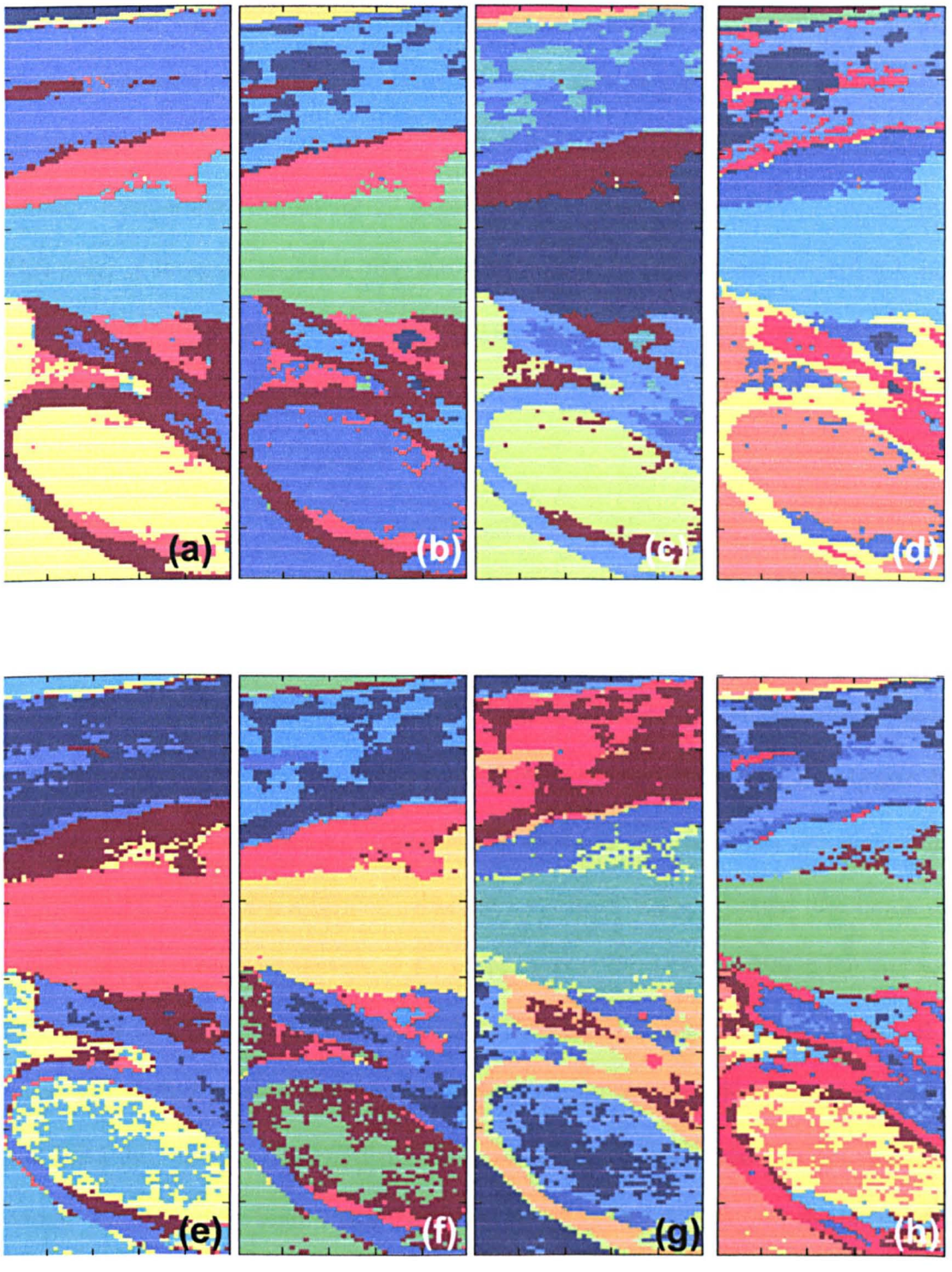

Figure 42: *Clustering results obtained when subjectively increasing the cluster number from 6 – 9. (a) – (d) display clustering structures for 6 – 9 clusters via KM clustering. (e) – (h) display clustering structures for 6 – 9 clusters via FCM clustering.*

subtypes (capsule and reticulum) and also the cortex tissue spectra into a separate cluster. When comparing these cluster images against the H&E stained parallel tissue section, it can be seen that the 5 cluster result obtained via the FCM method is directly comparable, displaying all tissue types defined via conventional histology. However, it is again apparent that the KM method was not able to partition the cancerous tissue spectra into a separate cluster. This is a very useful observation and could prove to demonstrate that k-means clustering is limited and unsuitable for diagnostic purposes.

Figure 42 displays the clustering results obtained for both KM and FCM methods when subjectively increasing the cluster number from 6 – 9. Starting at 6 clusters, the K-Means algorithm begins to separate the cancerous and secondary follicle tissue spectra (see figure 42a). The FCM algorithm on the other hand has begun to partition tissue spectra that exist around the edges of the reticulum and cancerous areas that may describe a further subset of cancerous tissue whereby the grade of malignancy may not be as severe or far reaching (see figure 42e). When the number of clusters is further increased to 7 and 8, both clustering algorithms partition the tissue spectra that exist within the capsule into several subsets which includes a layer that lines the region. The FCM method also classified more mixed types of tissue within the cancerous region (see figures 42b, c, f, g). Finally, when the cluster number was increased to 9, the results from both KM and FCM showed more and more tissue types being mixed together (see figure 42d and h). These additional subsets of tissue spectra characterised by the clustering algorithms may be representative of potential subtypes of tissue that can not currently be identified via

conventional histopathology and thus be useful for diagnostic purposes. However, they could also be attributed to clustering noise.

In conclusion, the FCM clustering method was far more effective at discriminating and further partitioning the cancerous tissue spectra at a much earlier stage in the clustering process. However, the fatty tissue spectra were not discernable at any stage in the study, no matter how high the cluster number was increased. This again highlights that outlier data can often be detrimental to efficient clustering analysis. As the number of clusters was increased, more and more information about the tissue section was revealed, possibly uncovering further subtypes of tissue presently unidentifiable by conventional histopathology.

2.3.4.3 A fully automated FCM based clustering algorithm

When using standard FCM algorithms, the number of clusters determined by the analysis has to be specified a priori. This can be a disadvantage in many real world applications where the correct or 'optimal' number of clusters that best describes a dataset is often an unknown measure. However, with the use of cluster validity indices, it is possible to discover the 'optimal' number of clusters within a given dataset dependent upon the clustering structure [52]. In short, clustering validity is a concept to evaluate the quality of each possible clustering structure and thus determine the number of clusters that best represents the given dataset. Many different cluster validity indices have been proposed for evaluating fuzzy clustering. Indices which utilise fuzzy membership values such as the partition coefficient and partition entropy have the advantage of being easy to compute [53], but are only

useful for a small number of highly separated clusters in multidimensional space. In order to overcome these problems, Xie and Beni defined a validity index which measures both the compactness and separation of clusters [54]. This validity index has been frequently utilised in recent research and has shown the ability to correctly define the 'optimal' amount of clusters in several different situations [55,56]. We therefore decided to apply this validity index to our FCM algorithms as to determine whether such an index could allow a fully automated and accurate classification of tissue spectra.

Pilot studies using the Xie – Beni validity index were again carried out upon the seven oral tissue datasets previously described in section 2.3.4.1. By use of the traditional FCM method, a recently developed Variable String Length Simulated Annealing (VFC-SA) algorithm [57], and a novel developed Simulated Annealing Fuzzy Clustering (SAFC) algorithm, the seven datasets were clustered and the results compared [37,38]. This was achieved by comparing the final cluster number calculated by the automated clustering techniques with clinical diagnosis, and scrutinising their resultant validity index values. It was shown that the SAFC algorithm produced the best validity index values with improved cluster compactness and separation. The algorithm also obtained the same amount of clusters as defined by clinical diagnosis in four out of the seven datasets. In the remaining datasets, the amount of clusters did differ from those established by histology, identifying an excessive number in most cases. This was due to a smaller and thus better validity measure being obtained when adopting these clustering structures. These contrasting results could be due to a variety of different factors. The additional clusters could be diagnostic of further subtypes of tissue that describe different severities of disease or

cellular change. But it is also very likely to be a consequence of the small amount of spectra contained within these datasets, and thus not sufficiently assessing the natural variation that may occur in these types of tissue. A sterner test was therefore required to assess the effectiveness of using such a validity measure upon larger more complex datasets. Although the simulated annealing clustering process performed well on these small datasets, the algorithm can become computationally expensive when applied to large datasets where spectra numbers exceed 1000. We therefore adopted an FCM based model selection algorithm for automated clustering of large spectral datasets. This algorithm is based upon the standard FCM method whereby c_{min} and c_{max} represent the minimal and maximal number of clusters that the dataset may contain. The final clustering structure (C) is returned based upon the optimal Xie-Beni validity index value (V_{XB}). The algorithm is performed in the following steps:

1) Set c_{min} and c_{max}

2) For c=c_{min} to c_{max} ;

   2.1 Initialise the cluster centres.

   2.2) Apply the standard FCM algorithm and obtain the new centre C and new fuzzy partition matrix *U*.

   2.3) After the FCM reaches its stop criteria, the cluster validity is calculated (e.g. V_{XB}).

3) Return the best data structure (C) that corresponds to the optimal cluster validity value (e.g. the minimum V_{XB}).

In our experiments the values of c_{min} and c_{max} were set to 2 and 10 clusters respectively. To reduce the number of variables involved in the analysis, the first ten principal components for each dataset were again calculated and used for clustering.

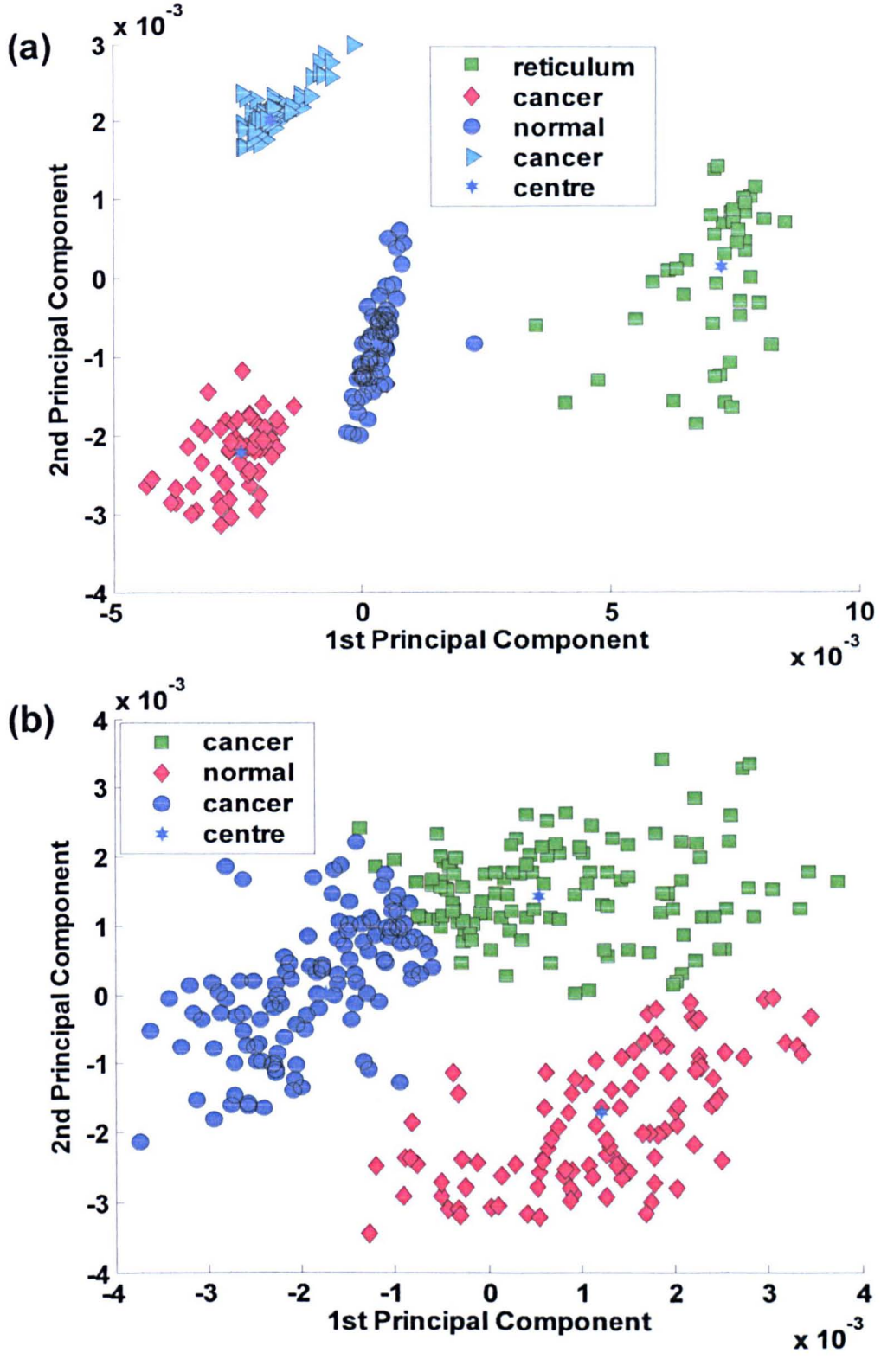

Figure 43: *Clustering results of lymph node tissue spectra obtained via the automated FCM based model selection algorithm. (a) Positive lymph node LNII5. (b) Positive lymph node LNII7.*

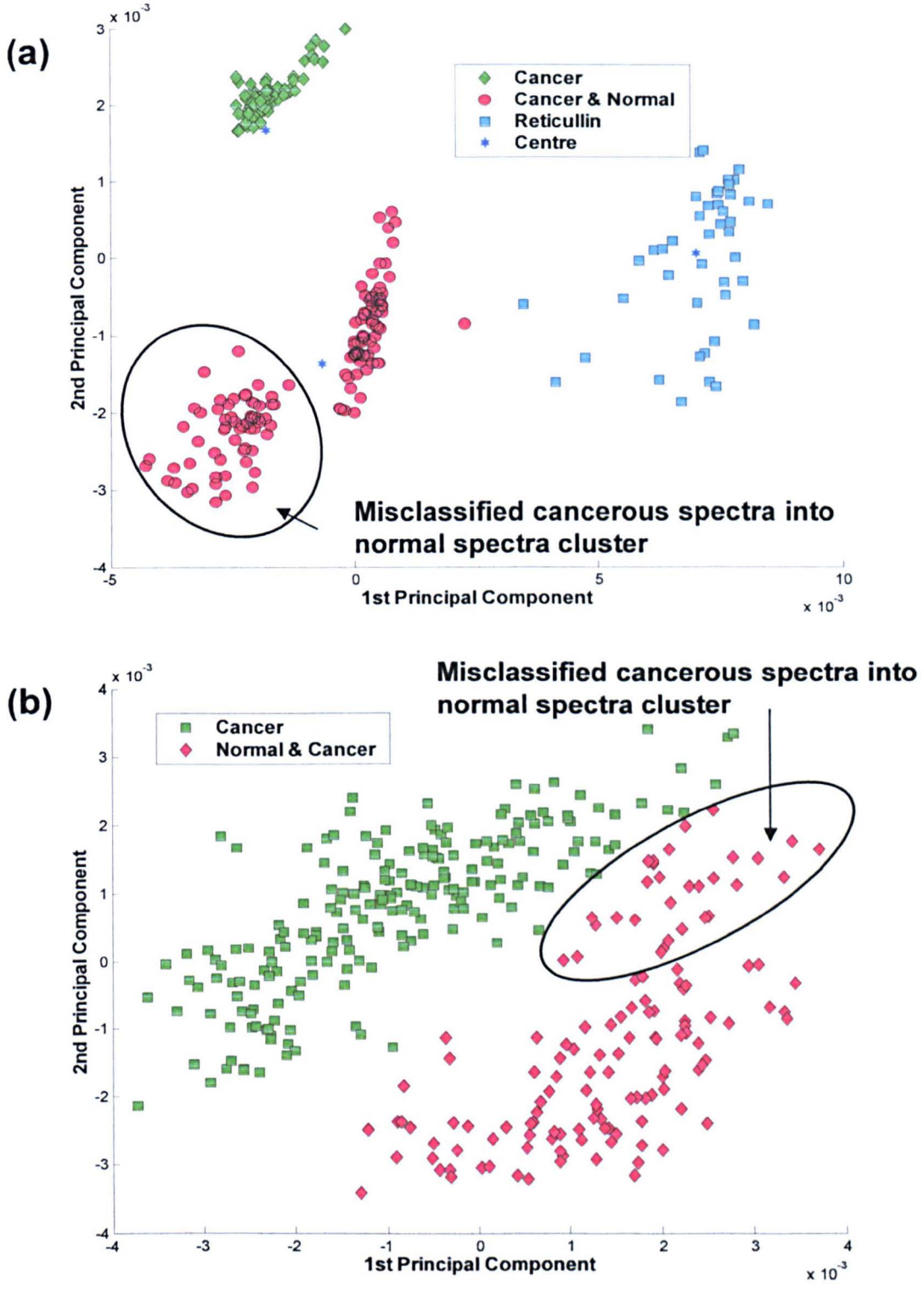

Figure 44: *Clustering results of lymph node tissue spectra obtained via the Standard FCM Algorithm. (a) Positive lymph node LNII5; 3 specified clusters. (b) Positive lymph node LNII7; 2 specified clusters. In both diagrams, it can be seen that a number of spectra have been misclassified into different clinical clusters (misclassified spectra have been encircled).*

In a more comprehensive study, the proposed automated algorithm was applied to a number of large spectral datasets that had been collected from a variety of different lymph nodes. The results from two particularly interesting positive (cancerous) lymph nodes, named LNII5 and LNII7 are shown in Figures 43a and 43b. By plotting the original data onto the first two principal component dimensions, which incorporate ca. 80% of the original variance, the approximate clustering structure can be visualised.

When studying these plots in more detail, it can be seen that although the algorithm has adopted the clustering structure that produces the best cluster validity value, an excessive amount of clusters have been identified when compared to clinical analysis. In both examples the algorithm identified two clusters that describe the cancerous tissue spectra. However, when comparing these clustering results with those calculated using the standard FCM algorithm (see Figures 44a and 44b), whereby the cluster number was chosen to match clinical diagnosis, it can be seen that the automated algorithm was not able to completely segregate the different types of tissue spectra into separate clusters.

Additional studies were carried out upon a variety of different lymph node spectral datasets, which included a number of different tissue types. These verified that the automated FCM method could give two clustering outcomes. It would either produce a clustering structure that matched histological analysis or generate an excessive number of clusters that partition the tissue types into multiple groups. When further examining the excessive cluster structures, it became apparent that this outcome may be due to two main reasons. The first was the cluster validity index

itself, where all distances between data points and cluster centres are calculated using the Euclidean distance. This could result with inefficient clustering when compared to histology if the shape of the clusters differed greatly from the ideal (spherical). The second was the possible identification of tissue subtypes. Characteristic IR cell signatures that describe different stages of cellular change may have been discovered by the clustering process. However, these could also be attributed to the natural variation in chemical composition of the clinically defined tissue types. Nevertheless, at this stage of study, we would prefer to cluster all tissue spectra with the same clinical diagnosis into the same cluster. In order to solve this problem, it is required to combine or "merge" the clusters with the same clinical diagnosis together. Although the separate clusters could represent different stages of cellular change, they may also have similar properties that make them discernable from different tissue types. This information is most likely to be held within the IR spectra themselves, rather than the clustering structure or cluster validity measures. A new method was therefore developed that utilised the chemical information contained within the IR spectra of the clusters to successfully merge separated clusters together. The detailed mechanism used by this algorithm will not be discussed in detail, but can be found in section 4.5.4.4. In short, the FCM based automatic selection algorithm initially partitions the spectra into different clusters, and then an automated merge method is used to assess the clustering structure. At this point the algorithm can either decide to accept the clustering structure or proceed to merge similar clusters together using their average spectra until a stop criterion has been reached.

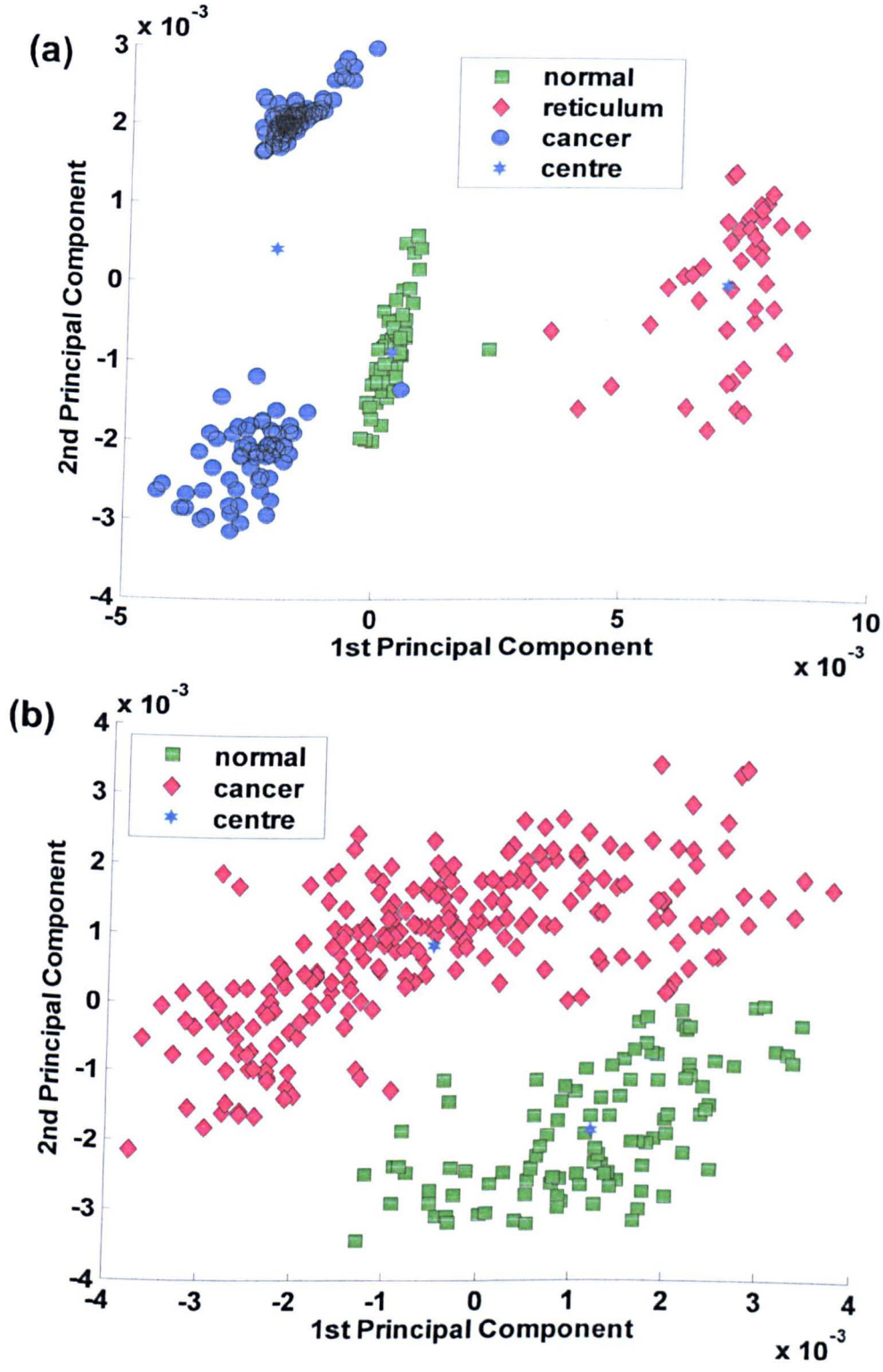

Figure 45: *Clustering results of lymph node tissue spectra obtained via the automated FCM based model selection algorithm and combined merge method procedure. (a) Positive lymph node LNII5. (b) Positive lymph node LNII7.*

This novel fully integrated FCM based merge method algorithm was again applied to the extracted spectral datasets collected from positive lymph nodes LNII5 and LNII7. The clustering results achieved are shown in figures 45a and 45b respectively. Again these have been plotted onto their first two principal components, thus aiding visualisation of the clustering structure. It can clearly be seen that in both datasets the tissue spectra have now been correctly partitioned into the same amount of clusters as defined by histology. The misclassified spectra previously grouped incorrectly via the standard FCM algorithm have also now been correctly classified into the same tissue type groups.

This novel clustering method was applied to all of our previously problematic infrared spectral datasets for which previous approaches could not obtain the correct number of clusters. For each dataset, the proposed method identified the correct amount of clusters. It should be noted that after merging clusters a small number of spectra were still misclassified (approximately 1 – 5 spectra per dataset). However, the overall clustering accuracy was significantly improved. An additional imaging example of the clustering benefits is shown in Figure 46. In this example a large IR map was collected from a positive lymph node tissue section and contained 5764 spectra. The spatial area examined characterised a region upon the tissue section where normal nodal tissue was met by invading cancerous tissue. A clear boundary between the two tissue types could be seen and is further depicted in the photomicrograph shown in Figure 46b. As previously described, the random initialisation procedure used by these types of algorithm can lead to different clustering structures. The experiments were therefore repeated 10 times and stopped after the first part of the new routine, whereby the clustering structure is initially

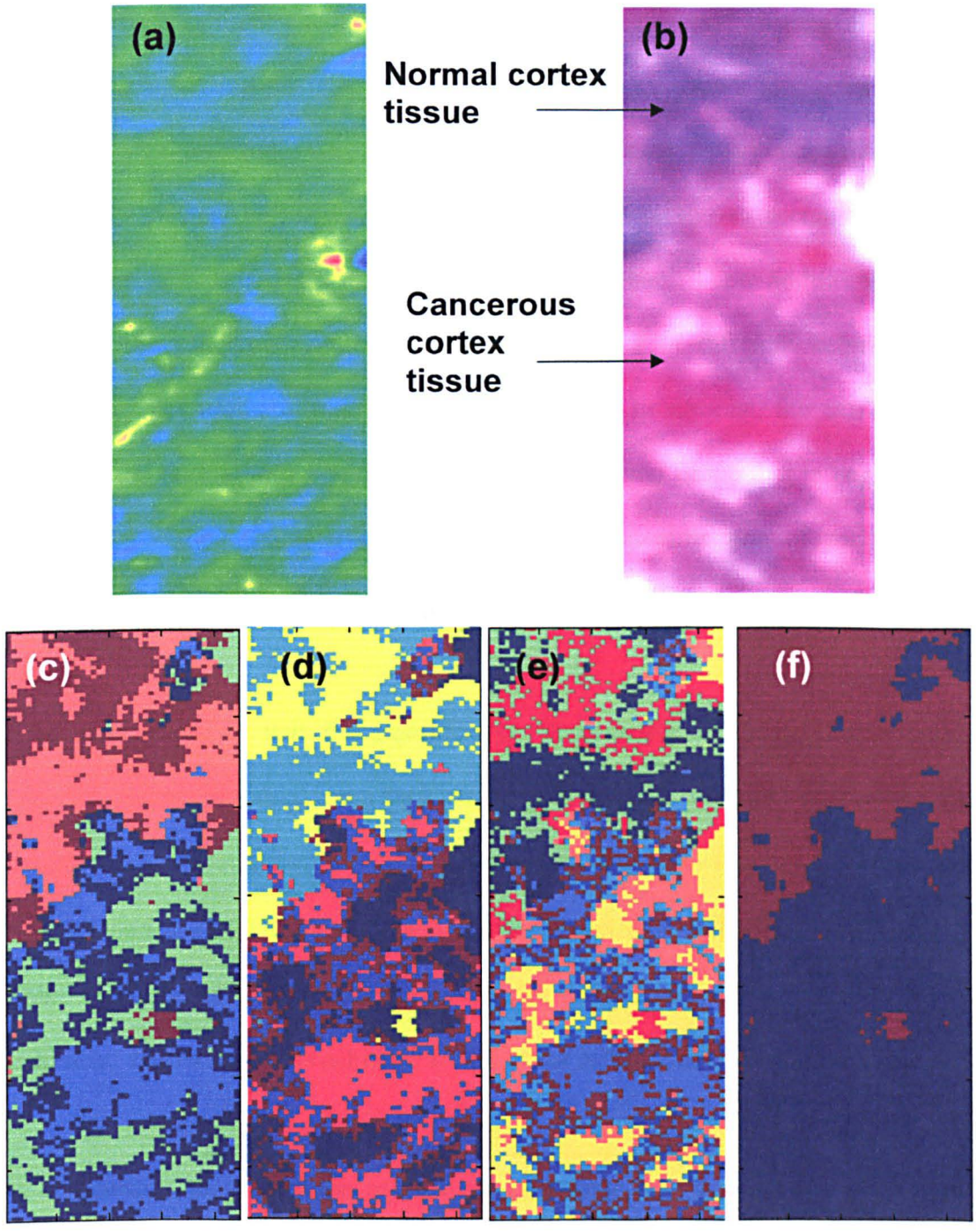

Figure 46: *Clustering results of an IR map collected from a positive lymph node. These were obtained via the automated FCM based model selection algorithm and combined merge method procedure. Sampled area was 275μm x 818.75μm in size. (a) Total absorbance IR image. (b) H&E stained photomicrograph of parallel tissue section. (c) – (d) Initial clustering structures obtained after FCM based selection algorithm. These describe a clustering structure of 5, 6 and 9 clusters respectively. (f) Final clustering result obtained via automated merge method procedure – this image contains two final clusters of IR spectra.*

defined by the optimal cluster validity value. Three different clustering structures were obtained during this step of the algorithm and are shown in figures 46c – e. These false colour images describe the formation of clustering structures that provide 5, 6 and 9 clusters respectively. All of these initial clustering scenarios describe multiple groups for both the normal and cancerous tissue. These additional clusters found in the cancerous area could be representative of different sub-classes of malignancy not normally recognised via histology. In contrast, the extra clusters found in the normal area could be descriptive of normal tissue that is beginning to take on cancerous characteristics. However, the identification of such a large number of tissue subtypes is highly unlikely.

It is more plausible that the algorithm is too sensitive to the natural variation occurring within the tissue and thus creating an excessive amount of clusters that describe these differences. The false colour image shown in Figure 46f displays the final clustering result that was reached after the merge method step was applied. This clustering structure describes both the cancerous and normal tissue via individual clusters and was achieved for all of the initial clustering results.

In conclusion, the application of classical clustering techniques for accurate tissue spectra classification can prove insufficient. This is due in part to the overwhelming complexity of biological systems and also to the mechanics of the clustering techniques. Conventional unsupervised clustering techniques are limited by the fact they require the optimal cluster number to be specified in advance. Although the clustering process is an unsupervised procedure, the dictation of the final cluster number can prove to render the analysis less efficient, often producing large numbers

of misclassified spectra. A novel algorithm has therefore been developed that clusters spectral datasets by scrutinising both the potential final clustering structures and also the spectral characteristics of the clusters themselves. Results indicate the successful classification of large tissue spectral datasets that were previously incorrectly classified via traditional clustering methods.

2.4 Conclusions

In this chapter we have used FTIR imaging to study lymph node tissue sections. To summarise the results I have:

- Demonstrated that frozen sectioning of lymph node tissue specimens does not adversely affect the sample characteristics. This preparation method negates standard procedures more commonly employed that include paraffin embedment and subsequent de-paraffinization.
- Mounted samples upon BaF_2 substrates that enable transmission spectra to be collected from the sample. These were free from contaminating dispersion artefacts that are often observed using transflection sampling methodologies. Subsequent multivariate analyses could therefore utilise the full spectral range of the data and classify spectra according to spectral features that were characteristic of the sample alone.
- Applied a variety of unsupervised multivariate analysis techniques to the collected spectral datasets. A comprehensive and detailed comparison between techniques for tissue discrimination was therefore achieved. When correlating the results to the known histology of the samples, FCM clustering achieved the best tissue characterisation.

- Collected spectral datasets from an array of different lymph nodes that describe a number of different pathological states and tissue types. The spectral characteristics that are descriptive for each tissue type have been reported in detail. Diseased or abnormal cells appear to exhibit subtle but distinct changes in their protein vibrations and nucleic acid features below 1400 cm^{-1}. It is therefore essential to collect spectra that are free from dispersion artefacts or correct for these contaminations, as such spectral differences (distortion of the amide I and II absorption bands) would be detrimental for accurate tissue discrimination.
- Reported the successful classification of a dataset comprising tissue spectra collected from a number of different lymph nodes. This would verify results from similar studies upon human tissues that report smaller spectral variations from patient-to-patient than those due to alternate tissue types and diagnoses.
- Charted the development of new FCM clustering algorithms that provide improved data analysis. By use of PCA, the dimensionality of the data can be reduced without a significant loss of information. Computational requirements and data analysis times are significantly reduced without a detrimental loss in sensitivity. A novel algorithm that automatically detects the optimal amount of clusters that best describes a dataset is also reported. Unlike previous automated algorithms, the proposed method utilises both the potential final clustering structures and the spectral characteristics of the data to define the final cluster number.

2.5 References

[1] S. H. Landis, T. Murray, S. Bolden, and P. A. Wingo, *CA-Cancer J. Clin.*, 1999, **49**, 8.

[2] J. M. Johnson, R. R. Dalton, S. M. Wester, J. Landercasper, and P. J. Lambert, *Arch. Surg.-Chicago,* 1999, **134**, 712.

[3] M. W. Kissin, G. Querci della Rovere, D. Easton, and G. Westbury, *Brit. J. Surg.,* 1986, **73**, 580.

[4] R. R. Turner, D. W. Ollila, D. L. Krasne, and A. E. Giuliano, *Ann. Surg.*, 1997, **226**, 271.

[5] Taken from Hologic breast cancer screening website, located at URL, http://www.hologic.com/lc/bhspsg.htm

[6] L. G. Luna, *Manual of Histologic Staining methods of the Armed Forces Institute of Pathology*, 1968, McGraw-Hill, New York.

[7] K. Contractor *et al., Eur. J. Surg. Oncol.*, 2002, **28**, 787.

[8] A. A. Salem *et al., Eur. J. Surg. Oncol.*, 2002, **28**, 789.

[9] D. A. Grabau, F. Rank and E. Friis, *Acta Path. Micro. Im. A.*, 2005, **113**, 7.

[10] L. Esserman and N. Weidner, *Cancer J. Sci. Am.*, 1997, **5**, 226.

[11] S. A. Gulec, J. Su, J. P. O'Leary, and A. Stolier, *Am. Surgeon*, 2001, **67**, 529.

[12] P. J. Van Diest, H. Torrenga, P. J. Borgstein, R. Pijpers, R. P. Bleichrodt, F. D. Rahusen, and S. Meijer, *Histopathology*, 1999, **35**, 14.

[13] U. Veronesi, S. Zurrid, G. Mazzarol, and G. Viale, *Ann. Surg. Oncol.*, 2002, **9**, 745.

[14] J. R. Mourant, J. Boyer, A. H. Hielscher, and I. J. Bigio, *Opt. Lett.*, 1996, **21**, 546.

[15] L. T. Perelman, V. Backman, M. Wallace, G. Zonios, R. Manoharan, A. Nusrat, S. Shields, M. Seiler, C. Lima, T. Hamano, I. Itzkan, J. Van Dam, J. M. Crawford, and M. S. Feld, *Phys. Rev. Lett.*, 1998, **80**, 627.

[16] J. R. Mourant, A. H. Hielscher, A. A. Eick, T. M. Johnson, and J. P. Fregen, *Cancer Cytopathol.*, 1998, **84**, 366.

[17] K. S. Johnson, D. W. Chicken, D. C. O. Pickard, A. C. Lee, G. Briggs, M. Falzon, I. J. Bigio, M. R. Keshtgar, and S. G. Bown, *J. Biomed. Opt.*, 2004, **9**, 1122.

[18] G. M. Briggs, A. C. Lee, D. C. O. Pickard, J. R. Sainsbury, M. Falzon, I. J. Bigio, P. J. Ell, S. G. Bown, and M. R. S. Keshtgar, *Eur. J. Surg. Oncol.*, 2002, **28**, 771.

[19] A. C. Lee, D. C. O. Pickard, G. Briggs, J. R. Sainsbury, M. Falzon, G. Kocjan, I. J. Bigio, P. J. Ell, S. G. Bown, and M. R. Keshtgar, *Brit. J. Surg.* 2002, **89**, 74.

[20] C. J. Frank, R. L. McCreery, D. C. B. Redd, and T. S. Gansler, *Appl. Spectrosc.*, 1993, **47**, 387.

[21] K. E. Shafter-Peltier, A. S. Haka, M. Fitzmaurice, J. Crone, J. Myles, R. R. Dasari, and M. S. Feld, *J. Raman Spectrosc.*, 2002, **33**, 552.

[22] N. Stone, C. Kendal, N. Shepard, P. Crow, and H. Barr, *J. Raman Spectrosc.*, 2002, **33**, 564.

[23] J. Smith, C. Kendall, A. Sammon, J. Christie-Brown, and N. Stone, *Technology in Cancer Research & Treatment*, 2003, **2**, 327.

[24] M. J. Romeo and M. Diem, *Vib. Spectrosc.*, 2005, **38**, 115.

[25] M. Diem, M. J. Romeo, S. Boydston-White and C. Matthaus, *IR spectroscopic imaging: From cells to tissue, Spectrochemical analysis using infrared multichannel detectors,* 2005 Blackwell Publishing, Oxford.

[26] A. Stevens and J. Lowe, *Histology*, 1992, Glower Medical Publishing, London.

[27] E. Malinowski, *Anal. Chem.*, 1977, **49**, 612.

[28] E. R. Malinowski, *J. Chemometr.*, 1987, **1**, 33.

[29] H. H. Mantsch and M. Jackson, *J. Mol. Struct.*, 1995, **347**, 187.

[30] M. Jackson and H. H. Mantsch, *Biomedical Applications of Spectroscopy*, 1996, Wiley, New York.

[31] M. Jackson and H. H. Mantsch, *Infrared Spectroscopy: Ex Vivi Tissue Analysis, Encyclopedia of Analytical Chemistry*, 2000, **1**, Wiley, Chichester.

[32] P. Lasch, W. Haensch, D. Naumann and M. Diem, *Biochim. Biophys. Acta,* 2004, **1688**, 176.

[33] P. Lasch, A. Pascifico and M. Diem, *Biopolym. Biospectrosc.*, 2002, **67**, 335.

[34] M. Diem, M. Romeo, C. Matthaus, M. Miljkovic, L. Miller and P. Lasch, *Infrared Phys. Techn.*, 2004, **45(5-6)**, 331.

[35] B. R. Wood, M. A. Quinn, F. R. Burden and D. McNaughton, *Biospectrosc.*, 1996, **2**, 143.

[36] B. R. Wood, L. Chiriboga, H. Yee, M. A. Quinn, D. McNaughton and M. Diem, *Gynecol. Oncol.*, 2004, **93**, 59.

[37] X. Y. Wang, G. Whitwell and J. Garibaldi, *The Application of a Simulated Annealing Fuzzy Clustering Algorithm for Cancer Diagnosis, in the Proceedings of the IEEE 4th International Conference on Intelligent System Design and Application,* Budapest, Hungary, 2004, 467.

[38] X. Y. Wang and J. Garibaldi, *Eur. J. Inform.*, 2005, **29**, 61.

[39] X. Y. Wang, J. M. Garibaldi, B. Bird and M. W. George, *Fuzzy Clustering in the Biochemical Analysis of Cancer Cells, in the Proceedings of the Fourth Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT)*, 2005, Barcelona, Spain, 1118.

[40] X. Y. Wang, J. M. Garibaldi, B. Bird and M. W. Geogre, *Appl. Intell.*, 2006 (In press).

[41] J. H. Ward, *J. Am. Stat. Assoc.*, 1963, **58**, 236.

[42] S. Boydston-White, T. Cherneuko, A. Regina, M. Miljkovic, C. Matthaus and M. Diem, *Vib. Spectrosc.*, 2005, **38(1-2)**, 169.

[43] C. Matthaus, S. Boydston-White, M. Miljkovic, M. Romeo and M. Diem, *Appl. Spectrosc.*, 2006, **60**, 1.

[44] R. Allibone *et al.*, *FTIR microscopy of oral and cervical tissue samples, Internal Report*, 2002, Derby City General Hospital.

[45] E. Garett-Mayer and G. Parmigiani, *John Hopkins University, Dept. of Biostatistics Working Press Papers,* John Hopkinds Univeraity, The Berkeley Electronic Press(bepress).

[46] A. K. Jain, M. N. Murty and P. J. Flynn, *ACM Comput. Surv.*, 1999 **31(3)**, 264.

[47] D. Jiang, C. Tang and A. Zhang, *IEEE T. Knowl. Data En.*, 2004, **16(11)**, 1370.

[48] D. R. Causton, *A Biologists Advanced Mathematics*, 1987, Allen & Unwin, London.

[49] I. T. Jolliffe, *Principal Component Analysis,* 1986, Sprnger-Verlag, New York.

[50] J. R. Mansfield, L. M. McIntosh, A. N. Crowson, H. H. Mantsch and M. Jackson, *Anal. Chem.*, 1997, **69(16)**, 3370.

[51] L. M. McIntosh, J. R. Mansfield, N. A. Crowson and H. H. Mantsch, *Biospectrosc.*, 1999, **5**, 265.

[52] M. R. Rezaee, B. P. F. Lelieveldt and J. H. C. ReiBer, *Pattern Recogn. Lett.*, 1998, **19**, 237.

[53] J. Beztek, *Pattern Recognition in Handbook of Fuzzy Computation,* 1998, IOP Publishing , Boston.

[54] X. L. Xie and G. Beni, *IEEE T. Pattern Anal.*, 1991, **13(8)**, 841.

[55] D. W. Kim, K. H. Lee and D. Lee, *Pattern Recogn.*, 2004, **37**, 2009.

[56] N. R. Pal and J. Beztek, *IEEE T. Fuzzy Syst.*, 1995, **3**, 370.

[57] S. Bandyopadhyay, *Simulated Annealing for Fuzzy Clustering: Variable Representation, Evolution of the Number of Clusters and Remote Sensing Applications,* 2003, unpublished, private communication.

Chapter 3

Cervical Cancer

3.1 Introduction

Until the early 1990's, cervical cancer was the most common malignancy found among women in developing countries [1]. At present, it is estimated that 493,000 new cases of cervical cancer are diagnosed worldwide each year [2]. In England alone, over 2800 new cases of invasive carcinoma and 19,000 cases of carcinoma *in situ* are reported each year [3], leading to a death rate of c.a. 1100 p.a. [4]. However, these numbers would be dramatically higher without the present National Health Service Screening Program (NHSCSP) [5]. Currently, screening for cervical disease is performed via the visual analysis of exfoliated cervical cells by a histopathologist (PAP test). Squamous and columnar epithelial cells are collected using an Ayre spatula or Cytobrush™ from the transformation zone of the cervix, fixed in ethanol, and stained using the Papanicolaou stain [6]. The stain colours their nuclei and cytoplasm different colours, making it possible to differentiate between healthy and abnormal cells using their relative nucleus to cytoplasm ratio. A more definitive diagnosis can be made by the examination of biopsy material, whereby samples are cut into thin sections and stained. This type of screening can provide a higher predictive value than that of the PAP test because the anatomical arrangement of the tissue is maintained. Therefore, evaluation of morphological changes occurring within the cells can be directly related to the histological architecture of the tissue. By use of the PAP test, abnormal cervical smears are classified using a two tier system, graded as either low or high grade squamous intraepithelial lesions (LSIL

and HSIL). Surgical samples are alternatively classified by a three tier system, graded as mild, moderate or severe cervical intraepithelial neoplasia (CIN I, II and III). The latter two stages of neoplasia display a high risk of developing into carcinoma *in situ* (CIS) and warrant the removal of abnormal tissue via diathermy or laser ablation. Recent studies have also indicated that the presence of human papilloma virus (HPV) can be associated with cervical dysplasia and its progression to malignancy [7,8]. These viral changes to cervical cells are now utilised as a preceding factor for the detection of cervical lesions within PAP smears [9,10].

Despite the success of cervical screening programs, the PAP test has limitations. The visual analysis and grading of smears employs human judgement that can be somewhat subjective. Research undertaken to assess the efficiency of this procedure revealed that 53% of patients with invasive carcinoma had previously attended a smear test which failed to identify abnormal changes [11]. False negative diagnosis rates for the PAP test have been reported to be as low as 1% and as high as 93% [12-14]. It has been suggested that several sensible factors may contribute to insufficient diagnoses [15]. These include the presence of contamination or inflammation that would mask diagnostic cells, poor sampling of the correct region in the cervix and poorly controlled technical processes. With the unfortunate recent decline in recruitment of qualified pathologists and cyto-technicians, there is a strain to complete an ever increasing and demanding workload, with a reported 4.4 million PAP smears analysed every year in the UK [16]. Taking into account all these factors, the incidence of reported misdiagnoses are understandable considering the procedure is so heavily dependent upon correct human judgement. In the modern

NHS there is a need for a less operator dependent and more accurate automated analysis of cervical smears.

A variety of different techniques have been examined that aim to eliminate the subjective diagnoses currently made during the cytological screening process. A popular approach was the use of automated image analysis systems coupled with artificial neural networks (ANNs) [17,18]. This technique has recently fallen out of favour and been replaced by liquid based cytology (LBC) methods that aim to improve the quality of cervical smear presentation upon slides [19,20]. Although such LBC techniques improve the cellular presentation, removing unwanted contaminates such as inflammatory or blood cells, the slides created are still diagnosed via PAP staining and assessed by subjective cytological screening. An alternative technique that shows distinct potential is the application of FTIR spectroscopy, particularly FTIR imaging. Over the past decade, the application of FTIR spectroscopy to disease diagnosis has received a large amount of attention since this method is sensitive to biochemical changes that occur within cells and could thus identify differences that accompany and precede the onset of disease [21-27]. However, research to date has emphasised the high degree of heterogeneity found within mammalian cells, cervical tissues proving to be one of the most complex. The first FTIR spectroscopic studies of cervical cells were undertaken by Wong *et al.* [28,29] who collected macroscopic spectra from exfoliated cell pellets. The authors reported a decrease in the intensity of glycogen bands and an increase in the intensity of symmetric (PO_2^-) and anti-symmetric (PO_2^-) bands associated with nucleic acids for dysplastic and cancerous samples. However, further studies undertaken by Cohenford *et al* [30 – 32], McNaugthton *et al* [33 – 36] and Diem *et*

al [37 – 44] indicated that these spectral changes observed by Wong may not be related to the number or molecular composition of dysplastic cells, but to confounding contributions made by different cell types present within a smear. Benign variations such as inflammation, the ratio of non-dividing to dividing cells, and the overall divisional activity of the cells can dramatically change the IR spectrum collected. As these problems were recognised, it also became apparent that other contaminates may effect the spectra, including blood, mucus, micro-organisms and semen. It was estimated by Chriboga *et al* [38] that around 30% of the samples they examined were contaminated by blood or mucus making correct interpretation of such spectra impossible. Liquid based techniques (LBC) of sample preparation have more recently been utilised to minimise spectral contributions from mucins and erythrocytes to cell pellet spectra [36]. Leukocytes can also be removed by the addition of white cell lysis buffer, but care must be taken not to damage diagnostic cells [45]. Glycogen levels may also change throughout the menstrual cycle, where it maximises at ovulation, and is known to decrease dramatically after menopause [46].

In conclusion, the collection of macroscopic spectra from cervical smear pellets can lead to spectral features that are not directly correlated to disease change. A detailed understanding is therefore required about each cell type present within a smear and also the natural variation they exhibit due to differentiation, maturation and stage within the mentrual cycle. In this study we have examined both cervical tissue sections and individual exfoliated cells to help interpret and assess the spectral variations that may occur within exfoliated smear material.

3.2 Cervix Histology

3.2.1 Basic Structure

As shown in the schematic of the female reproductive system (Figure 1a), the cervix is located at the bottom of the vaginal canal and forms the lower part of the uterus. A photomicrograph displaying the three main parts of the uterus is shown in Figure 1b. Both the fundus and body parts of the uterus have similar histological structures and are both lined with columnar epithelium. In contrast, the cervix displays two main regions where stromal tissue is lined by both columnar and squamous epithelium separately. These two regions are more commonly termed the endocervix and ectocervix respectively, and display a distinctly different histological architecture. As illustrated in the photomicrograph, the cervix is both cylindrical and symmetrical in shape, being approximately 3cm in length and 2cm in diameter. Small changes in these dimensions can occur after pregnancy whereby the endocervical canal can become more barrel shaped. Cervical stroma is primarily composed of fibrocollagenous tissue that incorporates some smooth muscle. The proportion of each tissue type can vary according to maturation. Both blood and lymphatic vessels are often prominent and numerous.

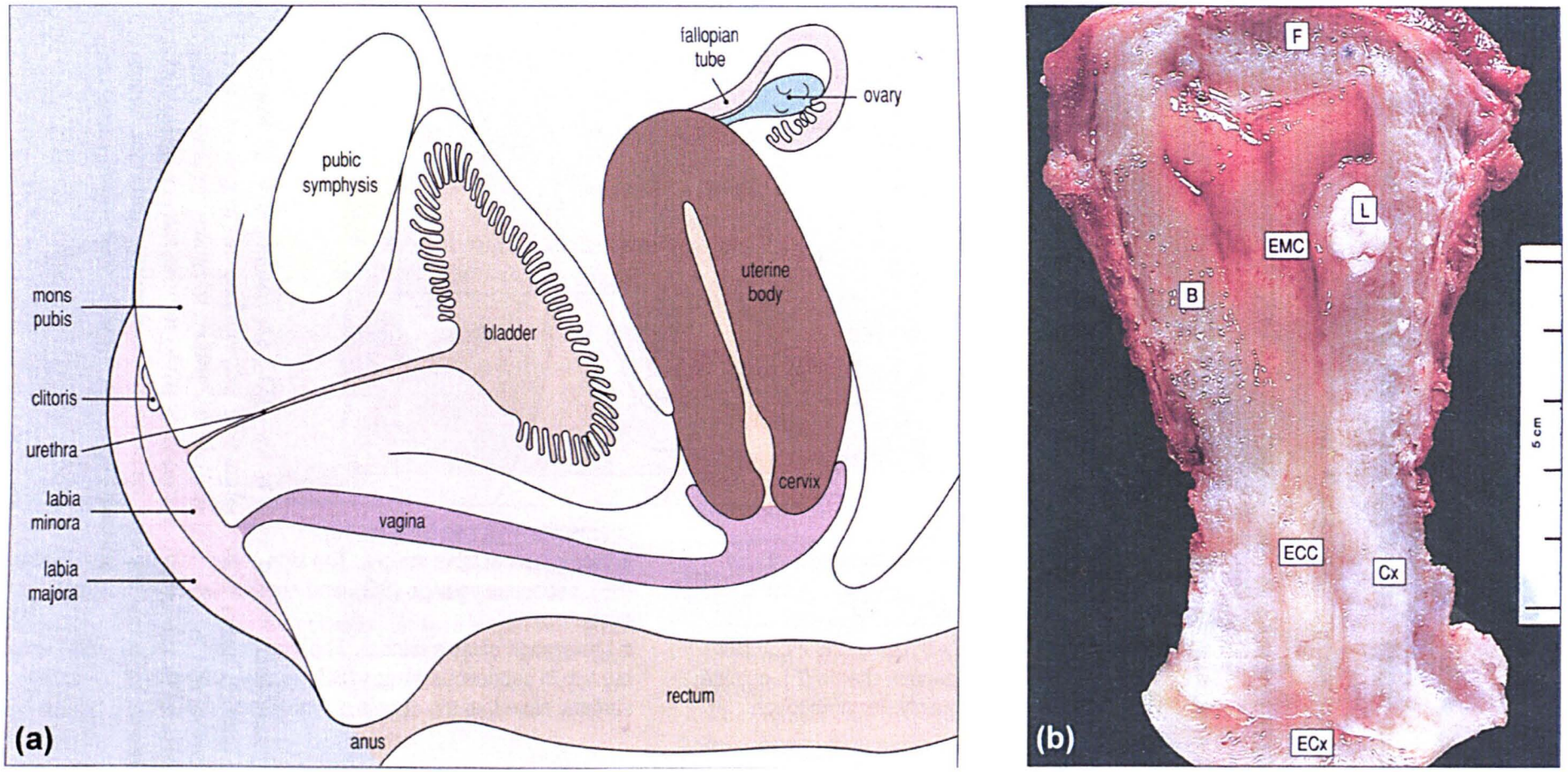

Figure *1:* *The female reproductive system [47]. a) A diagram displaying a lateral view of female genitalia. The cervix is located at the end of the vaginal canal and provides the opening into the uterine body. b) A photomicrograph of a section through the uterus. This clearly shows the fundus (f), body (B) and cervix (Cx) parts of the uterus. Note the endometrial cavity (EMC), endocervical canal (ECC) and ectocervix (ECx). The smooth muscle of the body contains a small tumour, a leiomyoma (L).*

3.2.2 Endocervix

The endocervix is lined with a single layer of tall columnar epithelium that allows the secretion of mucus into the endocervical canal. Tissue sections cut both along and across the canal have indicated the existence of large invaginations. These extend from clefts deep within the stroma that comprise mucus glands. Tubules rise from this region to the surface epithelium and provide a large surface area for mucus production and secretion. Stained photomicrographs displaying these histological features are shown in Figures 2a – b.

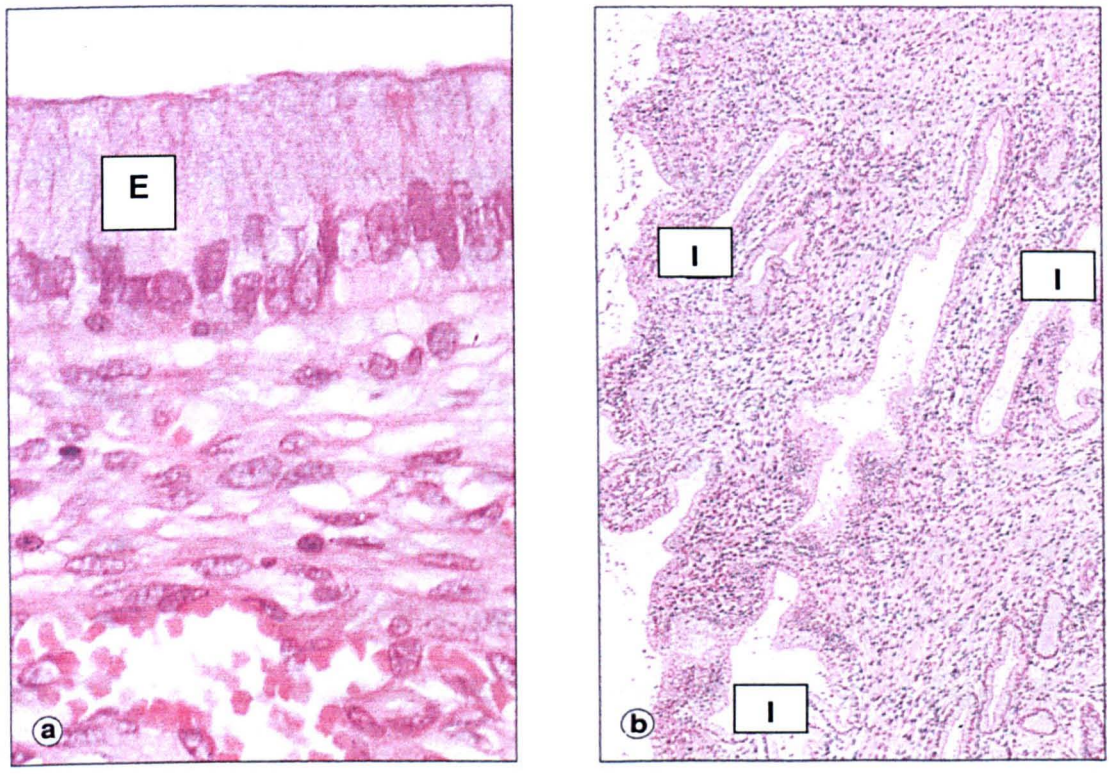

Figure 2: *The endocervix [47]. a) A high power photomicrograph displaying a cross section of the endocervix. The endocervical canal is lined by a single layer of tall columnar mucus-secreting epithelium (E). b) A low power photomicrograph displaying a cross section of the endocervix. Numerous deep invaginations (I) of the mucus secreting epithelium extend deep into the cervical stroma and thus greatly increases the surface area for mucus production.*

The movement of mucin into the endocervical canal is facilitated by ciliated columnar epithelial cells that are scattered across the mucus-secreting endocervical cells. This transport generally occurs at the upper end of the canal close to the endometrium junction. The chemical properties of cervical mucin are subject to marked changes during the menstral cycle providing properties that either facilitate or deter the movement of spermatozoa. However, the endocervical columnar epithelium displays little microscopic variation.

3.2.3 Ectocervix

The region at which the cervix protrudes into the vaginal cavity is more commonly termed the ectocervix, and is shown in Figure 3a. This part of the cervix is lined with non-keratinising, stratified, squamous epithelium, similar to that of the vagina. The structure of this epithelium varies with age and hormonal activity. Both before puberty and after menopause the epithelium is thin. However, during the reproductive years the epithelium thickens due to the release of oestrogens. An increased mitotic activity is observed for cells that exist within the basal and parabasal layers of the epithelium. The superficial layers of cells display a marked increase in both population and size. This histological change characterises the accumulation of stored glycogen and lipids within the cytoplasm of these cells. A stained photomicrograph illustrating these histological features is shown in Figure 3b. During times of ovulation, the glycogen content of these cells is maximal with some glycogen rich surface cells being shed into the vaginal cavity at the end of the ovulation period. These cells can then be broken down by commensal lactobacilli, producing lactic acid and restricting bacterial invasion via an acid pH.

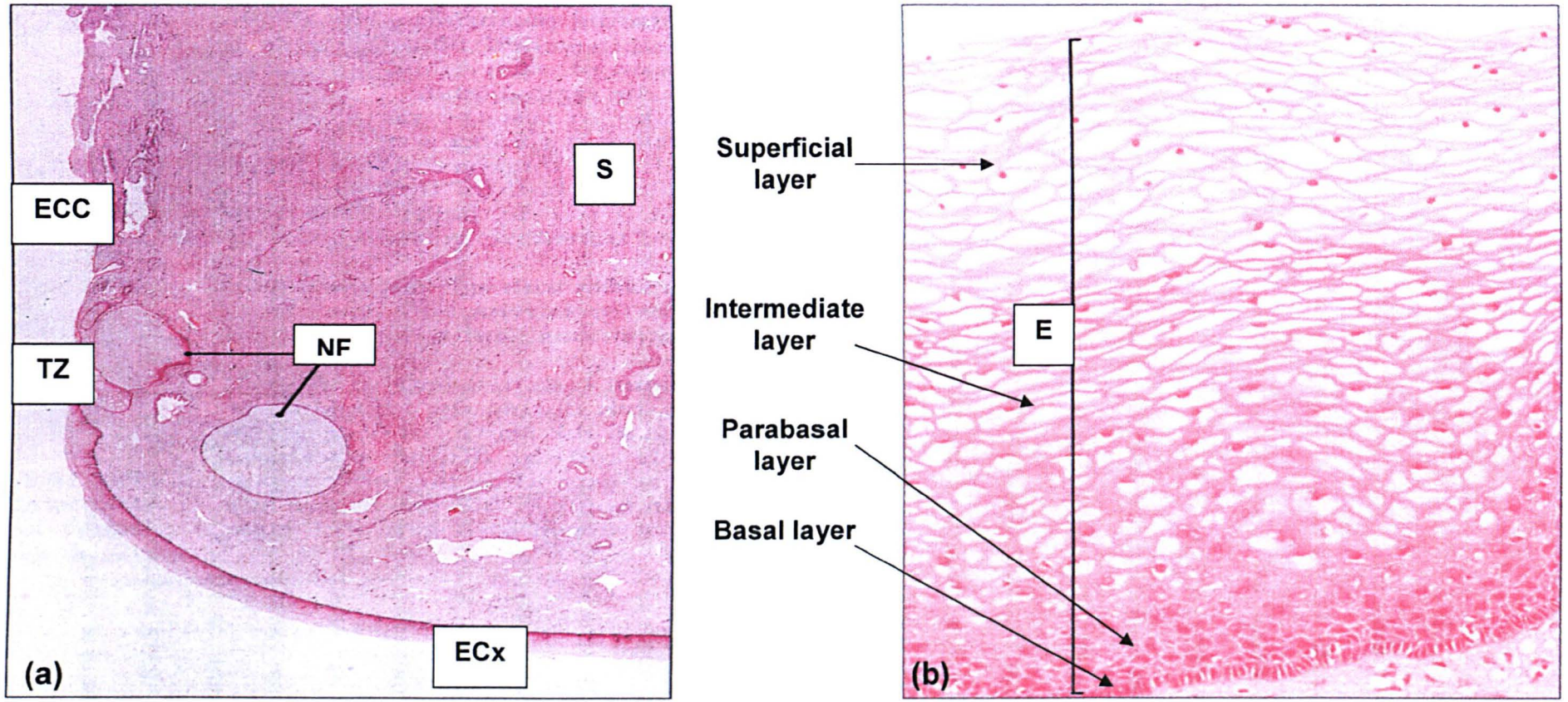

Figure 3: *The ectocervix. [47] a) Low power photomicrograph of the cervix. The stroma (S) is composed of smooth muscle fibres embedded in collagen. The proportions of muscle and fibrous tissue vary according to age. The ectocervix (ECx) is covered with stratified squamous epithelium, while the endocervical canal is lined by tall columnar epithelium. The junction between the squamous and columnar epithelium is located in the region of the external os. In this example, there is a transformation zone (TZ) of squamous epithelium which has extended into the endocervical canal. Also note the Nabothian foliciles (NF) produced after squamous metaplasia. b) A high power photomicrograph showing the squamous epithelium of the ectocervix (E).*

3.2.4 Squamo – columnar Junction of Cervix

The point at which columnar epithelial cells of the endocervix meet squamous epithelial cells of the ectocervix is known as the squamo-columnar junction. This zone is the site of many important pathological changes that accompany the progression of age and the onset of disease. The location of the squamo-columnar junction is initially found at the ectocervix's opening called the external os, as shown in Figure 4a. During puberty the columnar epithelium extends into the ectocervix forming an ectropion or cervical erosion (Figure 4b). This change is again governed by hormonal activity and can be markedly increased by a first pregnancy. The breakdown of glycogen contained within superficial cells of the vaginal and cervical squamous epithelium creates an acidic pH within this region. As a consequence squamous metaplasia is induced, and thus creates a transformation zone between the endocervical columnar epithelium and ectocervical squamous epithelium (Figure 4c). The transformation zone now comprises new squamous epithelium in an area previously dominated by columnar epithelium. The size of this zone is thus dictated by the original ectropian that invaded into the ectocervix. However, in older women the transformation zone often retreats back into the endocervical canal. Another invariable consequence of squmaous metaplasia is the obliteration of invaginations located close to the external os. Mucin now becomes trapped within the clefts and forms spherical cystic masses of inspissated mucus that are lined with flattened endocervical mucus-secreting epithelium. These cystic masses are more commonly termed Nabothian follicles and are shown in Figure 4c. Further consequences of this constant change of epithelium type and junction position can also include the development of abnormal epithelium that may progress to cancer.

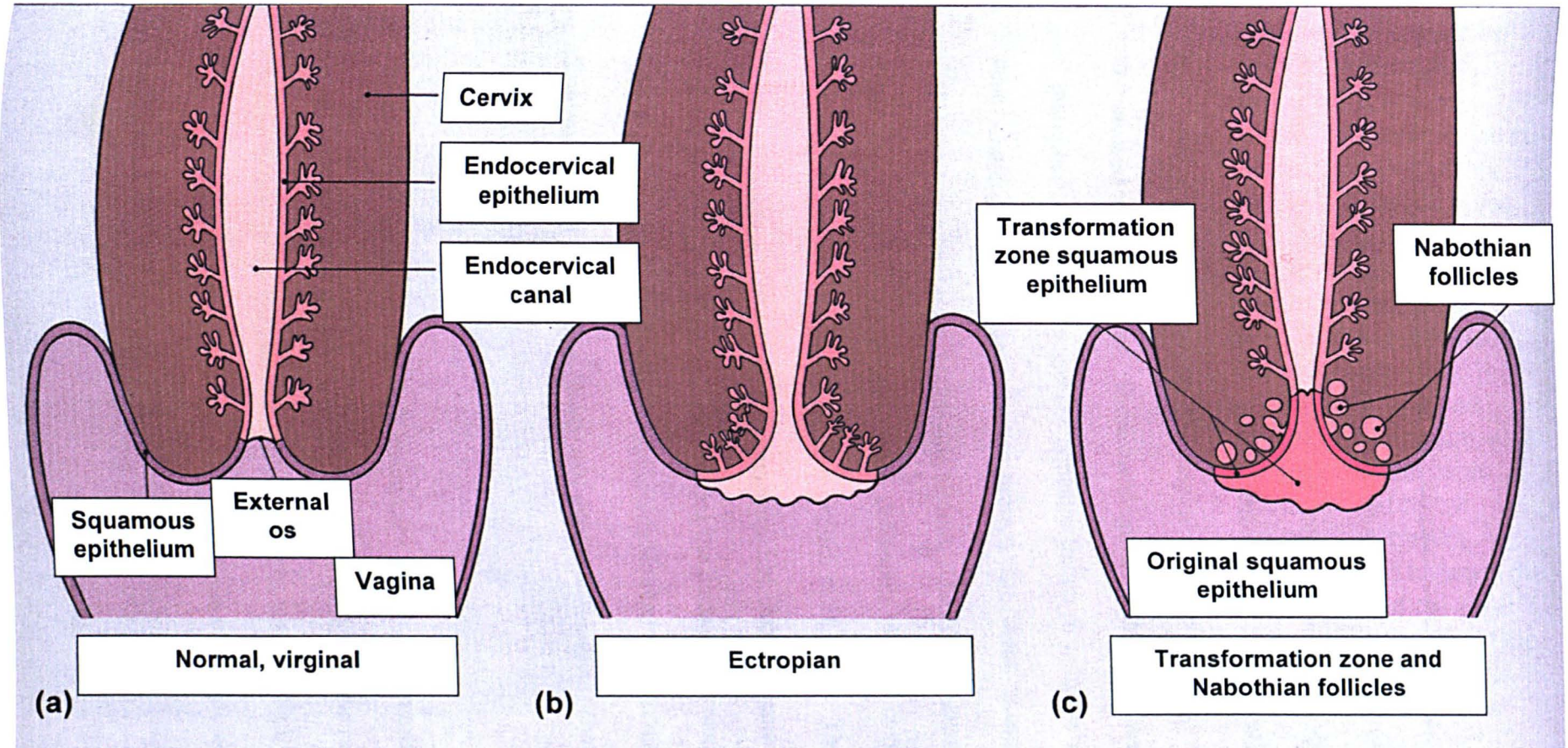

Figure 4: *The Squamo-columnar junction of the cervix. [47] a) The squamo-columnar junction is originally situated in the region of the external os before puberty. b) At puberty the endocervical epithelium extends distally into the acid environment of the vagina and forms an ectropian. c) A transformation zone forms as squamous epithelium regrows over the ectropian. The openings of the crypts may be obliterated in the process, and result in the formation of mucus filled Nabothian follicles.*

3.2.5 Carcinoma of the Cervix

The most common origin site of cervical carcinoma is found within the transformation zone of the cervical epithelium. An invasive carcinoma is usually preceded by histological abnormality that occurs within the squamous epithelial cells of this zone. The abnormal cells show a loss in their regular stratified pattern, have a high nucleus to cytoplasm ratio, and display a variation in their shape and size with an increased mitotic activity. These histological features are classical characteristics of malignant tumour cells and are typically linked to invasive activity. However, these histological changes may be present for many years before abnormal epithelium begins to invade underlying stroma. The progression of abnormal change across the squamous epithelium, more commonly termed cervical intraepithelial neoplasia (C.I.N), can be characterised into three main stages. These are named mild (C.I.N. I), moderate (C.I.N II) and severe (C.I.N. III) neoplasia respectively. It is during these stages of abnormal change that the disease is termed carcinoma-in-situ, having yet to penetrate the barrier between epithelial and stromal cervix tissue. A photomicrograph displaying these malignant histological changes is shown in Figures 5a – b.

The abnormal epithelial cells will eventually breach the basement membrane and invade into the cervical stroma, as shown in Figure 6. At this point malignant cells can gain access to the lymphatic and blood vessels, enabling their passage around the body and formation of multiple tumours. This stage of disease is known as invasive carcinoma.

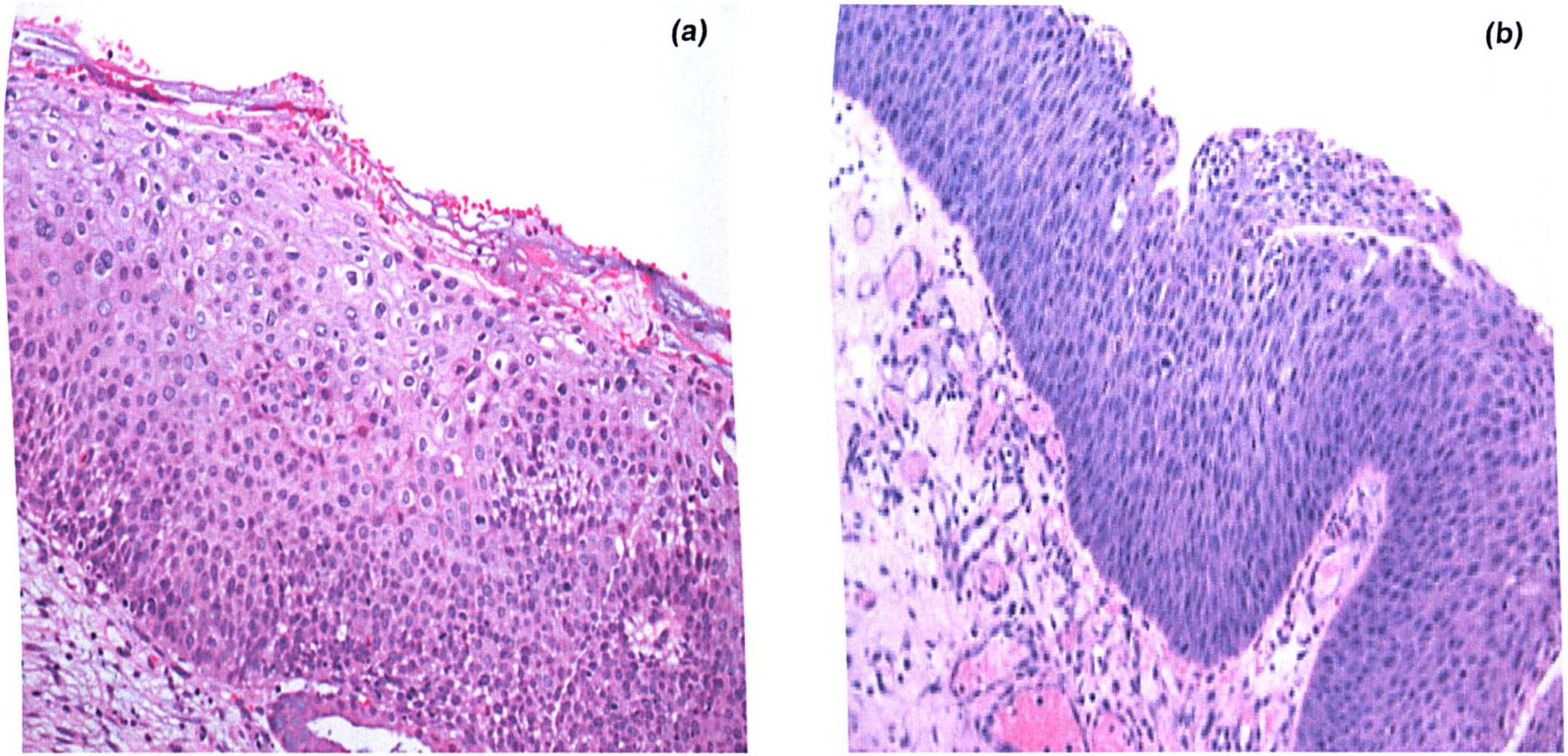

Figure 5: *Cervical intraepithelial neoplasia. [47] a) A high power photomicrograph displaying an area of moderate cervical neoplasia. Note the disorderly development of cells from the basement membrane that exhibit large nuclei. This histological change in the squamous epithelium would normally be termed C.I.N II. b) A high power photomicrograph of severe cervical neoplasia or C.I.N. III. Note the high nucleus to cytoplasm ratio of these cells that have now fully infiltrated the squamous epithelium.*

Diagnosis of the early stages of disease (C.I.N) can be achieved via cervical cytology. Superficial squamous epithelial cells are scraped from the transformation zone in the region of the external os. These exfoliated cells are then prepared onto slides and examined microscopically for the appearance of abnormal cells. An example microscopic area taken from such a slide is shown in Figure 7. Cytological features of dysplasia can be noted for the cells that exist toward the centre, displaying an increased nuclear to cytoplasmic ratio with darker and more irregular nuclei. Healthy squamous cells alternatively display small nuclei and large amounts of cytoplasm. The identification of such abnormalities can then allow the rapid treatment of these neoplastic regions upon the squamous epithelium. These commonly include surgical removal via diathermy or laser ablation, thus preventing future development of invasive carcinoma.

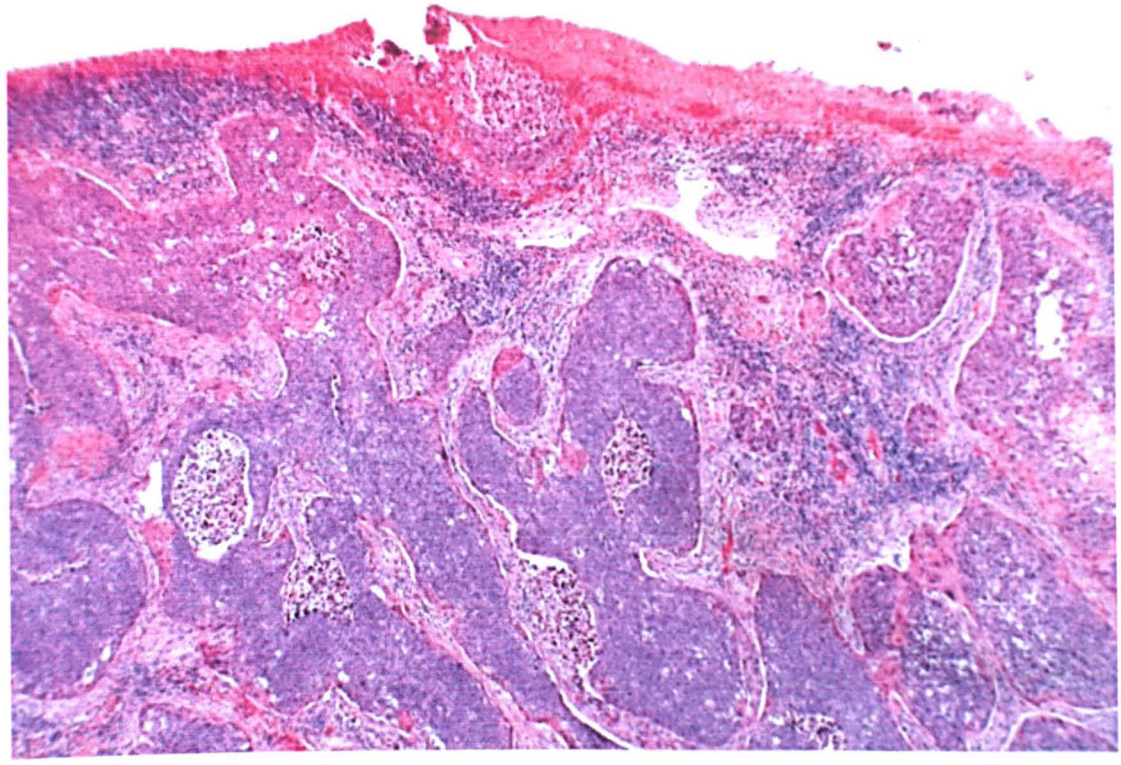

Figure 6: *A low power photomicrograph displaying an invasive carcinoma of the cervix [47]. The abnormal cells have now breached the basement membrane and are forming tumours within the stroma of the cervix.*

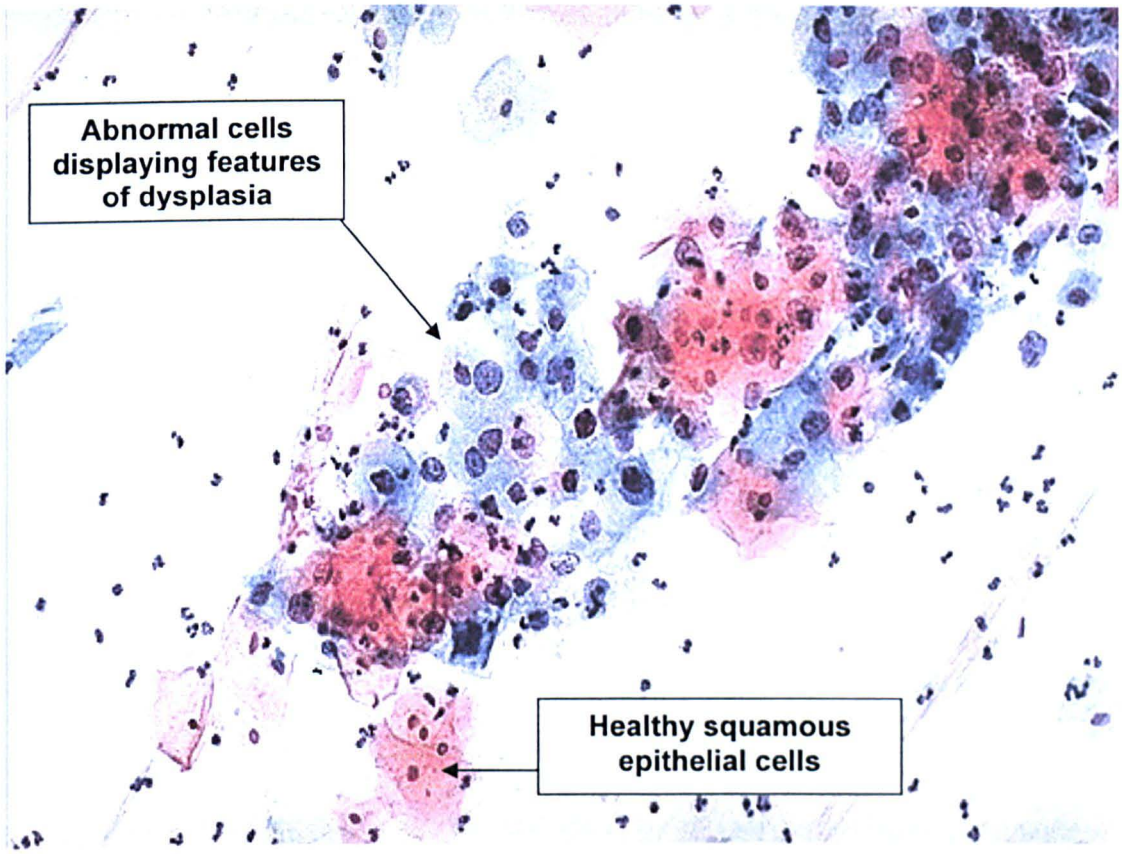

Figure 7: *A high power photomicrograph taken from a small region upon a traditionally prepared Pap smear slide. Note the cells at the centre of the image displaying features of dysplasia. These include an increased nuclear to cytoplasm ratio, with darkened more irregular shaped nuclei. In contrast healthy squamous epithelial cells display small nuclei and large cytoplasms. From ref [48].*

3.3 Results

In this work we have the following objectives:

(i) Assess the feasibility of using vibrational spectroscopy for accurate disease diagnosis of the cervix.

(ii) Compare and contrast the ability of unsupervised multivariate analysis techniques to discriminate different cervical tissue types, whether they are diseased or healthy in nature.

(iii) Establish spectral characteristics that are descriptive for each tissue type and seek features that could be used for future supervised pattern recognition.

In order to demonstrate this I will:

(i) Present results obtained from FTIR microscopic mapping of cervical tissue sections that incorporate the transformation zone. This will include a comparison of the multivariate techniques used to scrutinise the IR micro-spectral datasets produced.

(ii) Display multivariate IR imaging results from a selection of different cervical tissue sections.

(iii) Compare IR spectra collected from individual healthy squamous epithelial cervical cells by use of a synchrotron source.

(iv) Describe novel experiments whereby FTIR microscopic maps have been collected from exfoliated cervical cells by use of a conventional source.

3.3.1 Evaluation of Cervical Tissue Sections using IR Multivariate Imaging

The overall goal of this research is to develop protocols by which exfoliated cervical cells can be scrutinised and diagnosed via FTIR microspectroscopy. In order to gain a detailed insight into the various cell types and IR signatures produced by exfoliates, it was deemed necessary to fully understand the origin of the major spectral types that can be observed. Therefore, we examined tissue sections cut from cervical

biopsy material that incorporated the transformation zone via FTIR microscopic mapping. This zone is the site of many important pathological changes that accompany the progression of age and the onset of disease. A rich understanding of the cell types present in this region can thus aid FTIR spectroscopic interpretation of cervical exfoliates. A variety of different unsupervised multivariate analysis techniques were applied to the IR micro-spectral datasets produced. These include PCA, MCR and a novel PCA-FCM Clustering algorithm. The ability of each technique to discriminate alternative tissue types was assessed via direct comparison to conventional histopathology.

3.3.1.1 FTIR Multivariate Imaging of Healthy Cervical Tissue Sections

Ectocervix

The initial aim of this study was to assess the natural variation in biochemistry that may occur within the cells of healthy cervix tissue. Biopsy material was therefore collected from a healthy patient that exhibited no previous abnormal cervical smears (case C771602). The tissue section that was cut for analysis is shown in Figure 8a (white light image) and clearly illustrates the transformation zone whereby squmaous epithelium meets columnar epithelium. The first region chosen for analysis is displayed in Figures 8b-c respectively, and describes the ectocervix. Unfortunately a parallel H&E stained section was not made available for this sample, but the main types of tissue can be visualised via contrast in light intensity of the tissue regions. These include the underlying stromal or connective tissue and the basal, parabasal, intermediate and superficial layers of the squamous epithelium (Figure 8c). Using a pixel size 6.25μm a total of 4264 individual IR spectra were collected from an area

of 650 x 256.25 μm. The multivariate imaging results produced for this dataset are shown in Figure 9. Each method applied has been allocated an individual panel and only displays imaging results that produce meaningful information about the tissue section and the technique that was used.

The first panel displays the PCA imaging results. It can be seen that over 97% of the total variance contained within the dataset is comprised within the first two PC's of the analysis. When studying the colour weighted image for the first PC in Figure (b), we can see that this PC clearly gives contrast between the stromal tissue (blue pigmentation), parabasal layer (cyan pigmentation) and superficial layer (yellow pigmentation) of the squamous epithelium. In addition, the outside region that contains no tissue is clearly marked with a red colouration, and a very small amount of contrast can be visualised between the underlying stroma (dark blue pigmentation) and basal layer of the epithlieum (light blue pigmentation). The second PC image shown in Figure (c) provides contrast upon the mapped area that reveals three regions. The first region, highlighted by a cyan colour is descriptive of the underlying stroma and basal layer of the squamous epithelium. In contrast, the second region seen with a dark blue colour describes the parabasal and superficial layers of the epithelium. The final region shown with a red colour again highlights the area where no tissue exists. Studying the third PC image in figure (d), the stroma and superficial layers of the epithelium both display strong correlations with this component and are highlighted by a red colouration, which provides contrast to the basal and parabasal layers of the epithelium (blue and cyan colours respectively). All subsequent PC images provide little information about the tissue but highlight the

area that contains no tissue or pixels that are likely to have been half on and off the tissue.

The MCR panel displays the resulting images constructed from both a 3 and 4 component analysis of the same dataset (images a – g). By comparison to the known histological tissue types in this region, the 4 component system gave the best characterisation of the tissue section. The first component in the analysis (image d), is representative of the intermediate and superficial layers of the squamous epithelium. The second component (image e) appears to be descriptive of the underlying stromal tissue. Examining the third component (image f), this clearly highlights the area whereby no tissue exists found at the right hand side of the image. Studying the fourth and final component (image f), this highlights the basal layer of the squamous epithelium and provides a small amount of contrast for the directly adjacent parabasal layer (yellow pigmentation).

The final panel displays images created via PCA-FCM Clustering. Images (a) to (f) were constructed by subjectively increasing the amount of clusters found by the analysis from 2 – 7 respectively. When comparing these created images directly against the known tissue type regions, the image constructed from a 7 cluster analysis appears to best mimic the histological architecture of the tissue section (image f). The squamous epithelium is now characterised by individual clusters that describe the basal (red), parabasal (dark blue), intermediate (royal blue) and superficial (cyan) layers that illustrate the maturation of these epithelial cells. In contrast, the yellow cluster of spectra highlights the underlying stromal tissue. The final green and brown clusters describe areas where there is no tissue or pixels that

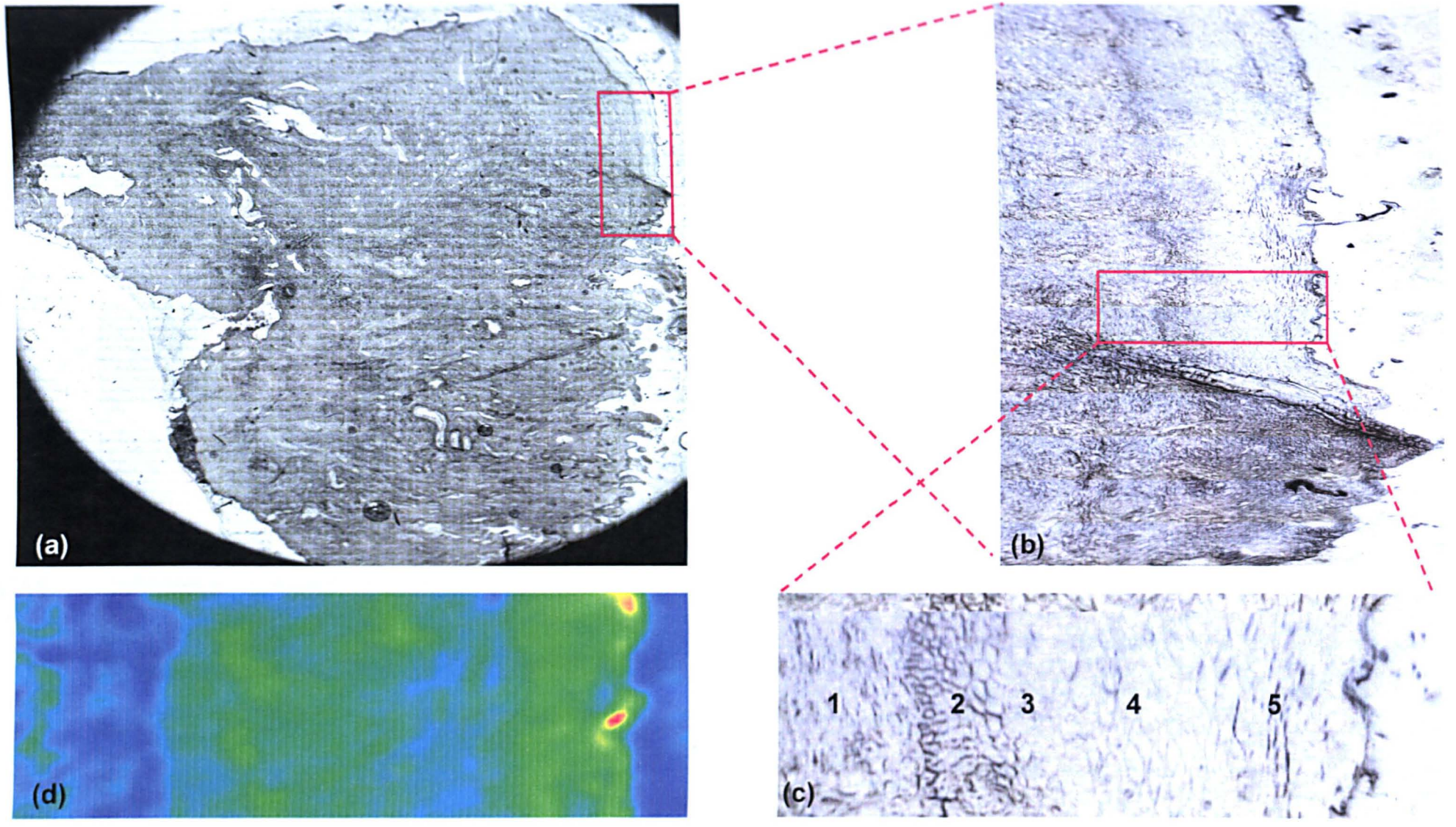

Figure 8: *a) White light image of entire cervical tissue section. b) White light image of transformation zone. c) Magnified region displaying benign anatomical features. (1) Underlying connective or stromal tissue, (2) basal layer, (3) parabasal layer, (4) intermediate layer and (5) superficial layer of squamous epithelium. d) IR imaged area (650 x 256.25 μm) mapped using a pixel size of 6.25 μm for a total 4264 individual IR spectra.*

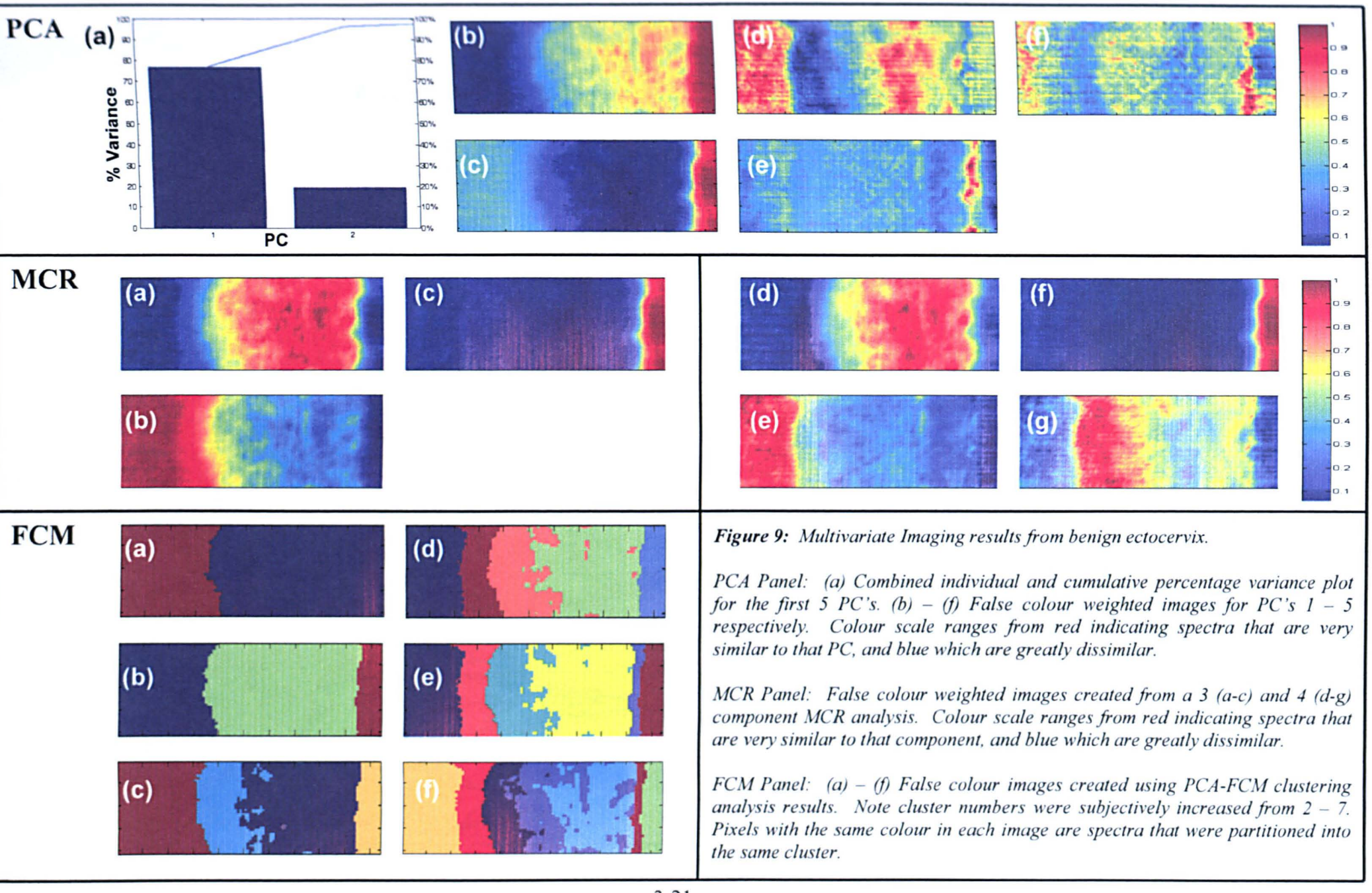

Figure 9: *Multivariate Imaging results from benign ectocervix.*

PCA Panel: (a) Combined individual and cumulative percentage variance plot for the first 5 PC's. (b) – (f) False colour weighted images for PC's 1 – 5 respectively. Colour scale ranges from red indicating spectra that are very similar to that PC, and blue which are greatly dissimilar.

MCR Panel: False colour weighted images created from a 3 (a-c) and 4 (d-g) component MCR analysis. Colour scale ranges from red indicating spectra that are very similar to that component, and blue which are greatly dissimilar.

FCM Panel: (a) – (f) False colour images created using PCA-FCM clustering analysis results. Note cluster numbers were subjectively increased from 2 – 7. Pixels with the same colour in each image are spectra that were partitioned into the same cluster.

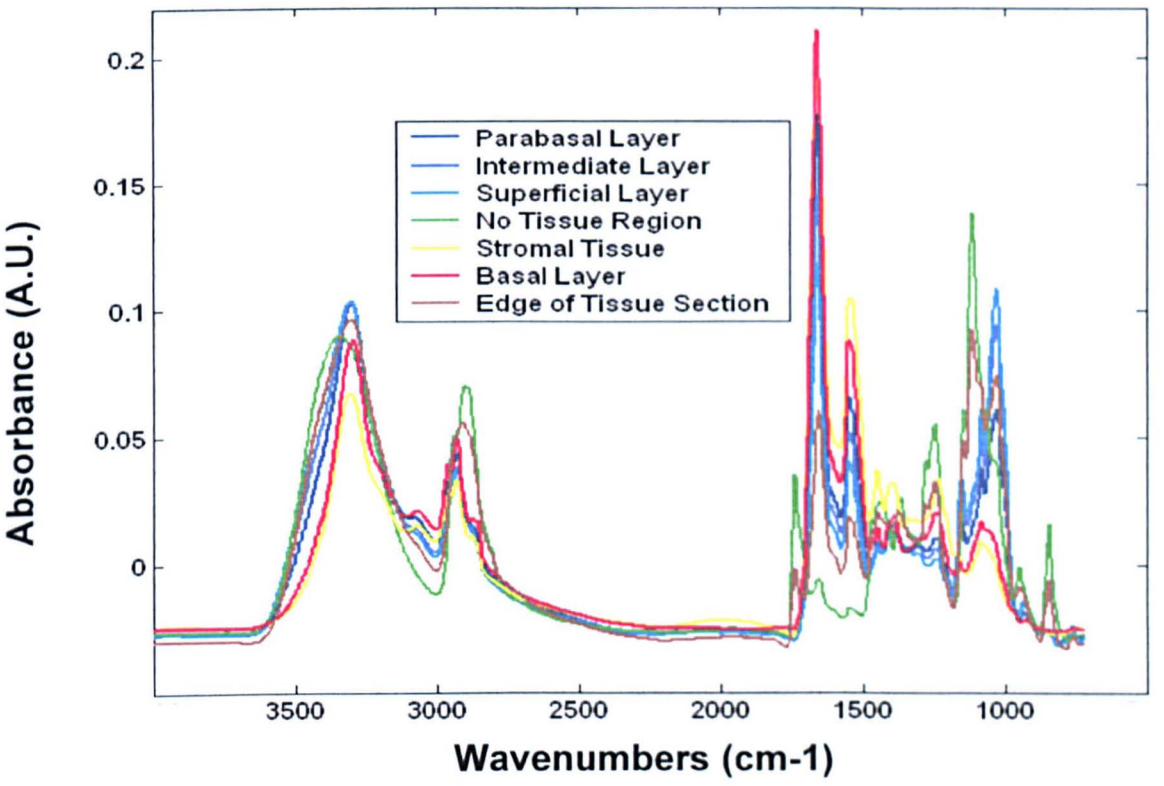

Figure 10: *7 Cluster PCA-FCM Analysis Results. Mean average spectra for each cluster in the analysis.*

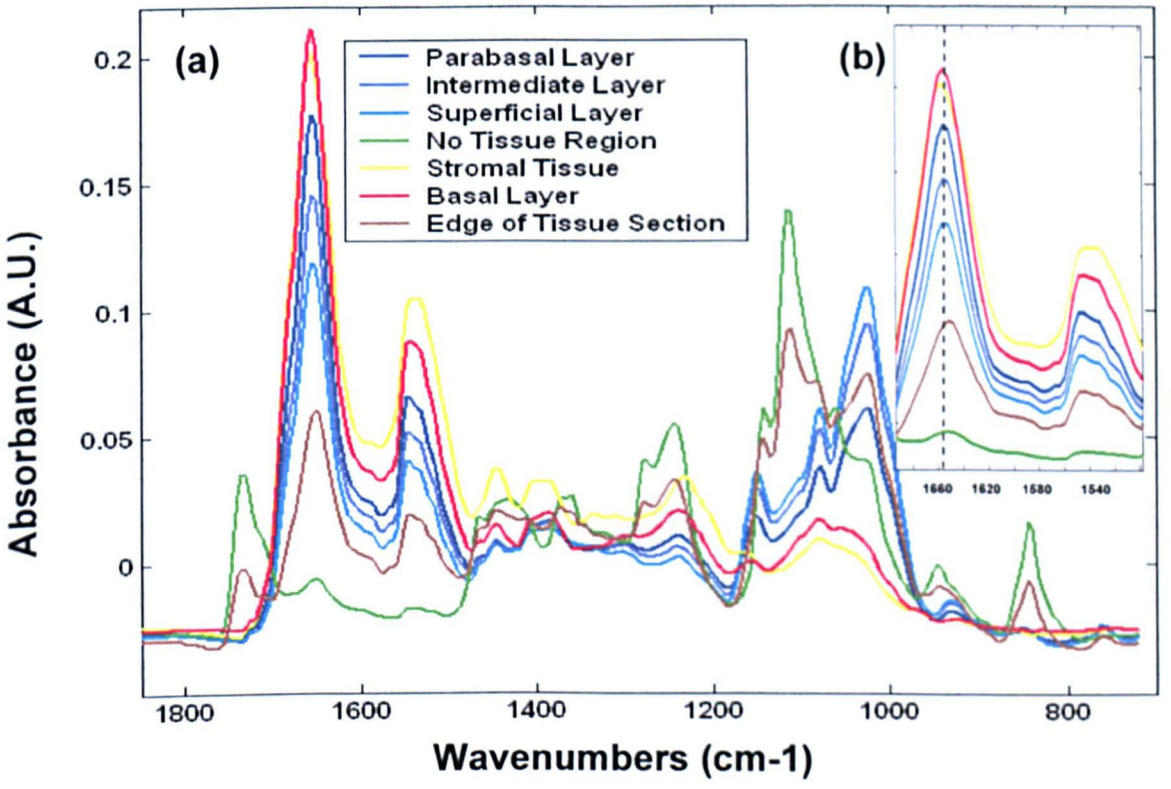

Figure 11: *7 Cluster PCA-FCM Analysis Results. (a) Spectral window displaying mean spectra between 1800 – 720cm^{-1}. (b) Spectral window displaying amide I and II region (1700-1500cm^{-1}).*

lie half on and off the tissue respectively. An additional advantage of the FCM clustering technique is that mean average spectra for each cluster in an analysis can be easily calculated and used to help interpret the biochemical differences that are occurring between them. The mean spectra calculated for the 7 cluster analysis are displayed in Figure 10. Although spectral changes are apparent across the entire spectrum, the most discernable occur within the spectral range 1800 – 720cm^{-1} (Figure 11). The stromal tissue (yellow spectrum) underlying the squamous epithelium is comprised mostly of vibrations due to structural proteins (e.g. collagen). A triad of peaks within the amide III region at 1205, 1232 and 1280 cm^{-1} is characteristic for these tissues, with a small broadening of the amide II peak at 1550 cm^{-1} also apparent [24,37,49]. The remaining spectral features found between 1150 and 1700 cm^{-1} are very similar between all spectra and are dominated by the spectral features of proteins. However all these peaks within the stromal spectra display markedly larger intensities, with the strong peak at c.a. 1450 cm^{-1} likely attributable to collagen [50]. Weak nucleic acid vibrations at 1030, 1060 and 1080 cm^{-1} are also evident characterising the small amount of nuclear material found within this tissue. All tissue sections examined during our investigations display a similar pattern for stromal tissue, with the triad of peaks found within the amide III region displaying a similar intensity ratio. The basal layer (red spectrum) of the squamous epithelium displays a similar spectral profile to the stromal tissue. However, these tissue types can be differentiated by examining the relative nucleic acid contributions found in the basal layer spectra. These display more prominent peaks at 1030, 1060 and 1080 cm^{-1}, demonstrating a greater nucleic acid contribution. The different intensity ratio of these peaks can also be used to easily discriminate these tissues apart. A decrease in the amide II / amide I ratio is also

apparent for basal layer spectra and is likely to reflect a contribution to the amide I intensity from a broad underlying DNA peak that displays peak maxima at 1690, 1655, and 1620cm^{-1} [37]. These observable spectral changes can be related back to the histological structures of the two cell types. Stromal tissue cells display very small nuclei containing tightly packed DNA and RNA strands. It has been estimated in previous studies of liver tissues [42] that dense nuclei such as those of the stromal tissue, are so tightly packed that they become almost opaque within the mid infrared range. Therefore, contributions from nuclear material to the spectra are negligible. In contrast the nuclei of basal layer cells are much larger and may occupy a volume 5 – 10 times larger. In this scenario the nucleus is far less tightly packed and may allow the transmission of light through the nuclear material and become apparent within the IR spectrum. The average spectra collected from the parabasal (dark blue spectrum), intermediate (royal blue spectrum) and superficial (cyan spectrum) layers of the squamous epithelium display similar spectral profiles. These spectral profiles can be differentiated from the stromal and basal layer cells by the appearance of strong glycogen absorptions. These occur at 1028, 1080 and 1152 cm^{-1} and are characteristic of the C–O–H deformation, C–C and C–O stretching modes of glycogen respectively [29, 33]. An additional small peak at 938 cm^{-1} is also likely attributable to glycogen [36]. As we move across the squamous epithelium from the parabasal to the superficial layer we can see a gradual increase in the intensity of the glycogen triplet. This quite nicely characterises the maturation of these epithelial cells, storing larger quantities of glycogen as they age. Only very small differences between the amide II / amide I ratio are observable between these layers. Therefore a better method to discriminate these tissue layers would be to use the ratio found between the amide I and 1080 cm^{-1} glycogen peaks. The average spectra calculated

for pixels that lie both half on and half off the tissue (brown spectrum), and entirely off the tissue section (green spectrum) display very strong absorptions due to mucus. Similar spectral profiles for pure cervical mucus have been reported previously in the literature [37,38], and display a characteristic triplet of peaks at 1060, 1115 and 1145 cm^{-1} respectively. These peaks can be assigned to the carbohydrate moieties of glycoproteins that comprise a major constituent of mucus. The discovery of such spectral profiles for pixels that lie off the tissue section would lead to the conclusion that a film of mucus must be surrounding the tissue section. This film of mucus is likely to have originated from the endocervical columnar epithelial cells, which naturally secrete mucus into the cervical canal.

Endocervix

The second region chosen for analysis from this tissue section is displayed in Figures 12a-c respectively, and describes the endocervix. At this part of the transformation zone, the underlying stromal tissue is now lined with a single layer of columnar epithelial cells (Figure 12c). By using a pixel size of 6.25 μm a total of 8806 individual IR spectra were collected from an area of 462.5 x 743.75 μm. The multivariate imaging results produced for this dataset are shown in Figure 13.

The first panel displays the PCA imaging results calculated from this dataset. It can be seen that over 95% of the total variance contained within this dataset is now comprised within the first five PC's of the analysis. When studying the colour weighted image constructed from the first PC in Figure (b), we can see that this PC clearly gives contrast between the tissue section (blue colour) and the surrounding

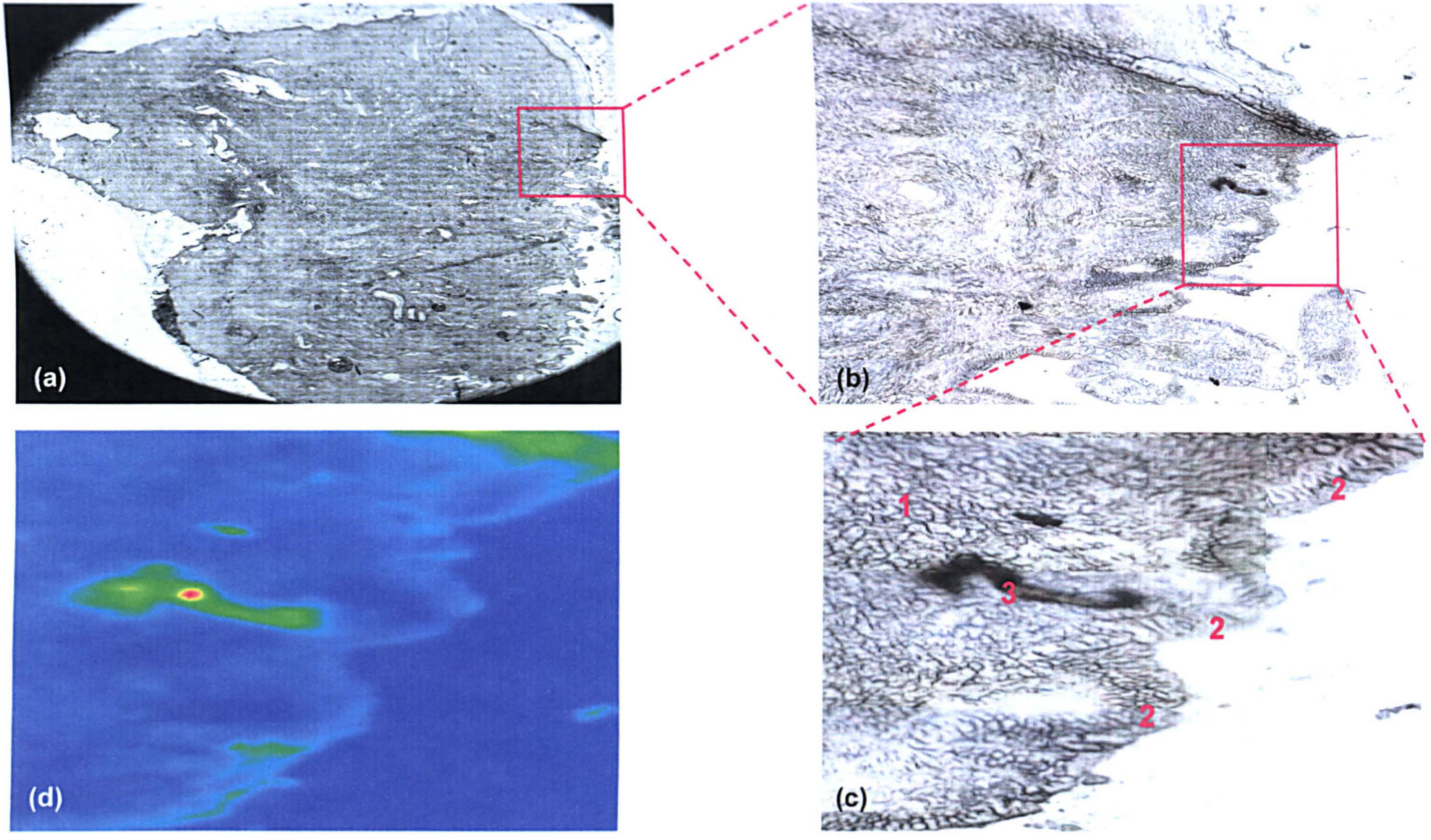

Figure 12: *a) White light image of entire cervical tissue section. b) White light image of transformation zone. c) Magnified region displaying benign anatomical features. (1) Underlying connective or stromal tissue, (2) columnar epithelium and (3) cellular debris. d) IR imaged area (462.5 x 743.75 μm) mapped using a pixel size of 6.25 μm for a total of 8806 individual IR spectra.*

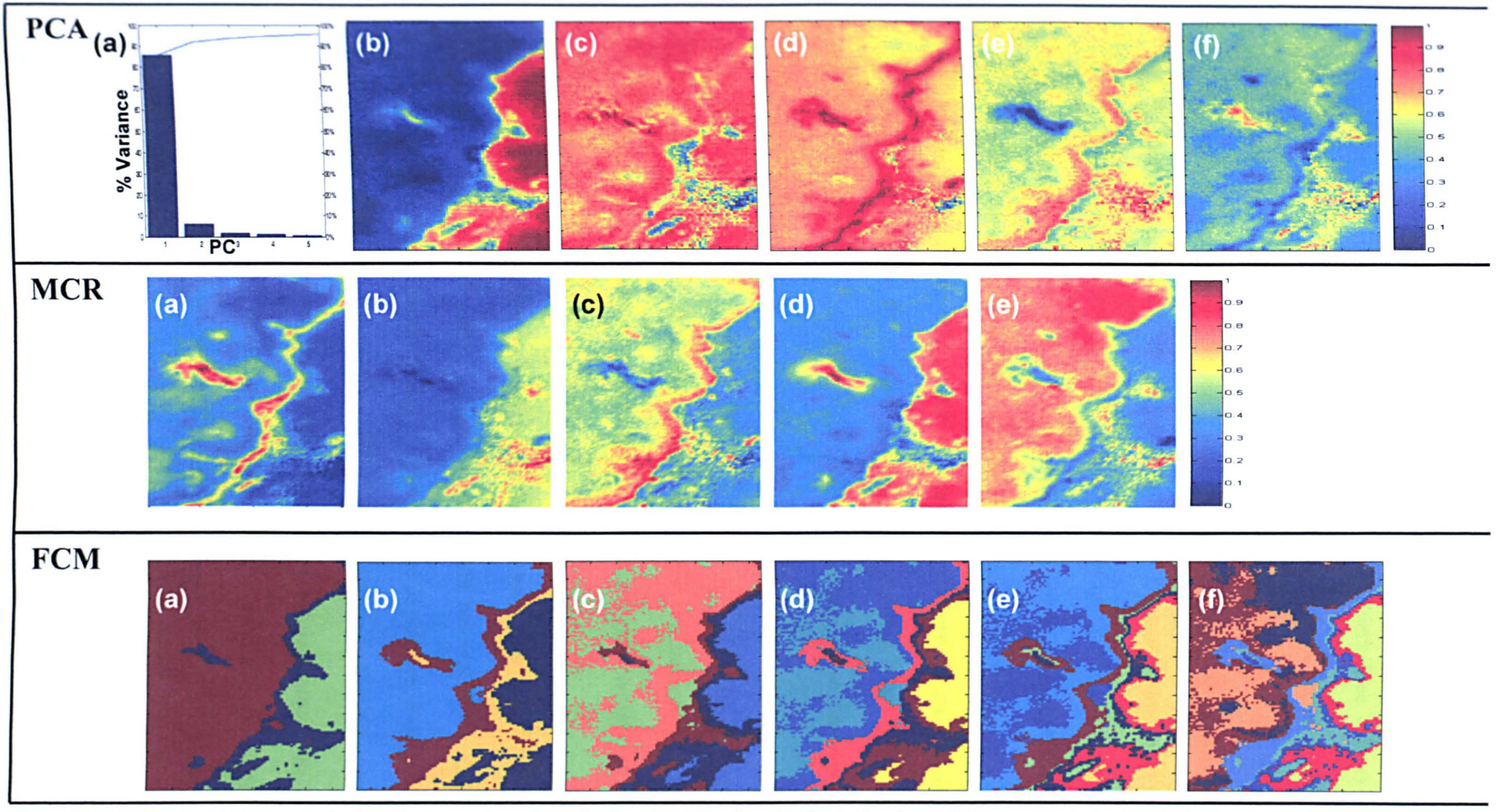

Figure 13: *Multivariate Imaging results from benign endocervix. PCA Panel: (a) Combined individual and cumulative percentage variance plot for the first 5 PC's. (b) – (f) False colour weighted images for PC's 1 – 5 respectively. Colour scale ranges from red indicating spectra that are very similar to that PC, and blue which are greatly dissimilar. MCR Panel: (a) – (e) False colour weighted images created from a 5 component MCR analysis. Colour scale ranges from red indicating spectra that are very similar to that component, and blue which are greatly dissimilar. FCM Panel: (a) – (f) False colour images created using PCA-FCM clustering analysis results. Note cluster numbers were subjectively increased from 3 – 8. Pixels with the same colour in each image are spectra that were partitioned into the same cluster.*

area where no tissue exists (red colour). This component also provides a small amount of contrast between the stromal tissue (dark blue colour) and columnar epithelium (cyan colour). The second PC image shown in Figure (c) provides contrast for an area off the tissue where a small piece of cellular debris exists (blue colour). In a similar fashion, the third component image (Figure d) again highlights this small piece of debris, but additionally provides a small amount of contrast for the pixels that lie close to or on the edge of the tissue section (dark red pigmentation). In comparison, the fourth PC image (Figure e) strongly highlights another piece of debris which has settled onto the top of the tissue section (blue pigmentation). This component image additionally provides contrast between the columnar epithelium (red colour) and the remaining mapped area (yellow colour). The fifth and all subsequent PC images provide little information about the tissue section and alternatively highlight regions that contain no tissue.

The MCR panel displays the resulting images constructed from a 5 component analysis of the same dataset (images a – e), which gave the best characterisation of the tissue section when compared against histology. The first component in the analysis (image a) clearly represents the squamous epithelium. However, this first component additionally highlights the debris that has settled onto the tissue section. The second component (image b) appears to be descriptive of areas where no tissue exists. In contrast, the third component (image c) clearly highlights the tissue section and provides some contrast between the stromal (yellow pigmentation) and columnar epithelium (red pigmentation). It is also noticeable that this component discriminates the cellular debris from the remaining tissue section. Examining the fourth component (image d), this again clearly highlights the area with no tissue as found by

the second component. However, on this occasion the fourth component also shows similar intensities for the area where cellular debris exists on the tissue. This is an interesting finding as it highlights that the spectra collected from these pixels must have spectral features that are both similar to columnar epithelium (first component) and the surrounding area where no tissue is apparent (component 4). The fifth and final component (image e) nicely highlights the stromal tissue (red pigmentation) and provides contrast between the columnar epithelium (yellow pigmentation).

The final panel displays images created via PCA-FCM Clustering. Images (a) to (f) were constructed by subjectively increasing the amount of clusters found by the analysis from 3 – 8 respectively. When comparing these created images directly against the known tissue type regions, the image constructed from a 6 cluster analysis appears to best mimic the histological architecture of the tissue section (image d). The stromal tissue spectra have been partitioned into two clusters that describe both immature (royal blue) and mature (cyan) stromal cells. In contrast, the columnar epithelium spectra have been partitioned into a single group that are characterised by pixels with an orange pigmentation. The spectra collected from areas where no tissue exists have alternatively been partitioned into three separate clusters (yellow, blue and maroon pigmentation).

As highlighted previously, an additional benefit of the FCM clustering technique is the ability to calculate mean average spectra for clusters produced by the analysis. The mean spectra calculated for the 6 cluster analysis are displayed in Figure 14. Spectral changes are apparent across the entire spectrum, but the most discernable occur within the spectral range 1800 – 720 cm^{-1} (Figure 15). Both the immature

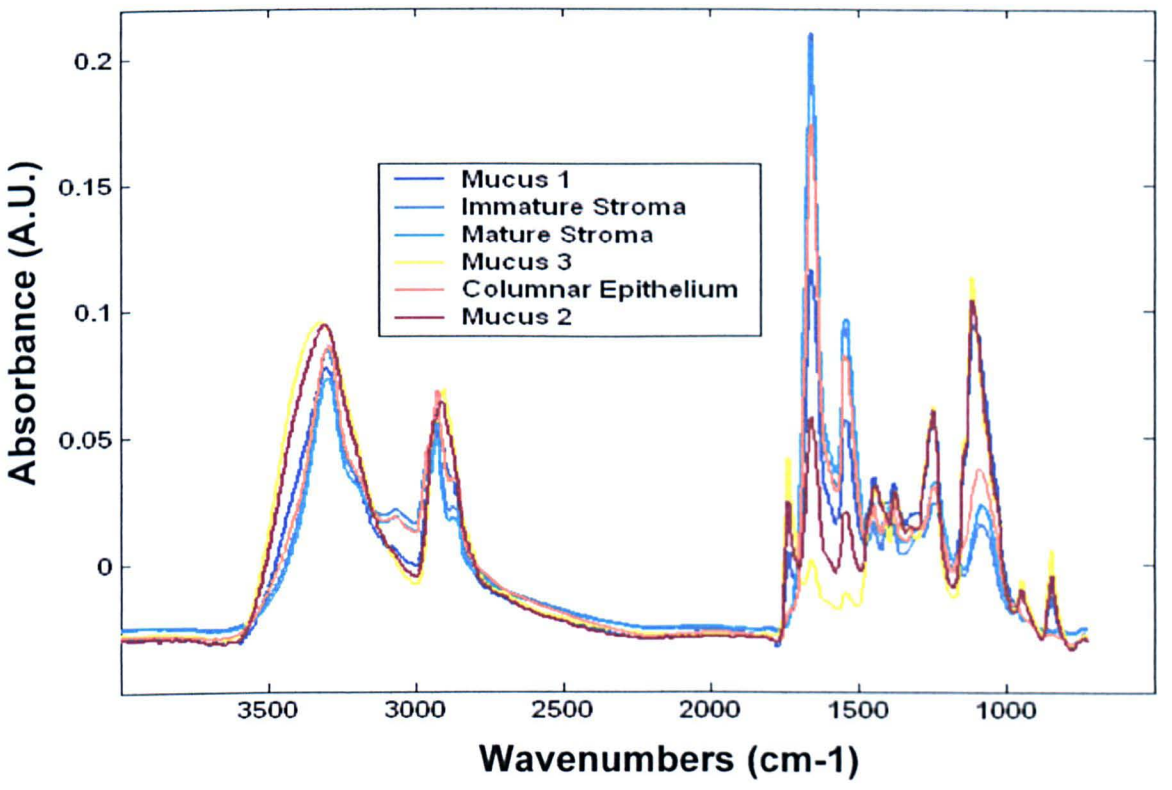

Figure 14: *6 Cluster PCA-FCM Analysis Results. Mean average spectra for each cluster in the analysis.*

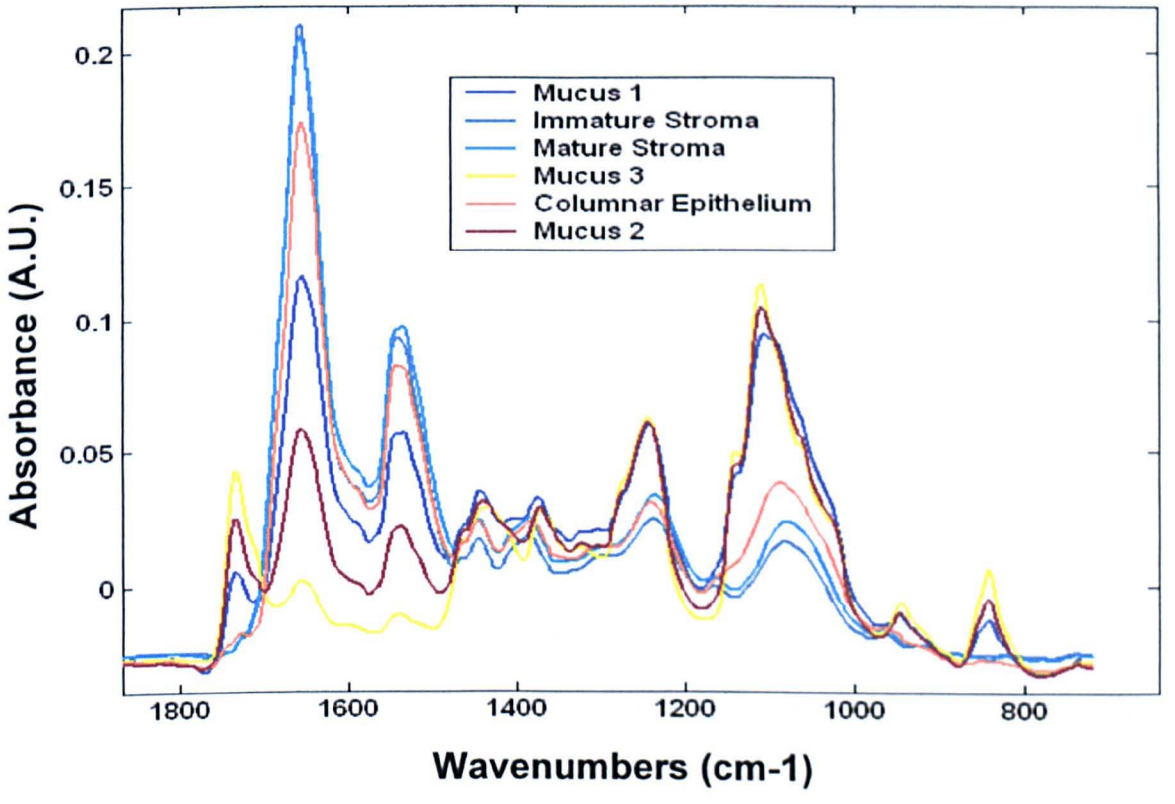

Figure 15: *6 Cluster PCA-FCM Analysis Results. Spectral window displaying mean spectra between 1800 – 720cm^{-1}.*

(blue spectrum) and mature (cyan spectrum) stromal tissue underlying the columnar epithelium display very similar spectral profiles. These tissues are again dominated by spectral features characteristic of structural proteins and exhibit a triad of peaks within the amide III region at 1205, 1232 and 1280 cm^{-1} as observed previously. When stromal cells mature they accumulate collagen within the cytoplasm and become enlarged. Thus the observed increase in intensity for this triad of peaks and the collagen band at 1450 cm^{-1} for mature stromal tissue spectra would make histological sense. A broadening of the amide II band is also noticeable with the peak maxima shifted to a lower wavenumber at 1536 cm^{-1}. In contrast to squamous epithelial cells, the columnar cells mature along the surface of the epithelium and produce a single cell layer. Furthermore, glycogen is not accumulated within the cytoplasm of these cells. The tissue spectra for columnar cells (orange spectrum), therefore, lack the spectral features of glycogen and better resemble the spectral profile shown previously for basal layer cells within the squamous epithelium. However, discrimination between these tissue types can be achieved by examining the 1000 – 1200 cm^{-1} spectral region. Columnar cells alternatively display a broad peak across this region, with a spectral pattern very similar to that previously revealed for mucus. This finding would also make histological sense as one of the physiological roles of this type of cell is to allow the secretion of mucus into the endocervical canal. The cytoplasm of these types of cell therefore appear to contain large amounts of glycoproteins opposed to glycogen found within mature squamous cells. A similar spectral profile was reported by Chriboga *et al* [39] who examined exfoliated endocervical material that was fractioned to only include cells larger than 5μm but smaller than 12μm. The brown, blue and yellow spectral profiles describe the three clusters of spectra that were partitioned from pixels lying off the tissue. As

seen previously in our analysis of the ectocervix, the tissue section has been surrounded by a film of cervical mucus. The three alternate clusters simply highlight a change in the concentration of glycoproteins within the surrounding mucus.

3.3.1.2 FTIR Multivariate Imaging of Diseased Cervical Tissue Sections

Ectocervix

The second aim of this study was to identify spectral variations that may characterise biochemical changes within diseased cervical tissue. Biopsy material was therefore collected from a patient that had previously exhibited an abnormal cervical PAP smear, displaying high grade intraepithelial lesions (HSIL). H&E stained images taken from the directly parallel section used for analysis are shown in Figures 16a – b respectively. These clearly illustrate the ectocervix region of the transformation zone and allow the visualisation of regions upon the squamous epithelium that were clinically diagnosed as being healthy and diseased in nature. White light images collected from the same region of the analysed tissue section (named C010406) are displayed in Figures 16c – d. To assess whether any notable biochemical changes were apparent within the directly adjacent healthy squamous epithelial cells, an IR map was collected from this region. A magnified image detailing the mapped region upon the tissue section is further displayed in Figure 16e. Examining the H&E stained images we can again visualise the main tissue types present. These include the basal, parabasal, intermediate and superficial layers of the squamous epithelium that surround the underlying stromal tissue. Using a pixel size of 10μm a total of 891

individual IR spectra were collected from an area of 720 x 90 μm. The multivariate imaging results produced for this dataset are shown in Figure 17.

The first panel displays the PCA imaging results calculated from this dataset. It can be seen that over 96% of the total variance contained within this dataset is now comprised within the first five PC's of the analysis. When studying the colour weighted image constructed from the first PC in Figure (b), we can see that this PC clearly gives contrast between the tissue section and the surrounding area where no tissue exists. This component also provides a small amount of contrast between the superficial layers of the squamous epithelium (cyan colour) and the remaining tissue types. The second PC image shown in Figure (c) again highlights the area off the tissue but alternatively provides contrast for the connective and basal layers of the tissue (cyan colour). Subsequent PC images provide little information about the tissue section and alternatively highlight regions that contain no tissue or pixels that lie half on and off the tissue section.

The MCR panel displays the resulting images constructed from both a 3 and 4 component analysis of the same dataset (images a – g). By direct comparison to histology, the overall best characterisation of the tissue section was achieved via the 3 component analysis. The first component in the analysis (image a) clearly represents the underlying connective or stromal tissue (red colour). However, this first component additionally provides a small amount of contrast for the remaining tissue types. The basal and parabasal layers of the epithelium are highlighted by a light red and yellow colouration, whereas the outer layers are more clearly discerned via a cyan pigmentation. The second component (image b) appears to be solely

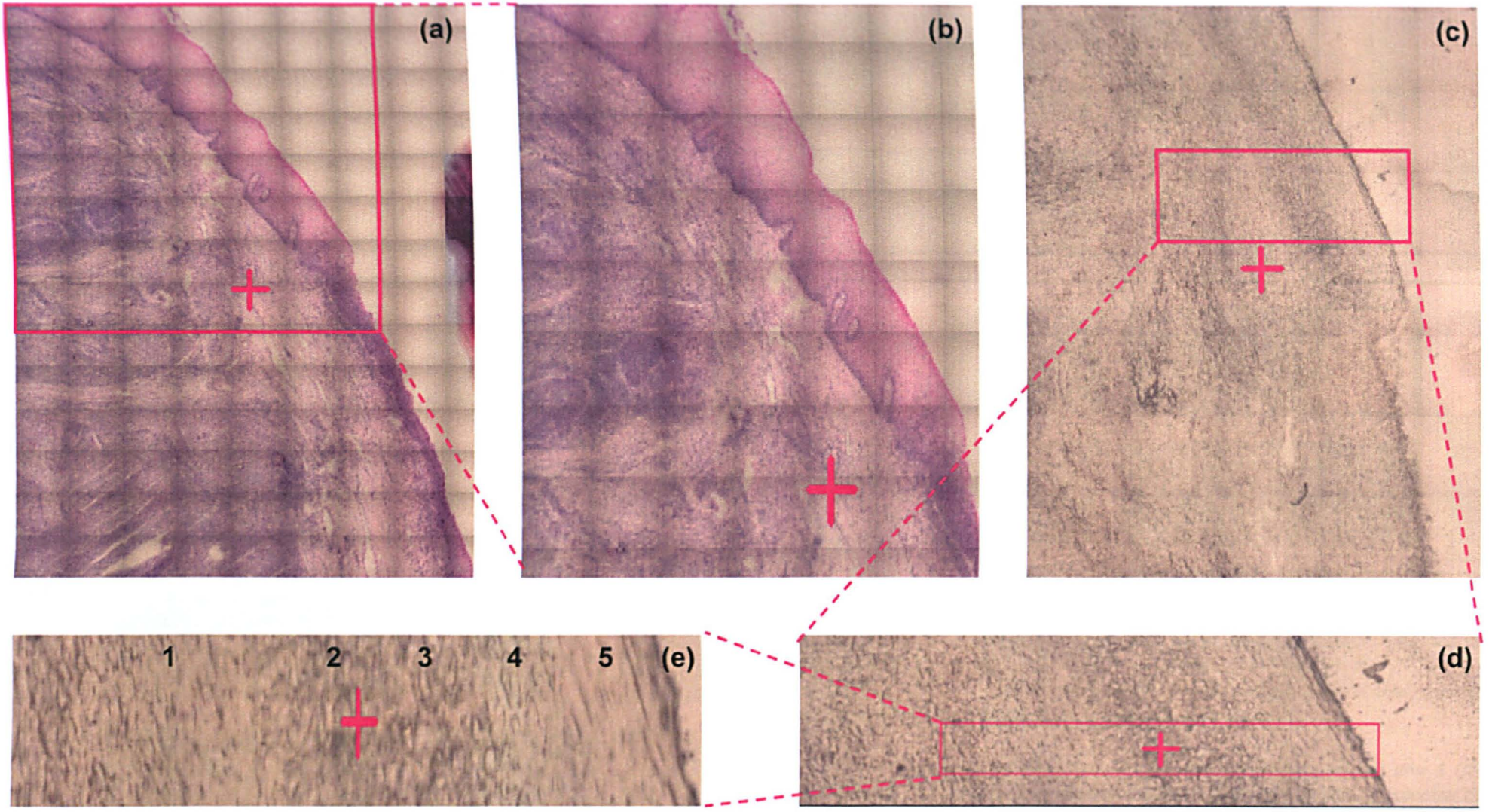

Figure 16: *a) H&E photomicrograph of entire cervical tissue section. b) H&E photomicrograph of squamous epithelium. c) White light image of same region upon analysed tissue section. d) Magnified cross section of squamous epithelium. e) Magnified image of examined region (720 x 90 µm) mapped using a pixel size of 10 µm for a total of 891 individual IR spectra. Both benign and malign anatomical features can be identified including (1) underlying connective or stromal tissue, (2) basal layer, (3) parabasal layer, (4) intermediate layer and (5) superficial layer of squamous epithelium.*

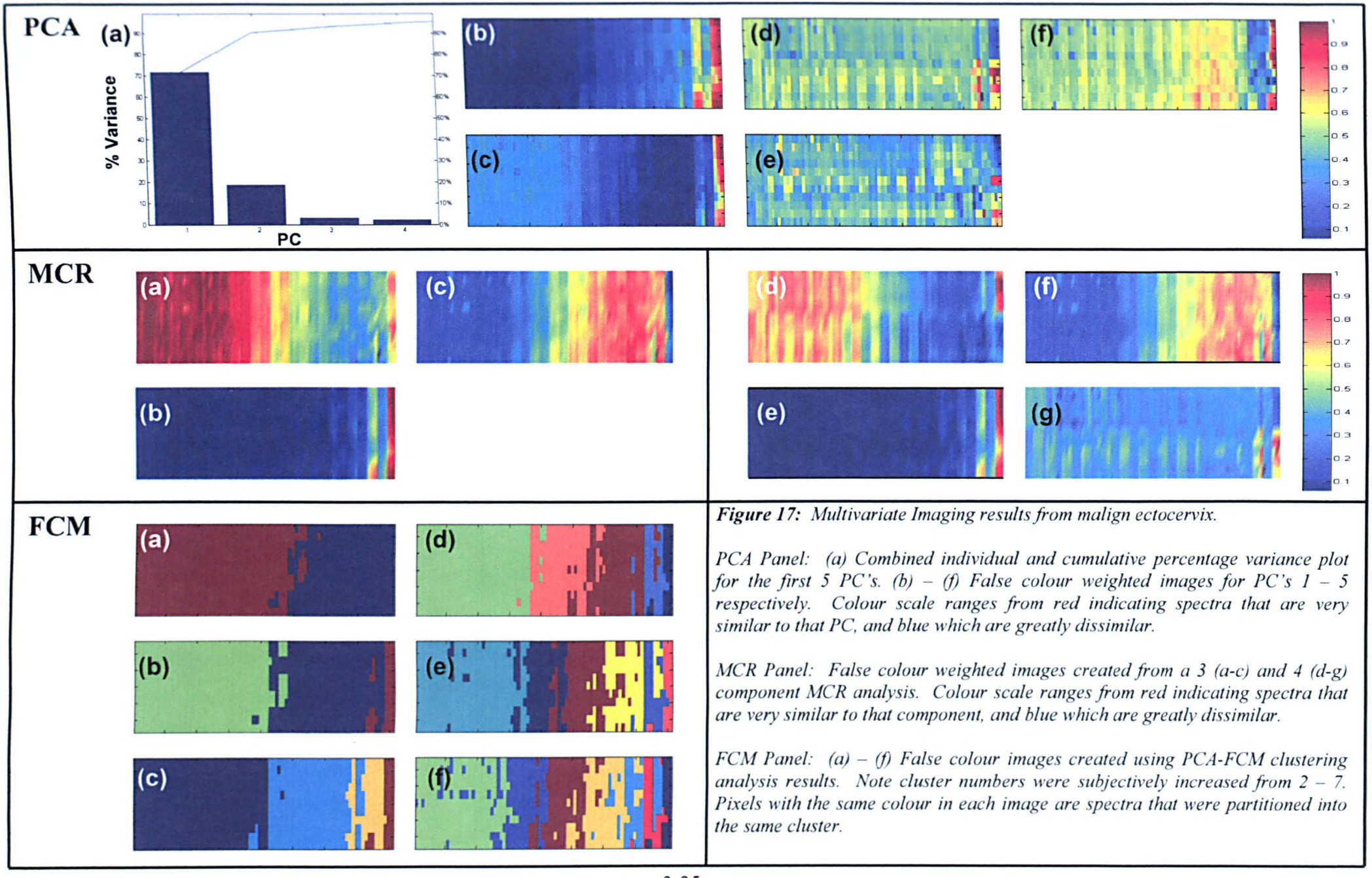

Figure 17: *Multivariate Imaging results from malign ectocervix.*

PCA Panel: (a) Combined individual and cumulative percentage variance plot for the first 5 PC's. (b) – (f) False colour weighted images for PC's 1 – 5 respectively. Colour scale ranges from red indicating spectra that are very similar to that PC, and blue which are greatly dissimilar.

MCR Panel: False colour weighted images created from a 3 (a-c) and 4 (d-g) component MCR analysis. Colour scale ranges from red indicating spectra that are very similar to that component, and blue which are greatly dissimilar.

FCM Panel: (a) – (f) False colour images created using PCA-FCM clustering analysis results. Note cluster numbers were subjectively increased from 2 – 7. Pixels with the same colour in each image are spectra that were partitioned into the same cluster.

descriptive of areas where no tissue exists. In contrast, the third component (image c) clearly highlights the tissue section and provides some contrast between the stroma (blue), basal (yellow) and superficial layers (red) of the epithelium.

The final panel displays images created via PCA-FCM Clustering. Images (a) to (f) were constructed by subjectively increasing the amount of clusters found by the analysis from 2 – 7 respectively. When comparing these created images directly against the known tissue type regions, the image constructed from a 7 cluster analysis appears to best mimic the histological architecture of the tissue section (image f). The squamous epithelium is now characterised by individual clusters that describe the basal (light blue), parabasal (maroon), intermediate (yellow) and superficial (cyan) layers that illustrate the maturation of these epithelial cells. In contrast, the green cluster of spectra highlights the underlying stromal tissue. The final blue and red clusters describe areas where there is no tissue or pixels that lie half on and off the tissue respectively.

Mean spectra calculated from the 7 cluster analysis are displayed in Figure 18 (1800 – 720 cm^{-1} spectral region). The overall spectral profiles of the tissue types present are very similar to those found previously in healthy cervical tissue. Stromal or connective tissue (green spectrum) again displays a pronounced triad of peaks within the 1300 – 1200 cm^{-1} spectral region that are characteristic of collagen contributions. The basal (light blue spectrum) and parabasal (maroon spectrum) layers of the epithelium are discernable via a reduction in the amide II / amide I ratio with little or no contribution from glycogen noticeable within the low-frequency window of 1200 – 900 cm^{-1}. Higher concentrations of glycogen are again apparent within spectra

representative of the intermediate (yellow spectrum) and superficial (cyan spectrum) layers of the squamous epithelium. However, these levels are substantially lower than those observed in tissue collected from healthy patients, as shown previously. The lowering or absence of glycogen in these mature layers of epithelium appears consistent among additional diseased tissue sections analysed and has been reported in similar studies of cervical tissue [51].

The second region examined upon this tissue section incorporated an area of diseased squamous epithelium, clinically diagnosed as being CIN II / III. H&E stained images taken from the same region of the directly parallel section used for analysis are shown in Figures 19a – b respectively. Areas of diseased squamous epithelium can be visualised with a dark purple pigmentation. These abnormal cells are more cubodial in shape, displaying large round nuclei with scant cytoplasm. White light images collected from the same region of the analysed tissue section are displayed in Figures 19c – d. A magnified image detailing the mapped region upon the tissue section is further displayed in Figure 19e. Using a pixel size of 10 μm a total of 605 individual IR spectra were collected from an area of 486 x 90μm. The multivariate imaging results produced for this dataset are shown in Figure 20.

The first panel displays the PCA imaging results calculated from this dataset. It can be seen that the overwhelming majority of the total variance contained within this dataset is now comprised by the first five PC's of the analysis. When studying the colour weighted image constructed from the first PC in Figure (b), we can see that this PC clearly highlights the area where no tissue exists. In contrast, the second PC image shown in Figure (c) highlights the tissue section itself. The third PC image

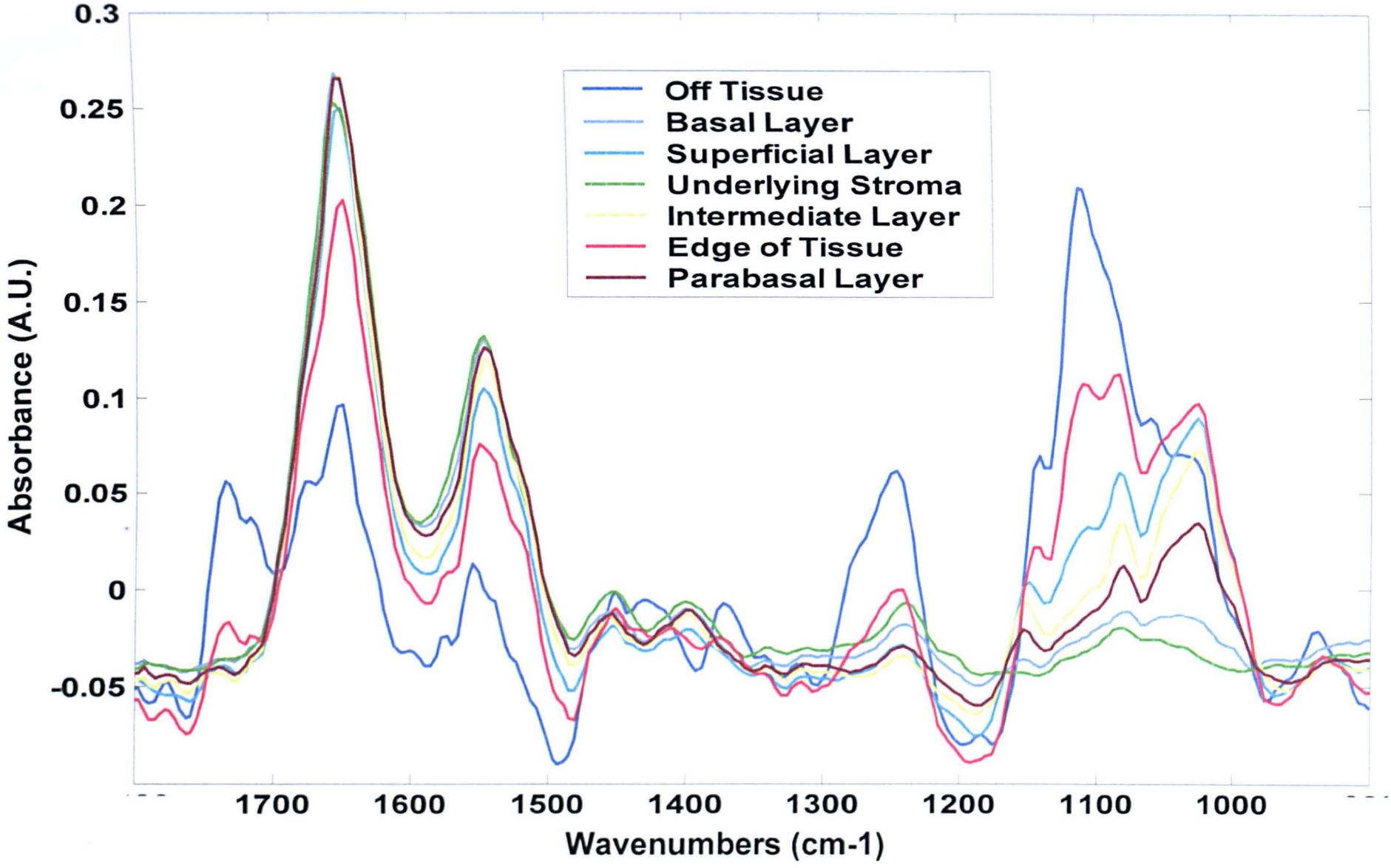

Figure 18: *7 Cluster PCA-FCM Analysis Results. Spectral window displaying mean spectra between 1800 – 720cm^{-1}.*

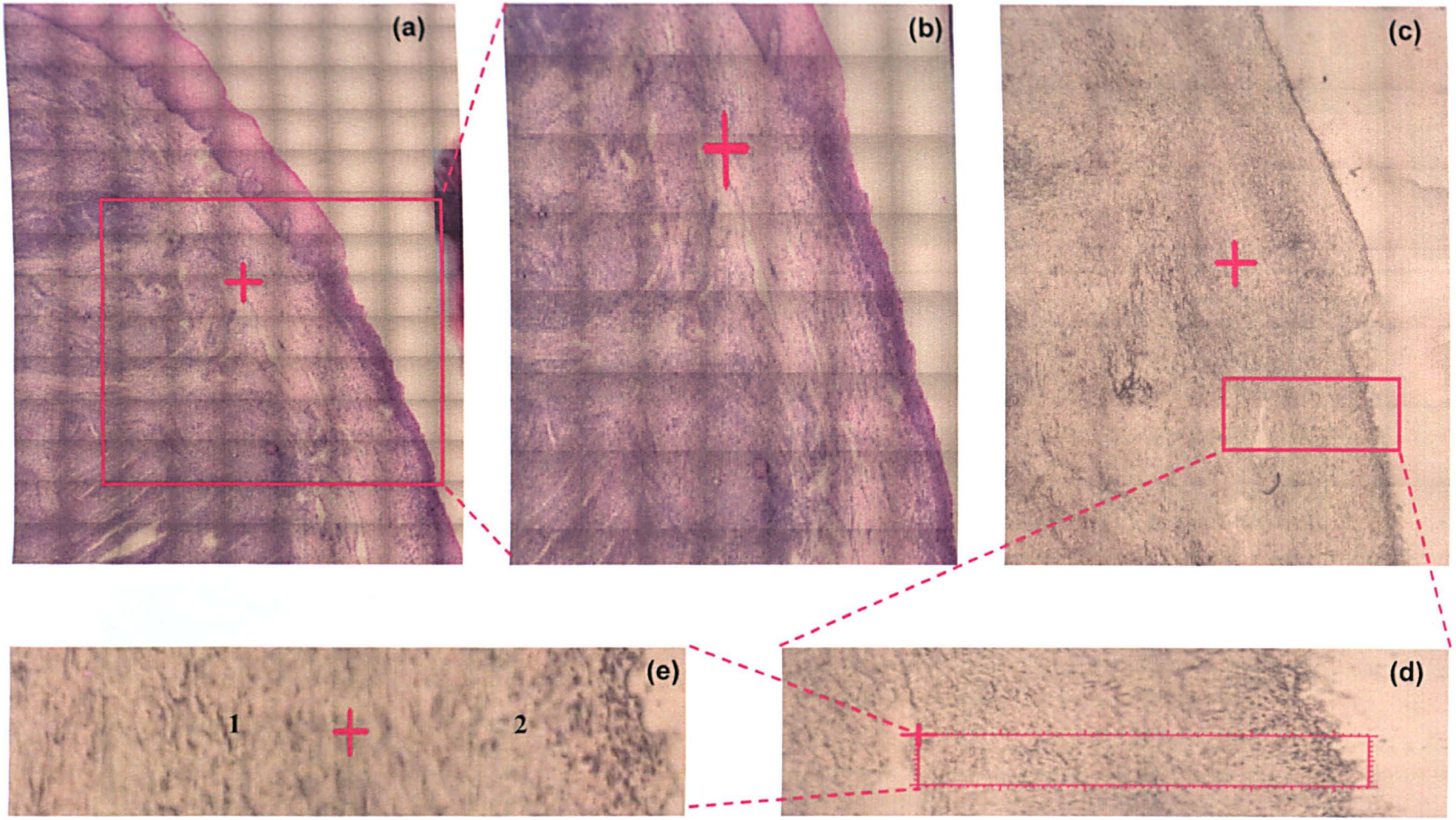

Figure 19: *a) H&E photomicrograph of entire cervical tissue section. b) H&E photomicrograph of malign squamous epithelium. c) White light image of same region upon analysed tissue section. d) Magnified image of examined region (486 x 90 µm) mapped using a pixel size of 10µm for a total of 605 individual IR spectra. Both benign and malign anatomical features can be identified including (1) underlying connective or stromal tissue and (2) malign squamous epithelium (C.I.N II / III).*

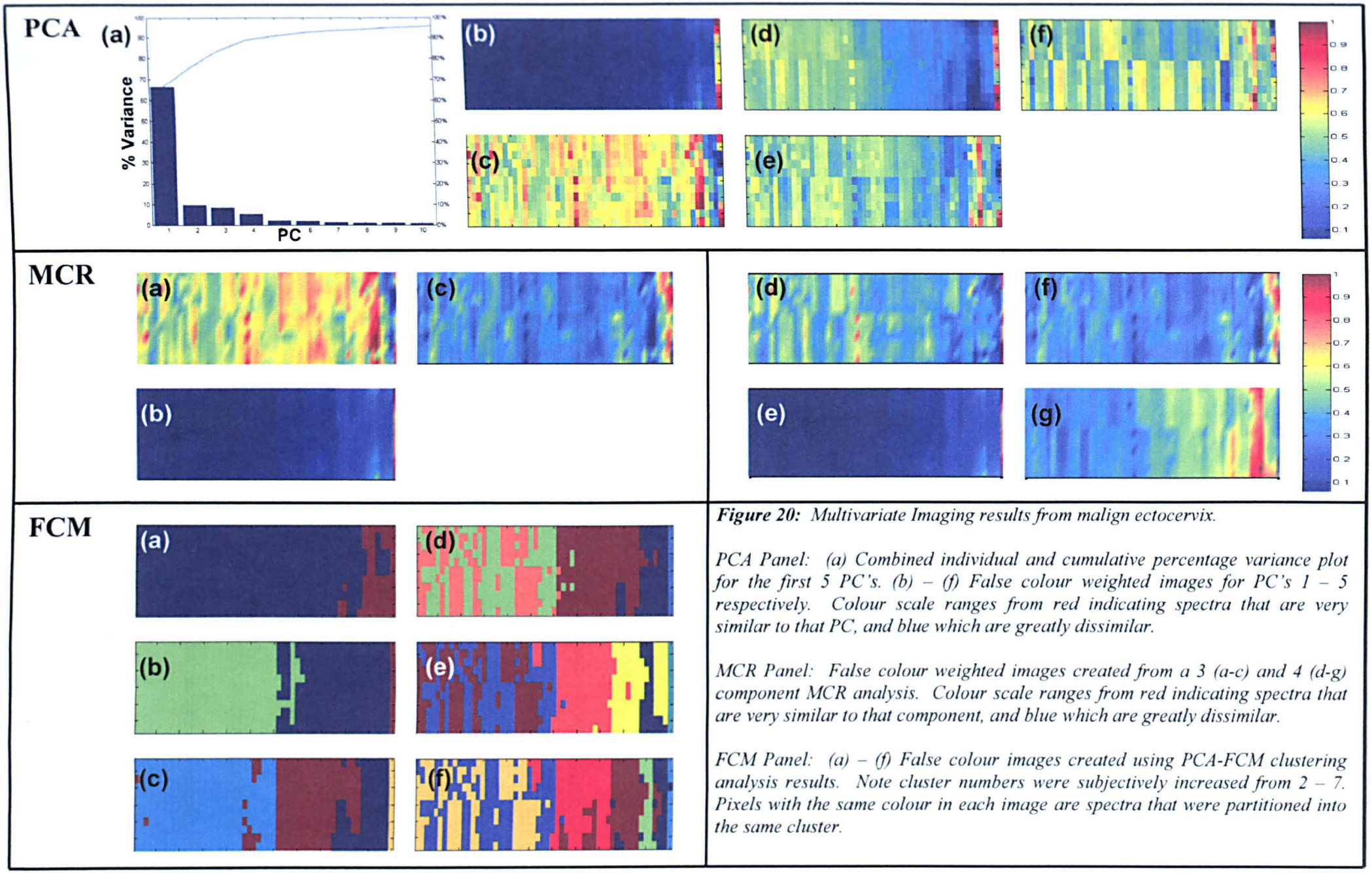

Figure 20: *Multivariate Imaging results from malign ectocervix.*

PCA Panel: (a) Combined individual and cumulative percentage variance plot for the first 5 PC's. (b) – (f) False colour weighted images for PC's 1 – 5 respectively. Colour scale ranges from red indicating spectra that are very similar to that PC, and blue which are greatly dissimilar.

MCR Panel: False colour weighted images created from a 3 (a-c) and 4 (d-g) component MCR analysis. Colour scale ranges from red indicating spectra that are very similar to that component, and blue which are greatly dissimilar.

FCM Panel: (a) – (f) False colour images created using PCA-FCM clustering analysis results. Note cluster numbers were subjectively increased from 2 – 7. Pixels with the same colour in each image are spectra that were partitioned into the same cluster.

provides a greater amount of tissue contrast and appears to highlight the underlying connective tissue in yellow and the diseased squamous epithelium with a cyan colour. Subsequent PC images appear confused and provide no real information about the tissue section that is beneficial for tissue discrimination.

The MCR panel displays the resulting images constructed from both a 3 and 4 component analysis of the same dataset (images a – g). By direct comparison to histology, the overall best characterisation of the tissue section was achieved via the 4 component analysis. The first component in the analysis (image d) provides a small amount of contrast for the underlying stromal tissue displaying a yellow and red colouration. However, the outlying pixels off the tissue are also highlighted with the same intensity. Both the second (image e) and third (image f) component images appear to highlight the region upon the mapped area where no tissue exists. In contrast, the fourth and final component (image g) more distinctly characterises the diseased squamous epithelium with a bright red and yellow colouration.

The final panel displays images created via PCA-FCM Clustering. Images (a) to (f) were constructed by subjectively increasing the amount of clusters found by the analysis from 2 – 7 respectively. When comparing these created images directly against the known tissue type regions, the image constructed from a 6 cluster analysis appears to best mimic the histological architecture of the tissue section (image e). The diseased squamous epithelium is characterised by three separate clusters. The orange cluster of spectra describe the basal layer, the yellow the intermediate and superficial layers, and the blue colour highlights a small cluster of spectra grouped within the superficial layer. In contrast, the underlying stromal

tissue is characterised by two clusters coloured maroon and light blue. The final cyan cluster of spectra characterises the region upon the map where no tissue existed.

Mean spectra calculated from the 6 cluster analysis are displayed in Figure 21 (1800 – 720 cm^{-1} spectral region). At first glance the spectral profiles of the different tissue types found in this map are very similar. However, distinct spectral differences are noticeable at frequencies below 1200 cm^{-1}. Since these tissue layers appear free of glycogen contribution, it is reasonable to assume that these differences are most likely due to vibrations of phosphate (PO_2^-) groups contained within RNA and DNA. The average spectra that describe the diseased squamous epithelium (orange, yellow and blue spectra) all display pronounced symmetric phosphate bands at 1080 cm^{-1}, which is likely to describe the increased nucleic acid concentration within these abnormal cells. As we move across the epithelium the intensity of this band increases and is coupled with a gradual reduction in the amide II / amide I ratio. This pattern may indicate that the blue cluster of spectra describes a group of cells that are distinctly more abnormal or malignant than those surrounding them. Both the symmetric and antisymmetric phosphate bands (1080 and 1240 cm^{-1} respectively) appear more intense than the methyl and methlene deformation modes (1450 – 1350 cm^{-1}) in these spectra, an observation not previously seen in our analysis of healthy squamous tissue. These findings are in agreement with early studies upon exfoliated cervical cells [52], and more recent tissue mapping experiments undertaken by Diem and coworkers [53]. The two clusters that describe the underlying stromal tissue (maroon and light blue clusters) are discernable via a significantly larger amide II / amide I ratio that is coupled with a smaller nucleic acid contribution to the spectra. The average spectrum that describes the region off the tissue section (cyan colour) is

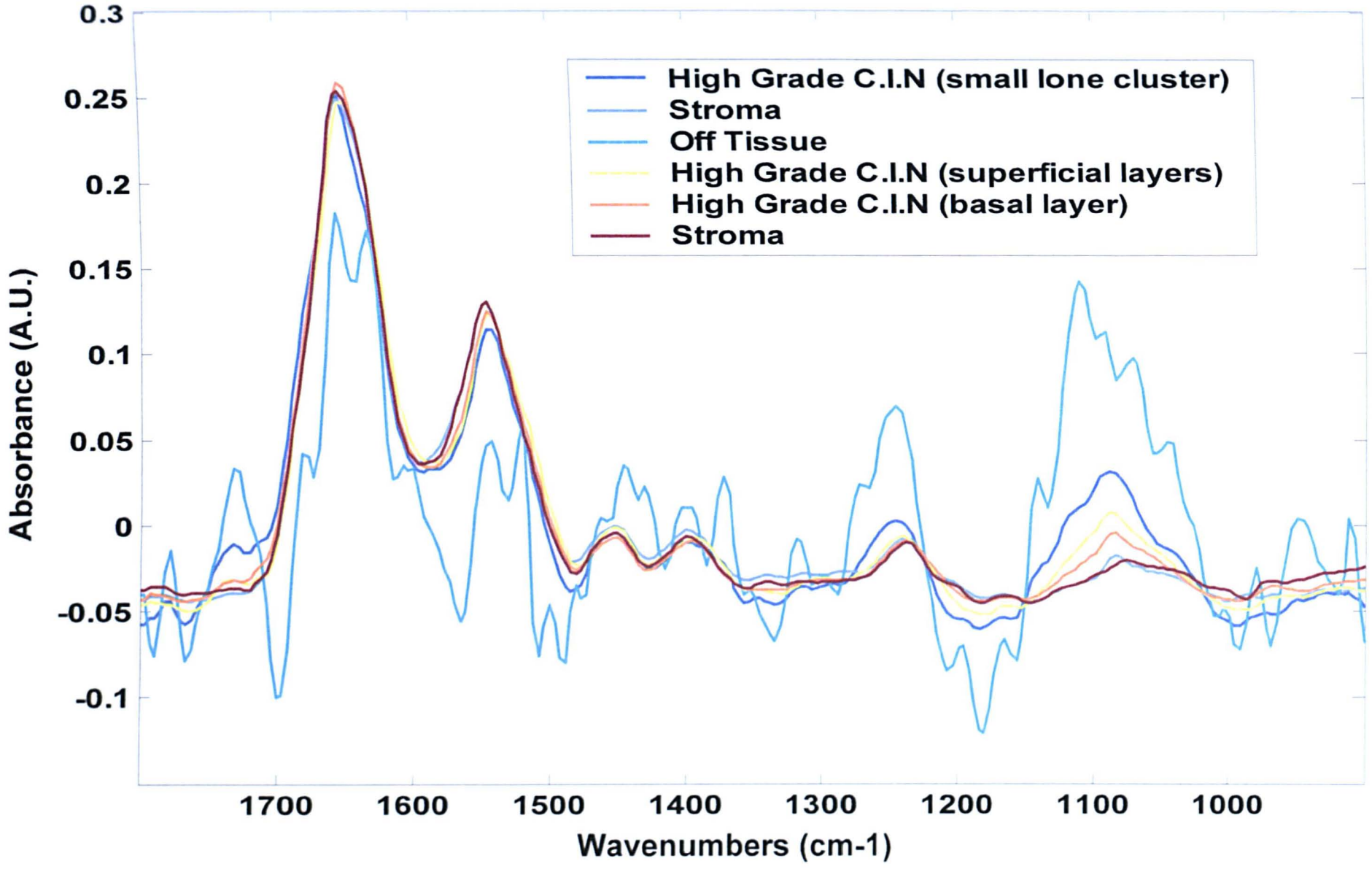

Figure 21: *6 Cluster PCA-FCM Analysis Results. Spectral window displaying mean spectra between 1800 – 720cm^{-1}.*

noticeably noisy, but does however reveal an outline that bares both tissue and cervical mucus characteristics.

Endocervix

In addition to the more common squamous carcinoma that originates within the ectocervix, a second type of malignancy can be found within the cervix named adenocarcinoma. This type of malignancy originates within the endocervix whereby columnar epithelium is infiltrated and replaced by abnormal cells. To again establish whether any marked biochemical changes are apparent with the onset of this type of disease, biopsy material was collected from a patient that had exhibited such abnormal changes in previous smear and tissue section screenings. H&E stained images taken from the directly parallel section used for analysis are shown in Figures 22a – b respectively. These clearly illustrate the endocervix region of the transformation zone and allow the visualisation of regions clinically diagnosed as being diseased in nature. A white light image collected from the same region of the analysed tissue section (named C100406) is displayed in Figure 22c. A magnified image detailing the mapped region upon the tissue section is further displayed in Figure 22d. Examining the H&E stained image in Figure 22b we can again visualise the main tissue types present. These include diseased columnar epithelium, healthy connective or stromal tissue and a small pocket of red blood cells. Using a pixel size of 10 μm a total of 2397 individual IR spectra were collected from an area of 414 x 450μm. The multivariate imaging results produced for this dataset are displayed in Figure 23.

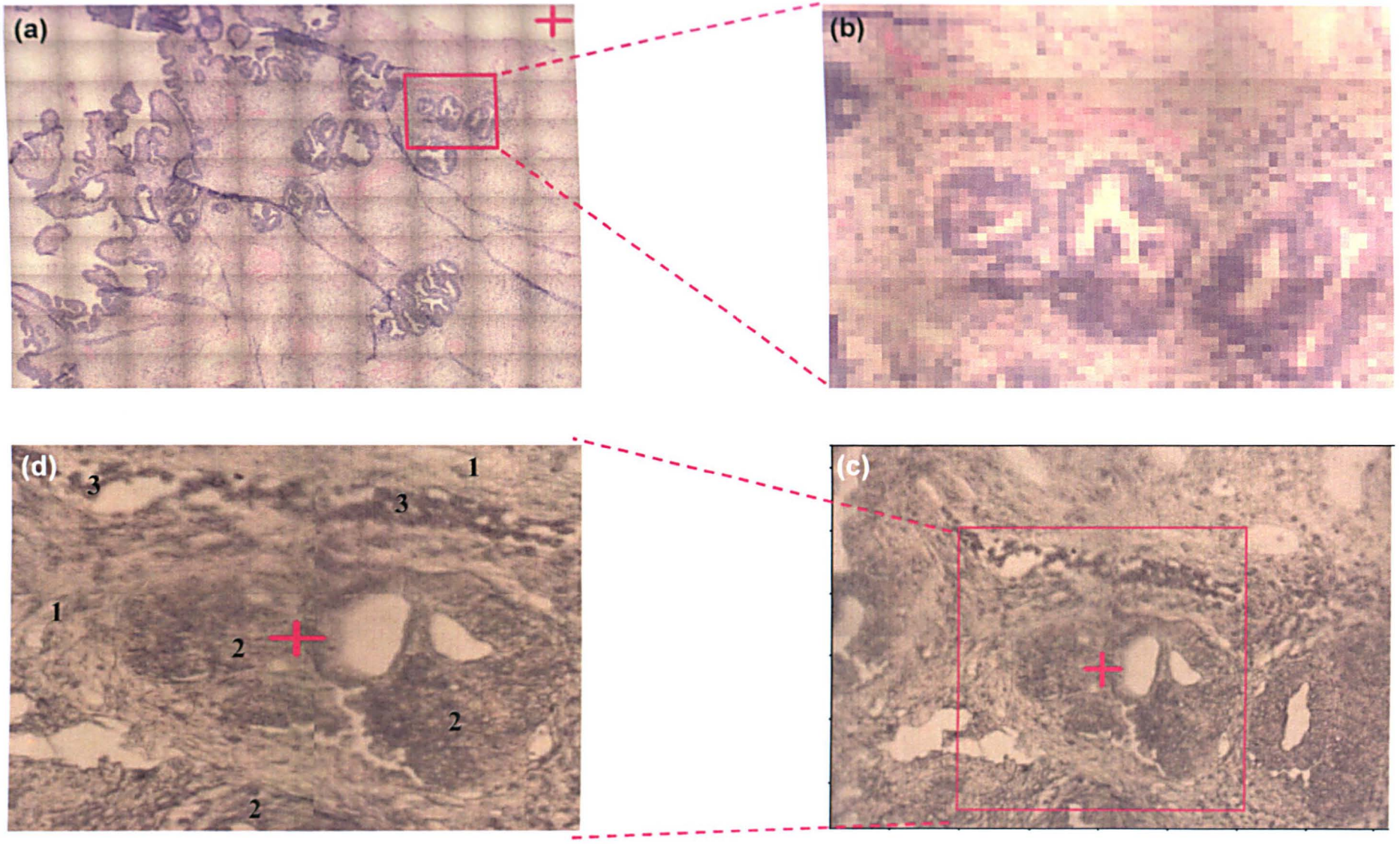

Figure 22: *a) H&E photomicrograph of entire cervical tissue section. b) H&E photomicrograph of malign columnar epithelium. c) White light image of same region upon analysed tissue section. e) Magnified image of examined region (414 x 450 μm) mapped using a pixel of 10μm for a total of 2397 individual IR spectra. Both benign and malign anatomical features can be identified including (1) connective or stromal tissue, (2) malign columnar epithelium (adenocarcinoma) and (3) blood cells.*

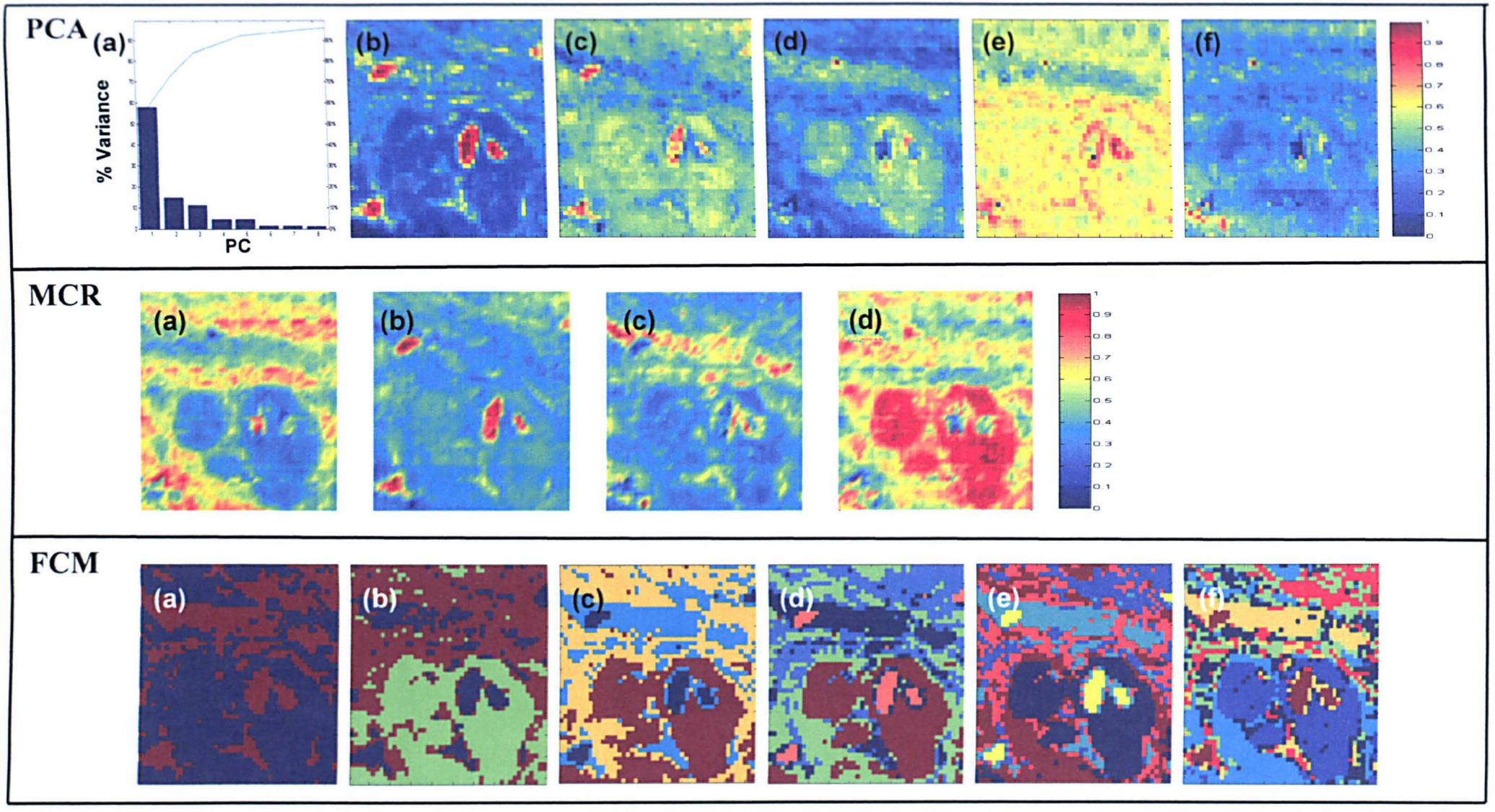

Figure 23: *Multivariate Imaging results from adenocarcinoma. PCA Panel: (a) Combined individual and cumulative percentage variance plot for the first 5 PC's. (b) – (f) False colour weighted images for PC's 1 – 5 respectively. Colour scale ranges from red indicating spectra that are very similar to that PC, and blue which are greatly dissimilar. MCR Panel: (a) – (e) False colour weighted images created from a 4 component MCR analysis. Colour scale ranges from red indicating spectra that are very similar to that component, and blue which are greatly dissimilar. FCM Panel: (a) – (f) False colour images created using PCA-FCM clustering analysis results. Note cluster numbers were subjectively increased from 2 – 7. Pixels with the same colour in each image are spectra that were partitioned into the same cluster.*

The first panel displays the PCA imaging results calculated from this dataset. It can be seen that the overwhelming majority of the total variance contained within this dataset is now comprised by the first five PC's of the analysis. When studying the colour weighted image constructed from the first PC in Figure (b), we can see that this PC marks areas upon the tissue section where holes are prevalent. In contrast, the second PC image displayed in Figure (c) appears more confused. Again the holes within the tissue are highlighted intensely with a red colouration, but the majority of the tissue section, including the adenocarcinoma, is now given contrast with a yellow colour. However, the regions where blood cells exist are not correlated highly with this component and are given a deep blue colour. The third PC image displayed in Figure (d) marks both the red blood cells and the region of adenocarinoma with a yellow colour. The remaining normal connective tissue displays a poor correlation to this component and is highlighted in blue. The fourth PC image (e) again highlights the majority of the tissue section with a yellow colouration, but does however give clear contrast for the red blood cells (cyan colour). The fifth (f) and subsequent PC images constructed by the analysis did not provide any further beneficial tissue discrimination and were dominated by spectra collected at areas where no tissue existed.

The MCR panel displays the resulting images constructed from a 4 component analysis of the same dataset (images a – d), which gave the best characterisation of the tissue section when compared against histological diagnosis. The first component in the analysis (image a) nicely highlights the connective or stromal tissue within the mapped area with a red and yellow colouration. The second component (image b) alternatively marks regions where no tissue exists (red colour).

In contrast, the third component in the analysis (image c) clearly marks the small pockets of red blood cells found within the tissue matrix. The fourth and final component (image d) highlights the central region of adenocarcinoma with a dark red colour, but does however mark several areas within the normal connective tissue with a similar correlation.

The final panel displays images created via PCA-FCM Clustering. Images (a) to (f) were constructed by subjectively increasing the amount of clusters found by the analysis from 2 – 7 respectively. When comparing these created images directly against the known tissue type regions, the image constructed from a 4 cluster analysis appears to best mimic the histological architecture of the tissue section (image c). The diseased columnar epithelium or adenocarcinoma is characterised by the red cluster of spectra, whereas the healthy stromal tissue spectra have been grouped into the yellow cluster. Blood cells and spectra collected from areas where holes within the tissue are apparent have alternatively been partitioned into the cyan and blue clusters respectively. When the number of clusters found by the analysis were increased above this level, the stromal tissue was further partitioned into multiple groups (images d – f).

The mean spectra calculated for the adenocarcinoma, red blood cells and stromal tissue are displayed in Figure 24 (1800 – 720 cm^{-1} spectral region). The stromal tissue spectra again display strong collagen characteristics, with a triad of peaks within the amide III region (1205, 1232 and 1280 cm^{-1} respectively) and an intense band at 1450 cm^{-1} attributed to the methyl/methylene deformation mode from amino acid side chains [51]. The blood cell spectra also display a similar spectral profile.

However, notable differences can be found within the amide modes, whereby the blood cells display a larger amide II / amide I ratio and a distinct broadening of the amide II band. In contrast, a dramatic decrease in the amide II / amide I ratio is found for spectra originating from the adenocarcinoma (red spectrum), which also displays a marked reduction when compared to the healthy columnar epithelium seen in earlier experiments (Figure 13). Strong mucin bands seen previously are also not apparent in the spectrum and are replaced by a more defined nucleic acid peak found at c.a. 1080 cm^{-1}. Since these types of cells do not actively store collagen we may assume that the strong band at 1240 cm^{-1} is more likely attributable to the antisymmetric vibration of phosphate (PO_2^-) groups found in RNA and DNA.

3.3.1.3 Discussion and Conclusions

The application of FTIR spectroscopy as a diagnostic tool for cervical dysplasia has to date been limited by the high degree of heterogeneity found among exfoliated cervical cells [31,34,38,]. Only a small number of studies have been conducted that methodically assess the natural variation found within these cell types. However, these investigations were carried out upon cultured cell lines [41] that do not ideally mimic conditions within the body, and cells isolated from peripheral blood [34], which can still include a diverse variety of cervical cell types. The successful partitioning of individual cell types found within cervical smear material is hard to achieve and has been attempted previously by fractioning the cells by their relative cell size [39]. However, such a process could never be appropriate for cervical screening as diagnostic cells often range in size dependent upon their site of origin within the cervix. Thus a more appropriate method to assess spectral differences

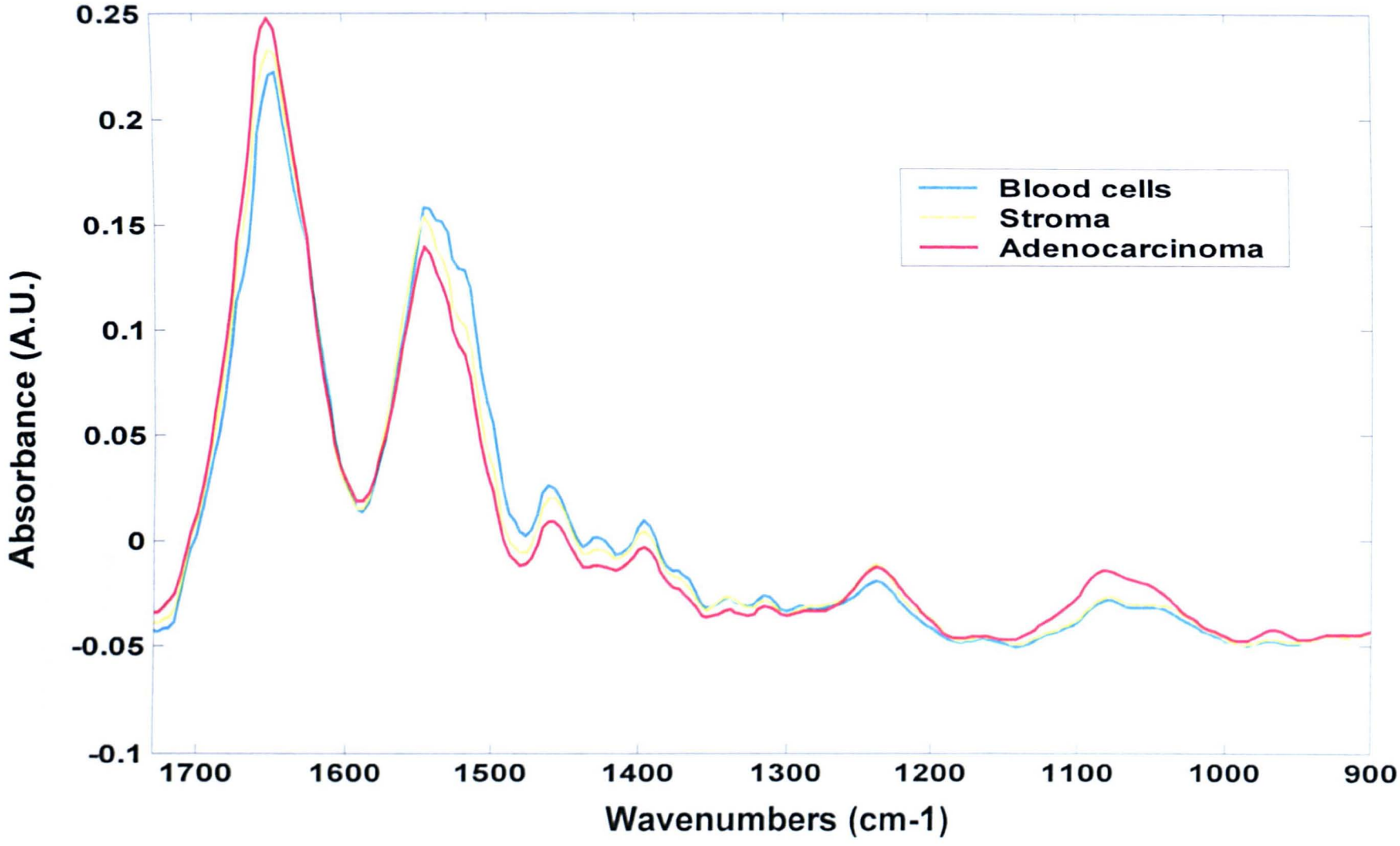

Figure 24: *4 Cluster PCA-FCM Analysis Results. Spectral window displaying mean spectra between 1800 – 720cm^{-1}. Only mean spectra calculated from tissue spectra clusters have been included. The mean spectrum of the cluster characterising the region off the tissue section was particularly noisy and clouded the spectral window.*

found between alternate cell types is to analyse tissue section material. This method enables anatomical and histological features to be easily identified and further scrutinised by FTIR spectroscopic imaging. To aid the discrimination between different tissue types found within the sections analysed, often unattainable using univariate images alone, a variety of multivariate statistical techniques have been applied and contrasted.

The construction of principal component images from the collected tissue datasets gave varied sensitivity for tissue discrimination. This type of multivariate imaging appears sensitive only to large spectral variations found between spectra and consequently provided limited tissue characterisation when compared to histological diagnosis. For example, the construction of PC images for healthy squamous epithelium (figure 9) gave rise to a number of components that enabled contrast between tissue pathology, since these spectra displayed marked spectral differences that accompanied alternate tissue pathology. However, a similar analysis carried out upon healthy squamous epithelium but taken from a diseased tissue section (figure 15), gave poor tissue discrimination as the analysis was dominated by large spectral variations occurring at pixels that lied off the tissue section. MCR imaging alternatively gave a more consistent number of factors that best described the major tissue types found within the datasets analysed. Acceptable contrast was made between the tissue pathologies present and individual factors could be assigned to individual tissue components. However, the application of PCA-FCM cluster imaging to the same datasets gave marked improvement upon tissue discrimination and allowed the major tissue types present to be partitioned into further subsets that mimicked histological characterisation more directly. By calculating mean average

spectra for the clusters produced, apparent biochemical changes between alternate tissue pathology could be directly assessed.

The adoption of a vector normalisation approach in our pre-processing routine appeared distinctly beneficial, allowing both the amide modes to be included in our multivariate analysis, which consequently proved an important region for tissue discrimination. A large degree of tissue classification was based upon variations that were identified between the intensity and positions of the amide I and amide II bands. These differences are likely to reflect changes in relative protein concentration and secondary structures. Within the healthy squamous epithelium, vibrations below 1300 cm^{-1} appeared most beneficial for tissue discrimination, glycogen content increasing with cell maturity. Spectra collected from areas of CIN displayed a distinct lack of glycogen and alternatively exhibited pronounced symmetric and antisymmetric phosphate (PO_2^-) bands at 1080 and 1240 cm^{-1}. These variations were additionally coupled with a distinct reduction in the amide II / amide I band intensity ratio. Similar spectral differences were identified within cells of adenocarcinoma, where the once prevalent mucin peaks found in healthy columnar epithelium were replaced by pronounced nucleic acid bands. Taking into account exfoliated cervical smear material is routinely collected from both these regions of the cervix, it is apparent that spectroscopic diagnosis of cell pellets would be greatly hindered by both glycogen and glycoprotein contributions to the collected spectra. These findings highlight the need for a cell preparation method that can allow analysis of single exfoliated cervical cells, if infrared spectroscopic diagnosis of such samples is to be successful.

In conclusion, IR multivariate imaging can accurately reproduce the morphological and histological architecture of the cervix, allowing both healthy and diseased tissues to be identified. With the advent of detector array systems it has become feasible to collect in excess of 10,000 individual spectra from a tissue section measuring 1 x 1 mm^2, a size close to that conventionally scrutinised via histopathology. A large tissue spectra databank could therefore be collected from alternate pathological states and enable the construction of a robust supervised pattern recognition method for automated spectroscopic diagnosis. Although we have shown that IR multivariate imaging can provide both accurate histological and biochemical information about the tissues analysed, with the distinct potential to discover earlier diseased states not identifiable via conventional histology, the time required to both collect and display the information is still a limiting factor. However, given the recent and expected rapid advance in both detector array and computer technology, a spectroscopic diagnostic tool for tissue section analysis appears plausible.

3.3.2 Multivariate Analysis of IR Imaging Results from Additional Cervical Tissue Sections

During this study multiple cervical tissue sections were analysed to help assess whether inter patient natural variation could produce markedly different spectral characteristics and thus render a spectroscopic method of diagnosis unattainable. However, the mean cluster spectra calculated from different patients with similar histopathology appear very similar. These results confirm previous findings by both Lasch *et al.* [54] and Wood *et al.* [53], who reported smaller patient to patient

variations than those observed for different tissue types and histological diagnoses. In this section I will display multivariate imaging results obtained from multiple cervical tissue sections. Only results gained via PCA-FCM cluster analysis will be discussed.

3.3.2.1 Cervical Tissue Section C19154

The first tissue section in our library, named C19154, was cut from a benign cervical biopsy and incorporated the entire transformation zone. This enabled a map to be collected from the entire region and characterise cells originating from both the ecto and endo cervix. A white light image of the entire tissue section and the region chosen for analysis are shown in figures 25a and 25d. Photomicrographs collected from the parallel H&E stained section for the same regions are displayed in figures 25b – c respectively. These allow the main tissue types to be visualised, which include a nabothian follicle surrounded by columnar epithelium, the endocervical canal of the cervix lined with columnar epithelium, squamous epithelium and the underlying connective tissues. By use of a step size and aperture of 25 μm, a total of 11,305 individual IR spectra were collected from a spatial area of 2375 x 2975 μm. The multivariate imaging results produced for this dataset are shown in Figures 26 and 27.

Figure 27 displays false colour images created via PCA-FCM Clustering and the H&E stained image from the same region to allow direct comparison. Images (b) to (j) were constructed by subjectively increasing the amount of clusters found by the analysis from 2 – 10 respectively. When comparing these constructed images

directly against the known tissue type regions, the image constructed from a 9 cluster analysis appears to best mimic the histological architecture of the tissue section (image i). All major tissue types are characterised within this image allowing the squamous epithelium (maroon), columnar epithelium that line nabothian follicles (cyan), columnar epithelium of the endocervical canal (yellow and light blue) and connective tissue (orange) to be identified. However, further subsets of spectra have been identified for the connective tissue. Cells that surround the columnar epithelium that line the nabothian follicles have been partitioned into a single group (dark cyan), those cells that directly underlie both types of epithelium and are highly nucleated (red), and finally those cells existing deeper within the cervix (blue). The outside region where no tissue exists has been characterised by the green cluster of spectra. The differentiation of tissue types was based on variations found within the band positions, intensities and half widths of the amide modes and the low frequency region of the collected spectra (< 1200 cm^{-1}). Columnar epithelium was dominated by contributions from glycoproteins (mucin), with those cells lining the endocervical canal displaying markedly increased concentrations. Squamous epithelium was alternatively dominated by glycogen contributions, displaying the characteristic glycogen triplet of peaks. Connective tissues were discriminated via their increased collagen contributions to the spectrum, and further subdivided by their nucleic acid concentrations and amide II / amide I peak intensity ratio. However, the analysis did struggle in separating the basal cell spectra of the squamous epithelium from the columnar epithelium that lined the nabothian follicles, even when cluster numbers were increased above this level. But these two types of cell do share very similar spectral profiles as reported previously, and their individual characterisation does not adversely affect the diagnostic information gained from the analysis.

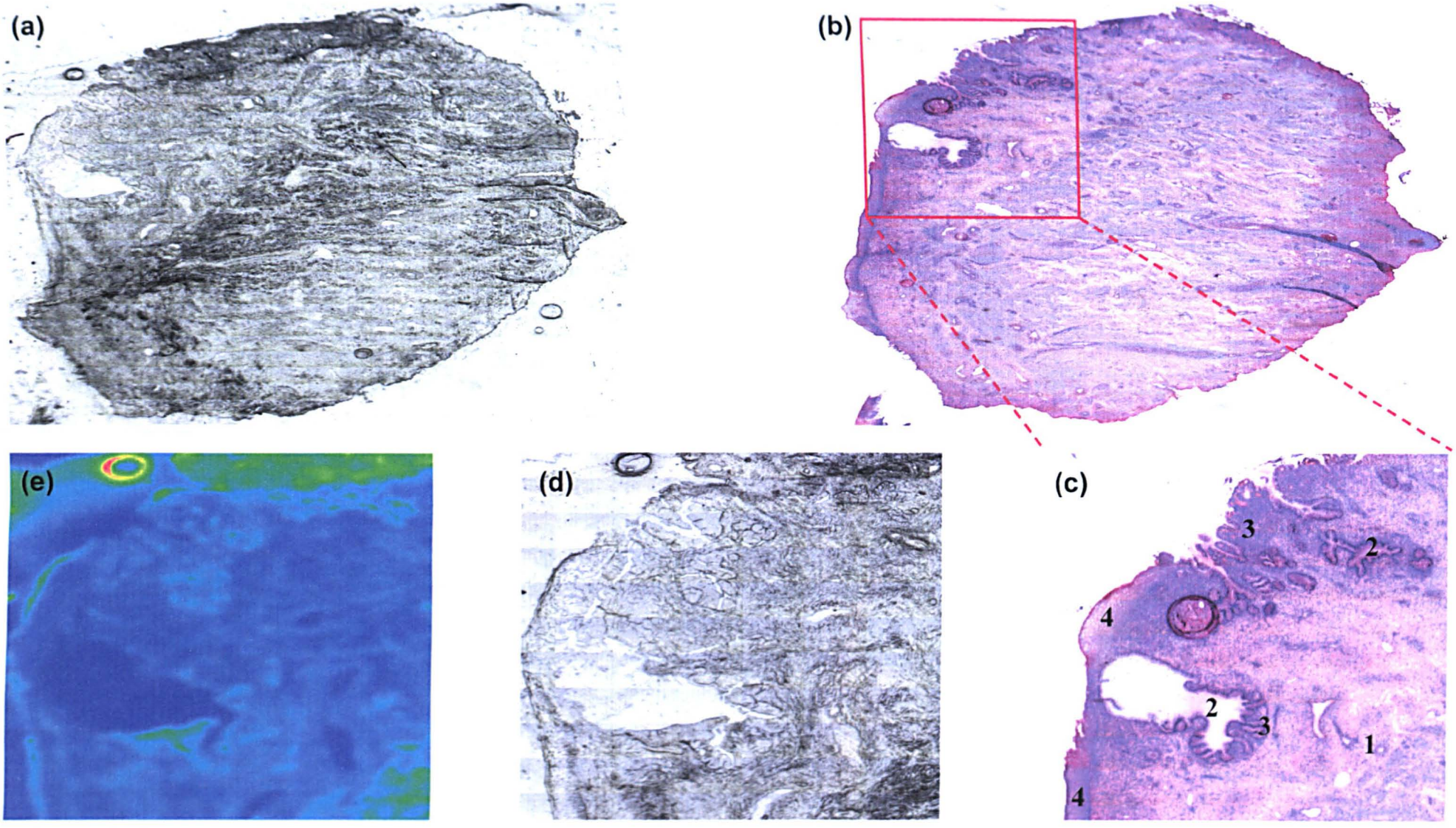

Figure 25: *a) White light image of entire tissue section. b) H&E photomicrograph of entire parallel tissue section. c) H&E photomicrograph of mapped region displaying benign anatomical features. (1) Connective or stromal tissues, (2) nabothian follicles, (3) columnar epithelium and (4) squamous epithelium. d) White light image of same region upon analysed tissue section. e) IR imaged area (2375 x 2975 μm) mapped using a step size and aperture of 25μm for a total 11,305 individual IR spectra.*

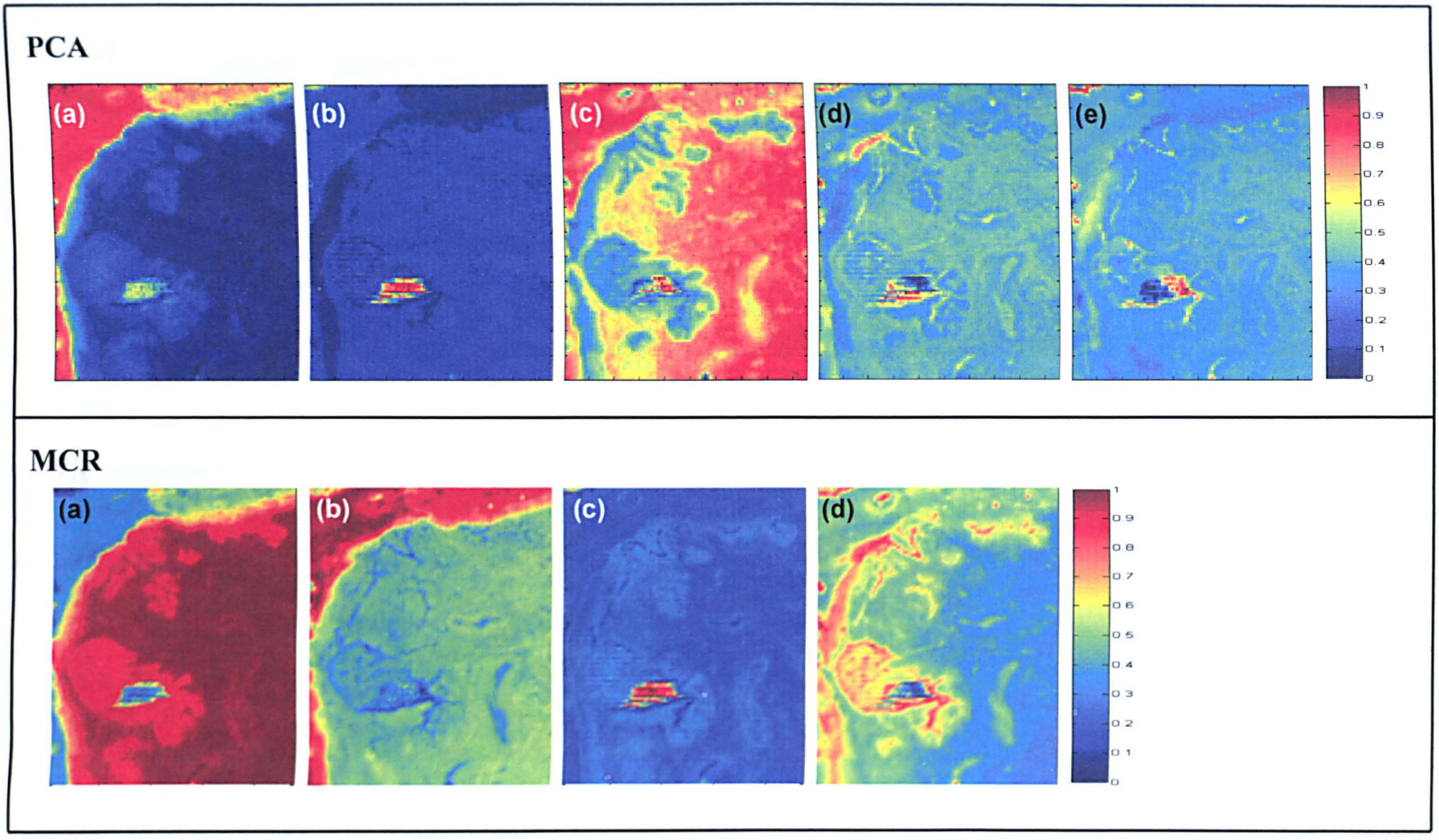

Figure 26: *Multivariate Imaging results from benign cervical tissue section C19154. PCA Panel: (a) – (e) False colour weighted images for PC's 1 – 5 respectively. Colour scale ranges from red indicating spectra that are very similar to that PC, and blue which are greatly dissimilar. MCR Panel: (a) – (d) False colour weighted images created from a 4 component MCR analysis. Colour scale ranges from red indicating spectra that are very similar to that component, and blue which are greatly dissimilar.*

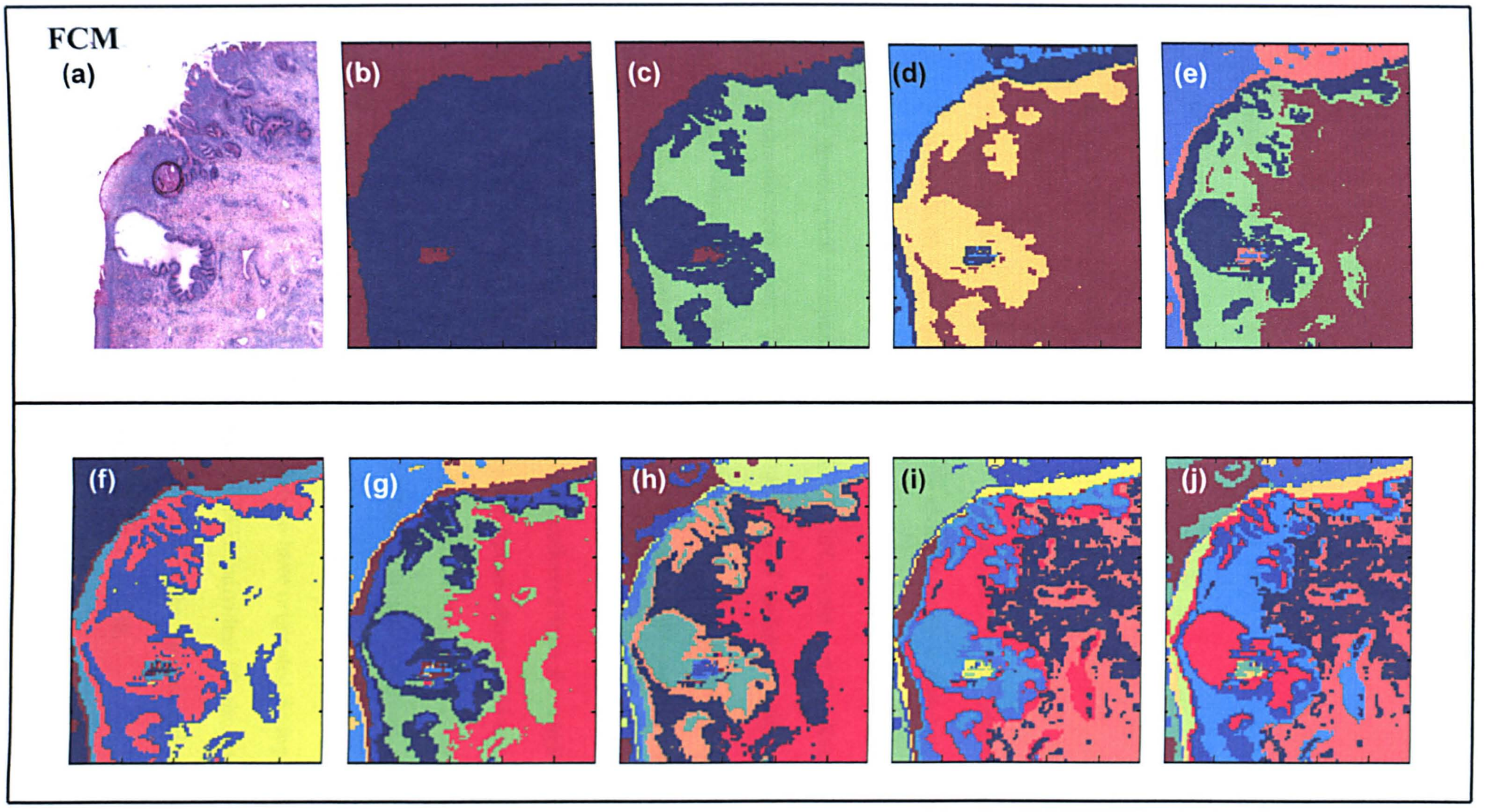

Figure 27: *PCA-FCM Imaging results from benign cervical tissue section C19154.FCM Panel: (a) – (j) False colour images created using FCM clustering analysis results. Note cluster numbers were subjectively increased from 2 – 10. Pixels with the same colour in each image are spectra that were partitioned into the same cluster.*

Moderate to large concentrations of glycogen were found within the squamous epithelium, which was coupled to the absence of metabolically active cells that commonly show reduced amide II / amide I ratios and pronounced nucleic acid bands. Therefore spectroscopic analysis of the tissue section can verify the absence of malignant cells.

3.3.2.2 Cervical Tissue Section C22727

The second tissue section in our library, named C22727, was collected and prepared from a patient who had revealed CIN characteristics within her previous smear screenings. However, upon analysis of the parallel stained tissue section that was cut for histological comparison, regions of CIN could not be located within the squamous epithelium. Although the absence of malignant cells was unfortunate, spectroscopic analysis could be used to assess the squamous epithelium and possibly identify pre-malignant changes that may accompany the onset of disease. A white light image of the entire tissue section and the region chosen for analysis are shown in figures 28a and 28c. Photomicrographs collected from the parallel H&E stained section for the same regions are displayed in figures 28b and 28d respectively. These allow the main tissue types to be visualised, which include the squamous epithelium and the underlying connective tissue. When additionally examining the white light image of the region chosen for spectroscopic analysis (figure 28c), a fold within the tissue section is also revealed that is likely to have originated during sectioning. By use of a pixel size of 6.25 μm a total of 8439 individual IR spectra were collected from a spatial area of 543.75 x 606.25 μm. The multivariate imaging results produced for this dataset are shown in Figures 29 and 30.

Figure 30 displays false colour images created via PCA-FCM Clustering and the H&E stained image from the same region to allow direct comparison. Images (b) to (j) were constructed by subjectively increasing the amount of clusters found by the analysis from 2 – 10 respectively. When comparing these constructed images directly against the known tissue type regions, the image constructed from a 9 cluster analysis appears to best mimic the histological architecture of the tissue section (image i). All major tissue types are characterised within this image allowing the squamous epithelium (royal blue), connective tissues (orange and green) and area of tissue folding (light blue) to be revealed. Within this analysis, the superficial squamous epithelium appeared to be contaminated by cervical mucus, displaying increased mucin bands as the layers approached the edge of the tissue (dark blue, maroon, yellow and red clusters respectively). Spectra collected from areas upon the mapped region where no tissue was apparent were partitioned into the cyan cluster and displayed characteristics of pure cervical mucus. These findings highlight a possible drawback with frozen sectioning, all sections analysed to some extent revealing a surrounding cervical mucus layer. However, this was the only section that revealed the distinct contamination of areas deep within the tissue section. Problematic pixels that surround the tissue section and obtain strong characteristic mucin bands could feasibly in the future be identified and systematically removed from the analysis. The subsets of connective tissue spectra were separated via their amide II / amide I peak intensity ratio, the tissues directly underlying the squamous epithelium displaying a reduced value (green). However, the most important finding of the analysis was the distinct lack of glycogen found within the histologically diagnosed healthy squamous epithelium. These cells instead displayed a slightly reduced amide II / amide I peak intensity ratio when compared to glycogenated

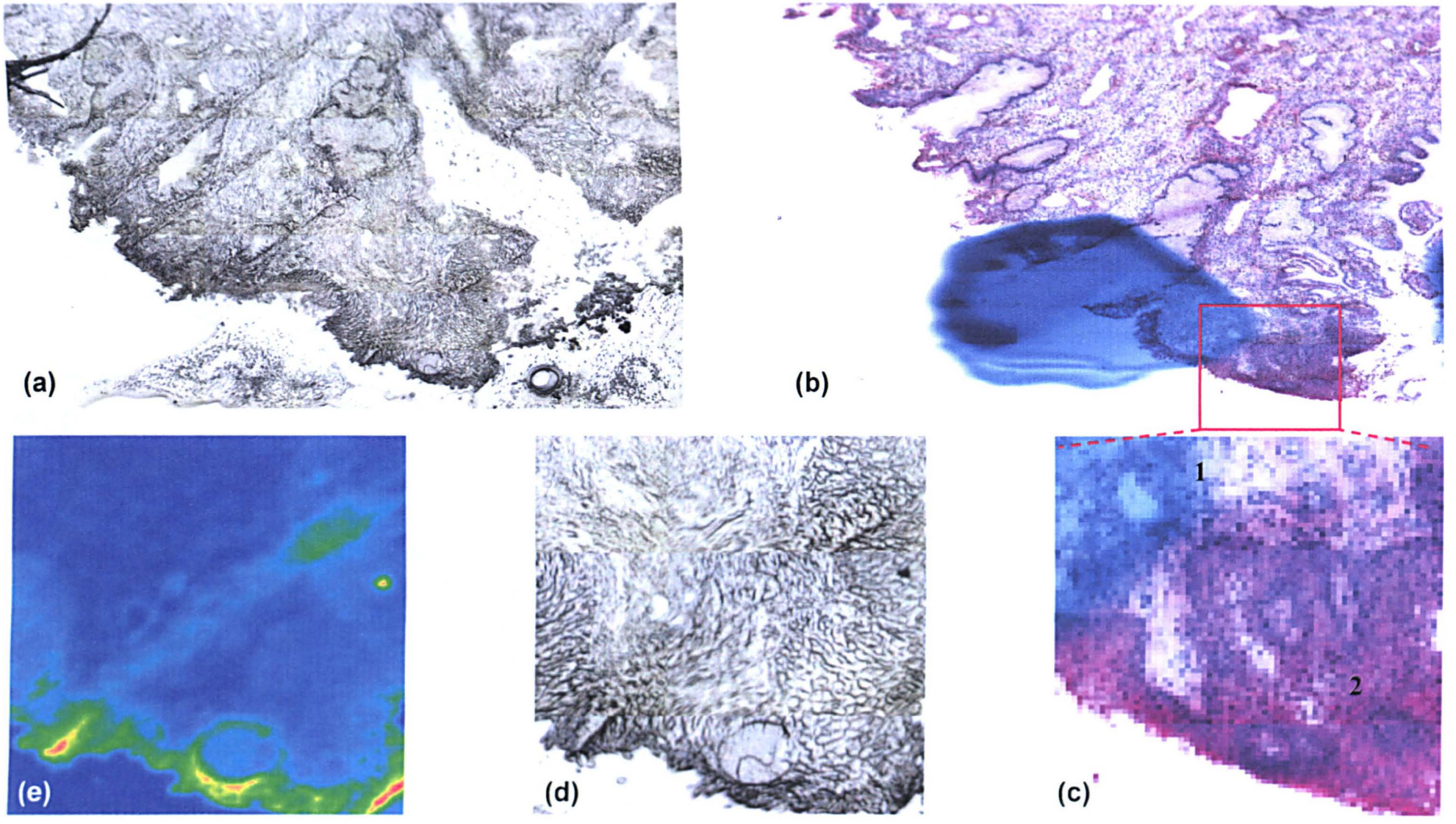

Figure 28: *a) White light image of entire tissue section. b) H&E photomicrograph of entire parallel tissue section. c) H&E photomicrograph of mapped region displaying malign anatomical features. (1) Connective or stromal tissues and (2) squamous epithelium. d) White light image of same region upon analysed tissue section. e) IR imaged area (543.75 x 606.25 μm) mapped using a pixel size of 6.25 μm for a total of 8439 individual IR spectra.*

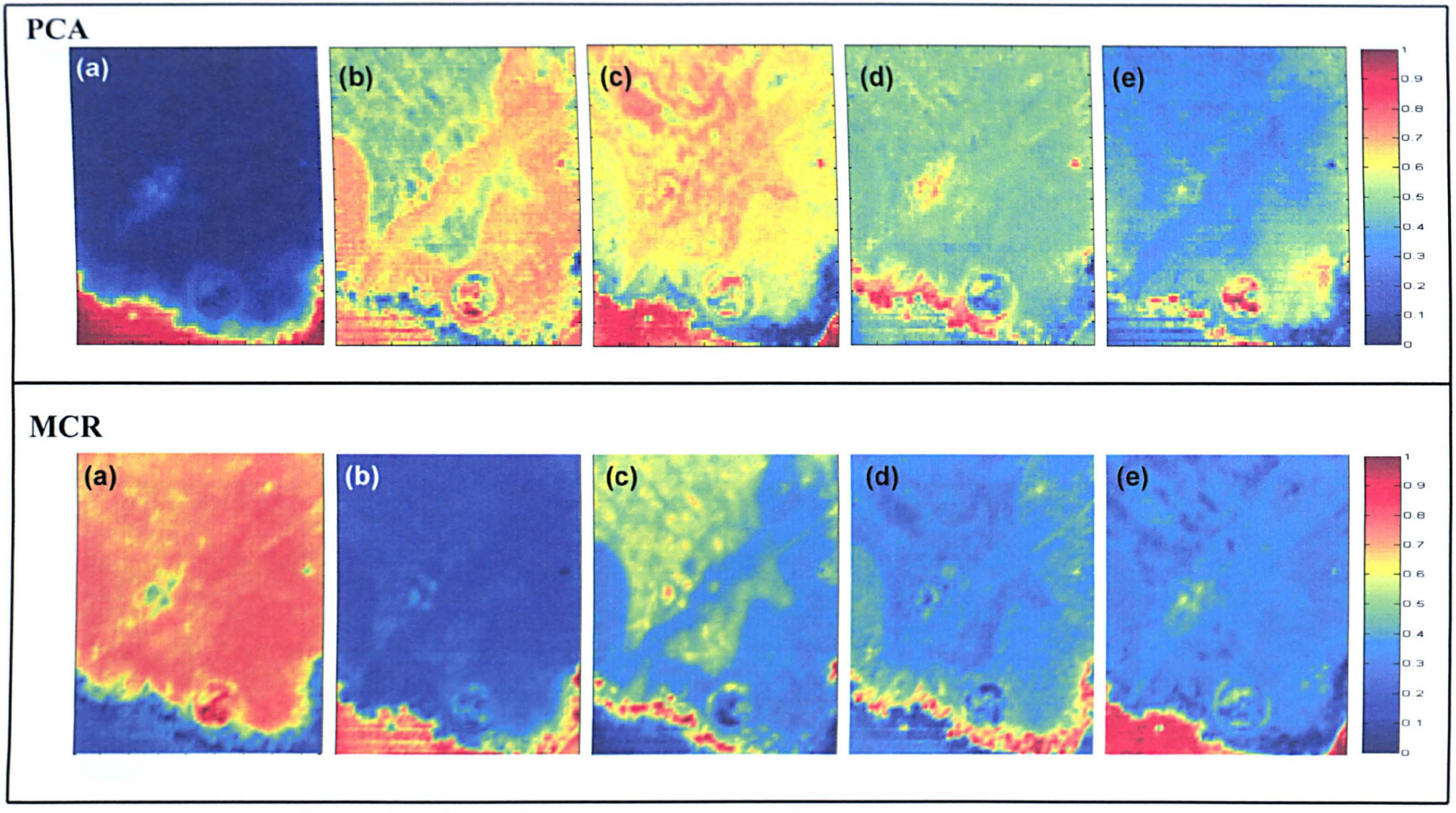

Figure 29: *Multivariate Imaging results from benign cervical tissue section C22727. PCA Panel: (a) – (e) False colour weighted images for PC's 1 – 5 respectively. Colour scale ranges from red indicating spectra that are very similar to that PC, and blue which are greatly dissimilar. MCR Panel: (a) – (e) False colour weighted images created from a 5 component MCR analysis. Colour scale ranges from red indicating spectra that are very similar to that component, and blue which are greatly dissimilar.*

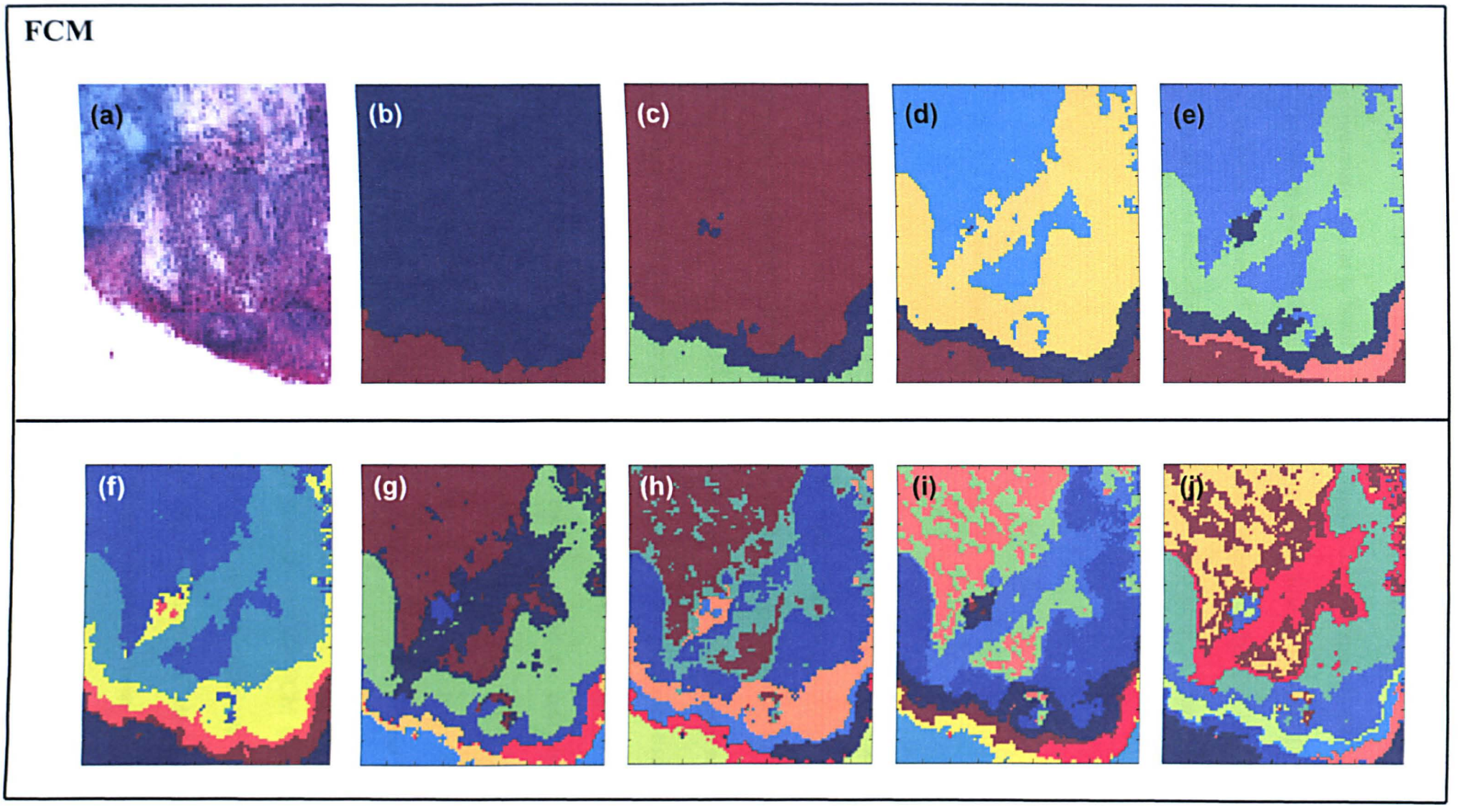

Figure 30: *PCA-FCM Imaging results from benign cervical tissue section C22727. FCM Panel: (a) – (j) False colour images created using FCM clustering analysis results. Note cluster numbers were subjectively increased from 2 – 10. Pixels with the same colour in each image are spectra that were partitioned into the same cluster.*

squamous cells, with symmetric and antisymmetric phosphate bands appearing more prevalent similar to basal cell spectra. These identified spectral differences were also observed within healthy squamous tissue lying close to regions of CIN in previous analyses (figures 15 and 16). Thus spectroscopic analysis of such tissue sections could allow the detection of small but distinct biochemical changes that accompany earlier stages of malignant change.

3.3.2.3 Cervical Tissue Section C19490

The third tissue section in our library, named C19490, was collected and prepared from a patient who had revealed no abnormal characteristics within her previous smear screenings. White light images of the entire tissue section and the region chosen for analysis are shown in figures 31a - b. Unfortunately no H&E stained images were made available for this tissue section, but the major tissue components can be visualised via small light intensity differences. These include both the mature and immature layers of the squamous epithelium and the underlying connective tissues. By use of a step size and aperture of 25μm, a total of 8103 individual IR spectra were collected from a spatial area of 2775 x 1825 μm. The multivariate imaging results produced for this dataset are shown in Figures 32 and 33.

Figure 33 displays false colour images created via PCA-FCM Clustering and the white light image from the same region to allow direct comparison. Images (b) to (j) were constructed by subjectively increasing the amount of clusters found by the analysis from 2 – 10 respectively. When comparing these constructed images

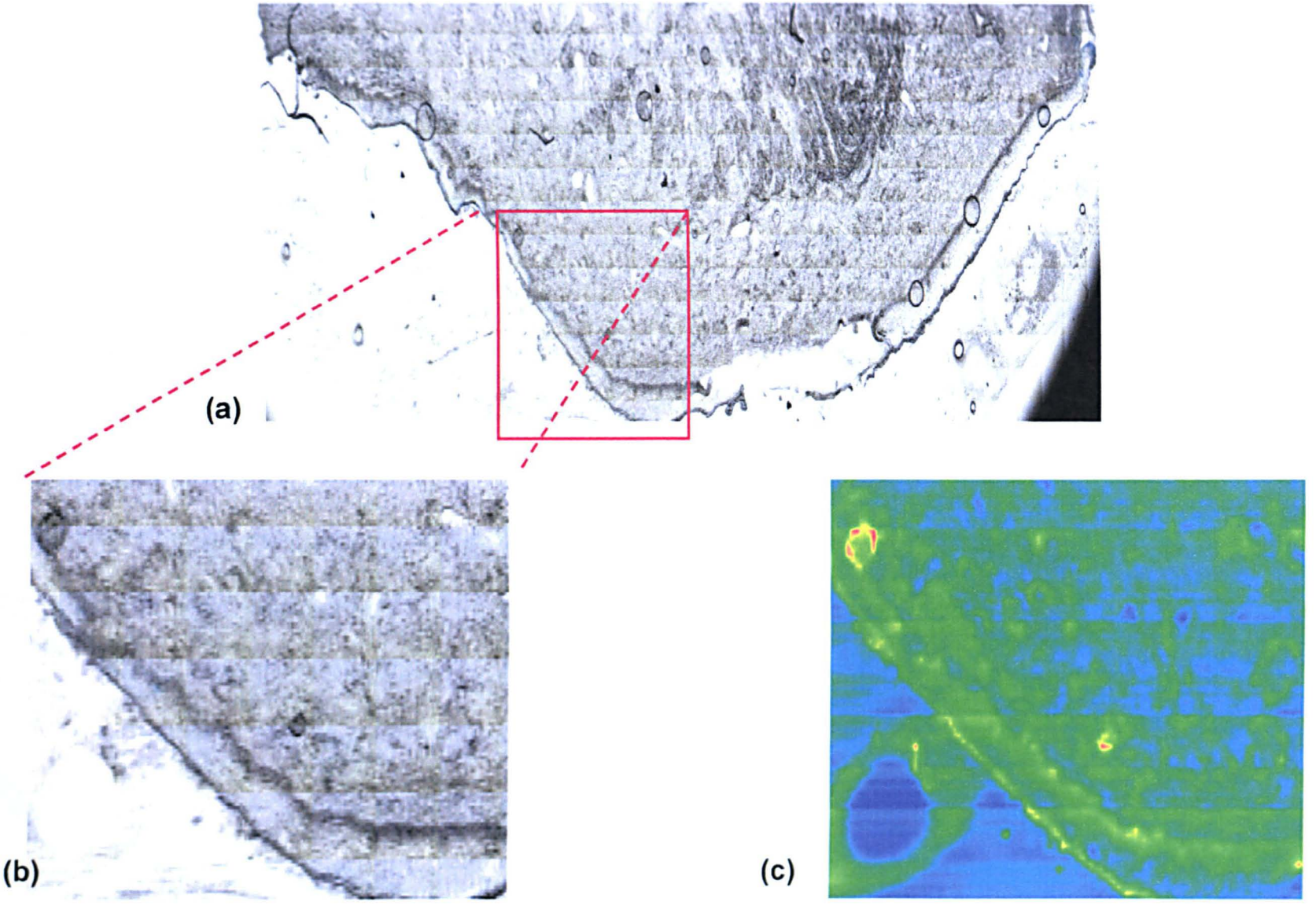

Figure 31: *a) White light image displaying region of squamous epithelium. b) White light image of mapped area displaying benign anatomical features. (1) Connective or stromal tissues and (2) squamous epithelium. c) Total absorbance image of mapped area (2775 x 1825 µm) using a step size and aperture of 25µm for a total 8103 individual IR spectra.*

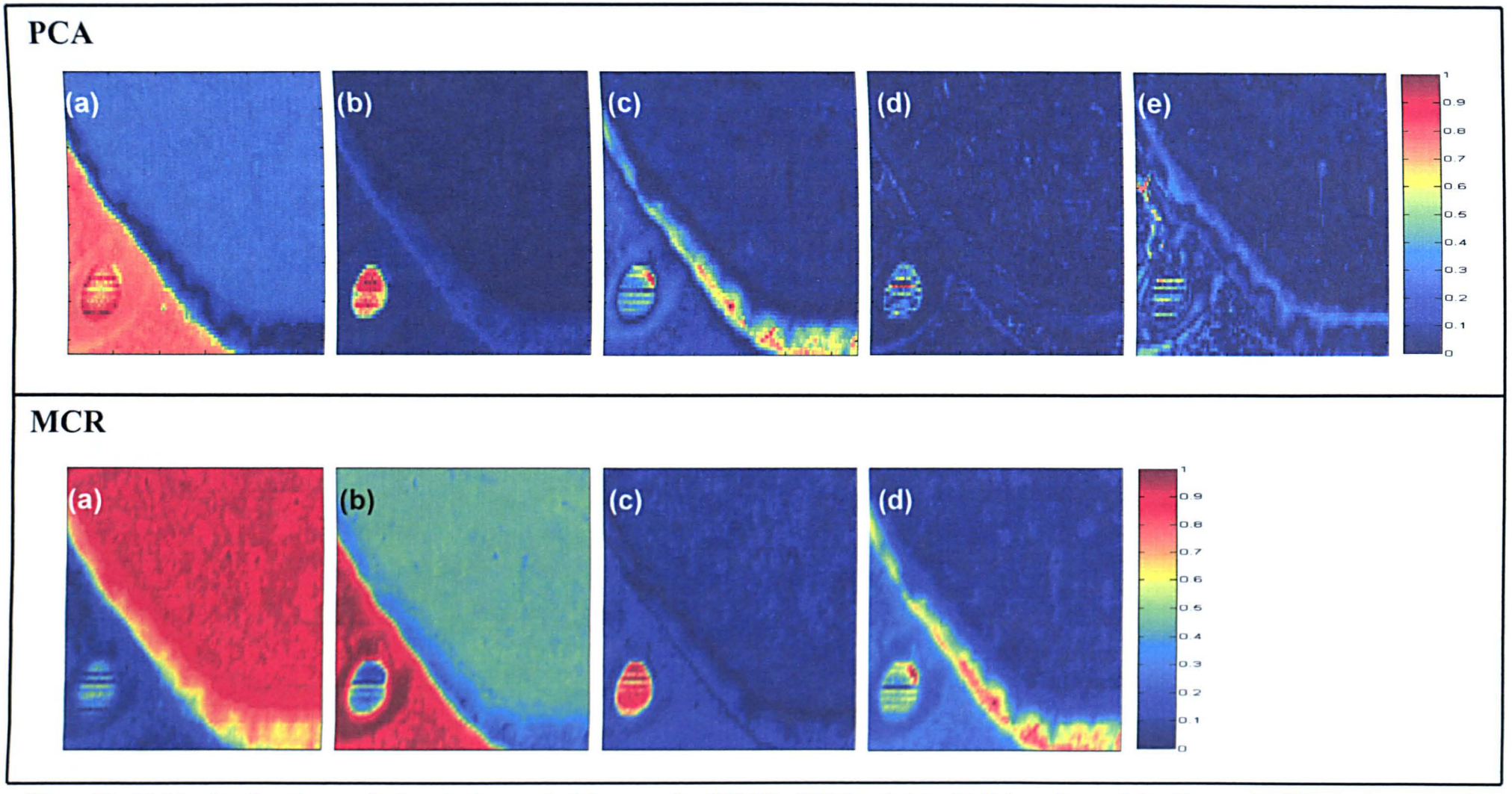

Figure 32: *Multivariate Imaging results from benign cervical tissue section C19490. PCA Panel: (a) – (e) False colour weighted images for PC's 1 – 5 respectively. Colour scale ranges from red indicating spectra that are very similar to that PC, and blue which are greatly dissimilar. MCR Panel: (a) – (d) False colour weighted images created from a 4 component MCR analysis. Colour scale ranges from red indicating spectra that are very similar to that component, and blue which are greatly dissimilar.*

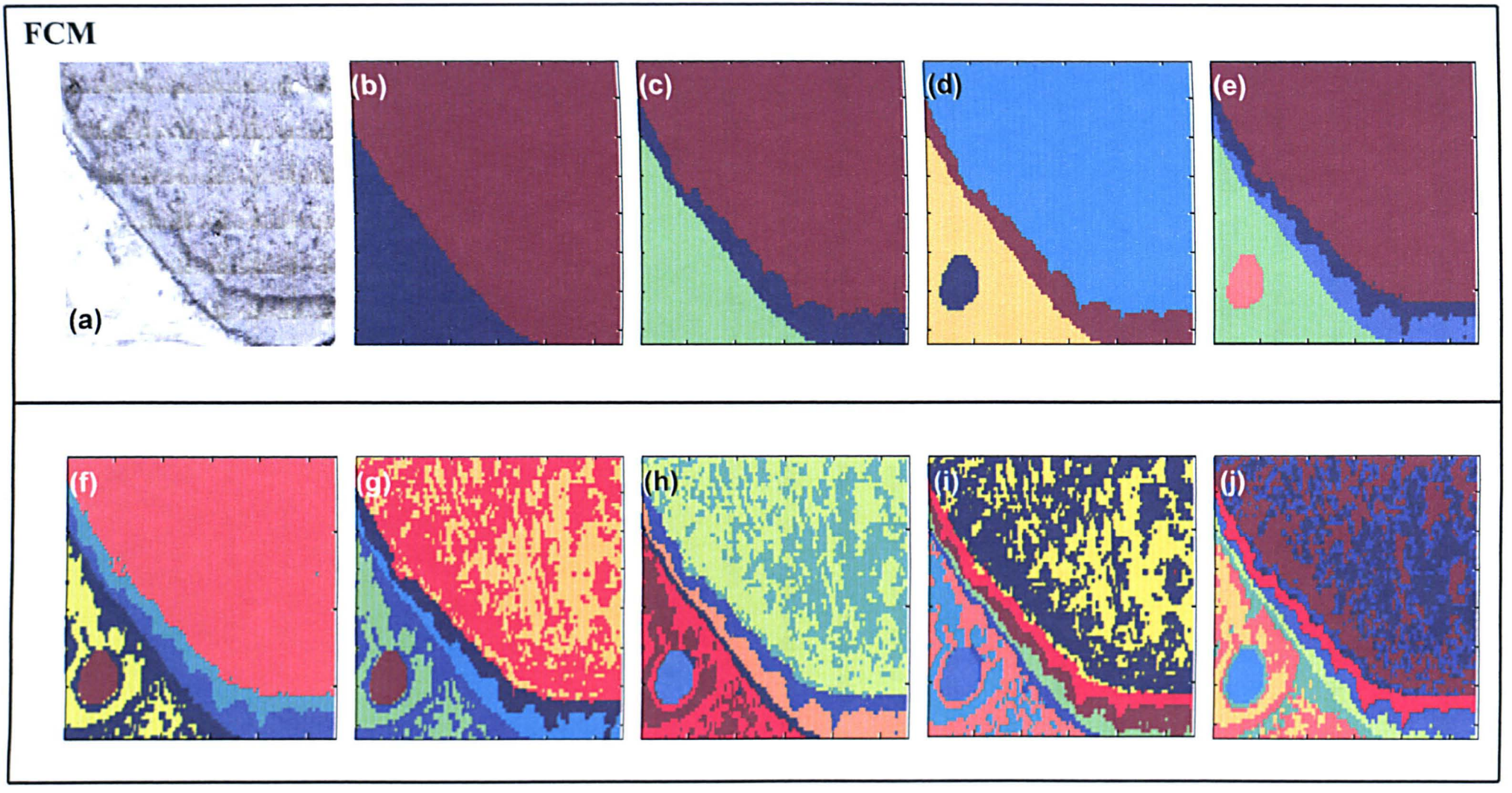

Figure 33: *PCA-FCM Imaging results from benign cervical tissue section C19490. FCM Panel: (a) – (j) False colour images created using FCM clustering analysis results. Note cluster numbers were subjectively increased from 2 – 10. Pixels with the same colour in each image are spectra that were partitioned into the same cluster.*

directly against the known tissue type regions, the image constructed from a 7 cluster analysis appears to best mimic the histological architecture of the tissue section (image g). All major tissue types are characterised within this image allowing the superficial (cyan) and basal (blue) layers of the squamous epithelium to be easily recognised. These were partitioned via the presence or absence of glycogen peaks within the tissue spectra. Connective tissues were partitioned via their increased collagen contributions and subdivided dependent upon their amide II / amide I ratio, indicative of protein abundance and secondary structure changes. The remaining clusters identified spectra that were collected from regions off the tissue but lie on areas rich with (light blue and green) or absent of (maroon) cervical mucus. In conclusion, large concentrations of glycogen were found within the squamous epithelium with no spectral differences identified that could suggest cancerous change. Thus in this case, spectroscopic analysis could confirm the absence of malignant or pre-malignant biochemical changes to the cells of the squamous epithelium.

3.3.3 IR Microscopic Analysis of Individual Exfoliated Cervical Cells by use of a Synchrotron Source

Early experiments undertaken during this study focused toward the collection of IR spectra from individual exfoliated squamous epithelial cervical cells. Conventional smear samples were collected at a specialist colposcopy clinic (Derby City General Hospital) and prepared onto reflective substrates via liquid based cytology (LBC) methods. Such techniques produce a monolayer cellular presentation upon the substrates and thus enable the collection of IR spectra from individual cells. However, the recommended preservative solution used in this process, PreservCyt®, produced inconsistent spectral artefacts that were problematic for reliable cell characterisation. A splitting of the amide I band was evident in some spectra, a finding inconsistent with parallel microscopic studies upon cervical tissue sections (figure 7, section 4.1.3.1). An alternative preservative solution, 70% ethanol, was thus adopted having gained support in earlier literature [33]. This solution did not produce the spectral artefacts observed previously with PreservCyt® and still provided cellular presentation that was suitable for spectroscopic analysis of individual cells (figure 8, section 4.1.3.2). All results presented in this section were collected from exfoliated cervical cells prepared in this manor.

Cervical smear samples collected in this study originated from two different categories of patient prognosis. The first group of patients were classed as being low risk having shown no previous abnormal smears. In contrast, the second group of patients had shown mild dysplasia (C.I.N I) in previous smear screenings and were classed as being high risk. Prepared samples were examined at the Daresbury SRS

Laboratory, where up to 100 spectra were collected from individual cells on each slide. A co-ordinate reference system was utilised to later relocate and diagnose the cells examined via conventional cytological PAP staining. Unfortunately all cells examined during the analyses were later classed by cytology as being normal in nature, showing no dysplastic characteristics. However, natural biochemical variation of squamous epithelial cells could be assessed. At this point in our study we were unable to scrutinise the collected spectra via MCR and FCM multivariate analyses since we were still developing our algorithms and their applications at that time. Therefore multivariate analysis was undertaken by use of Pirouette®, a proprietary piece of software developed by Infometrix (Woodinville, W.A., USA). This enabled the collected spectra to be analysed via Hierarchical Cluster Analysis (HCA), a multivariate method of data analysis described in full in section 4.5.4. Spectra collected from the same slide were compiled into single datasets and analysed collectively by this technique. All datasets examined via HCA displayed similar results and partitioning of spectra. An example dendrogram from one such analysis is shown in Figure 34.

Examining the agglomerative dendrogram produced by the analysis (Figure 34), it appears the collected spectra have been partitioned into two main groups that are negatively correlated to each other (similarity index is below 0). By "cutting" the dendrogram at this point, we reveal two main clusters of spectra that are now highlighted by red and pink colours respectively. By classing the data in such a way, all spectra can now be plotted into a spectral window and coloured according to their cluster membership (Figure 35).

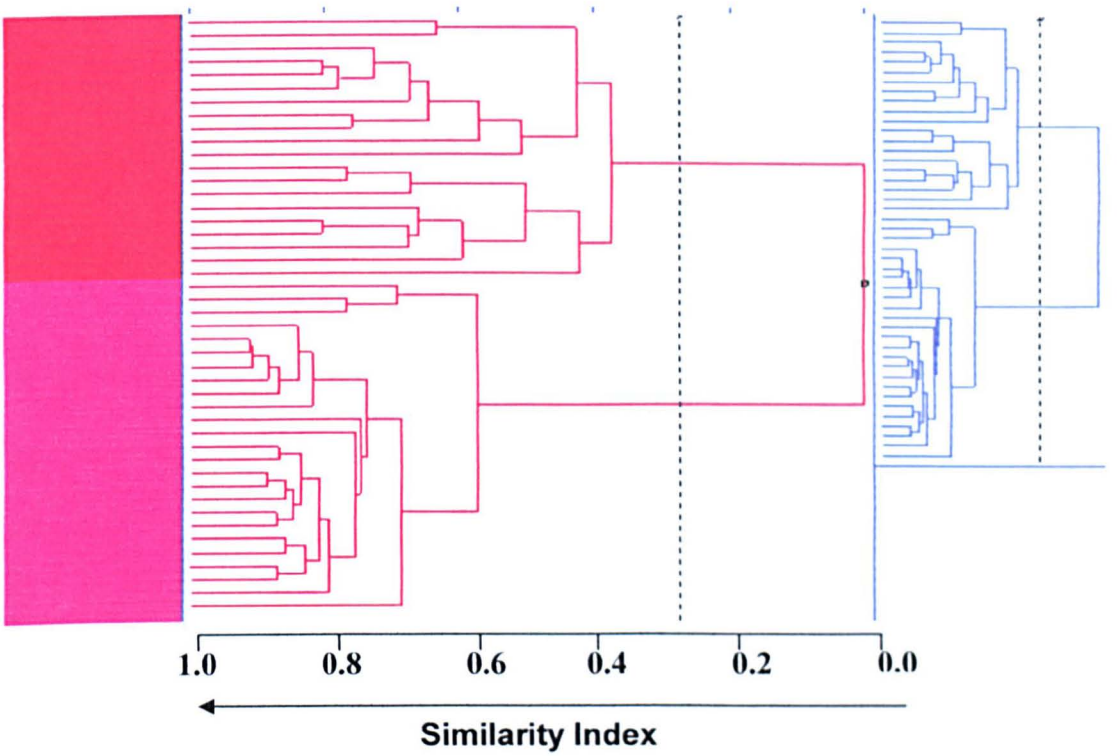

Figure 34: *HCA Dendrogram of the spectral dataset collected from a high risk patient (Cervical slide IR03-0008). The analysis was restricted to only include values in the spectral range 1800 – 900 cm^{-1}. Data pre-processing included amide II peak maxima normalisation and mean centring. Clustering was achieved by use of a Euclidean distance metric and a group average linkage method.*

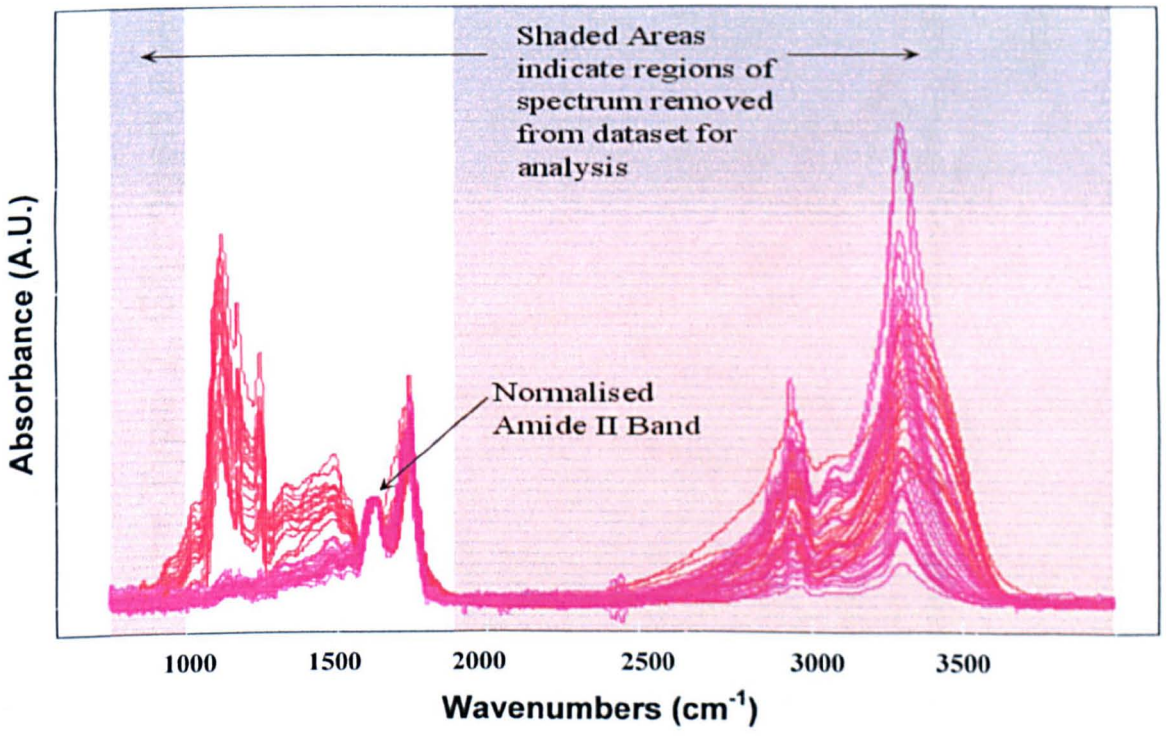

Figure 35: *Spectral window displaying the two main clusters formed via HCA . The red cluster represents spectra that are glycogen rich, whereas the pink cluster characterises glycogen absent spectra.*

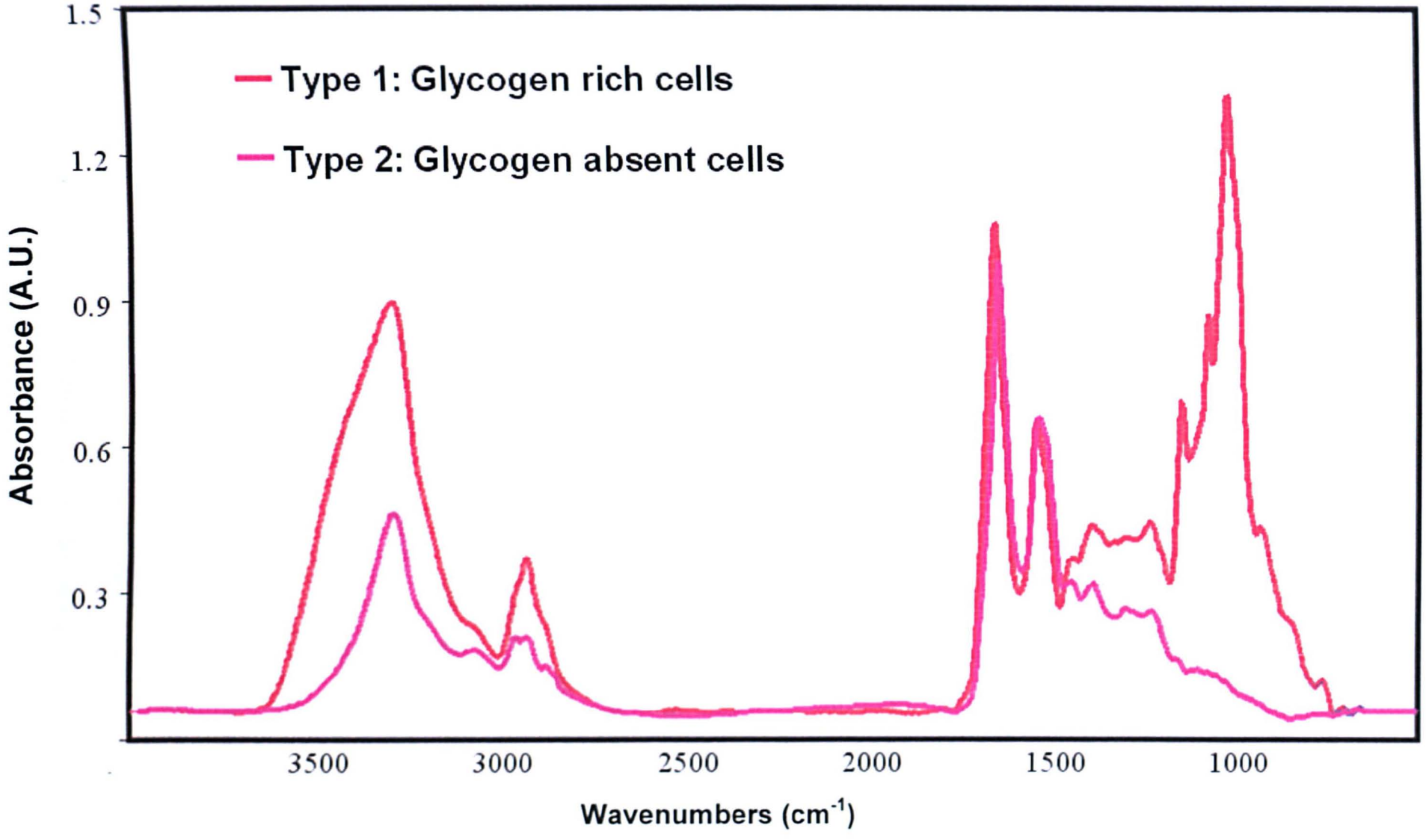

Figure 36: *Spectral window displaying the mean spectra calculated from the two main clusters partitioned via HCA . The red cluster represents spectra that are glycogen rich, whereas the pink cluster characterises glycogen absent spectra.*

By plotting the spectra in such a way, spectral similarities apparent for data held within the same cluster, and dissimilarities with those partitioned into different clusters can be identified. When studying the spectra more closely, it is apparent they have been partitioned by spectral differences occurring below 1200 cm^{-1}. To aid further characterisation of the two distinct spectral profiles, mean average spectra for each cluster were calculated and are shown in Figure 36.

The first cluster of spectra, which I have named Type 1, is clearly defined by a triplet of peaks commonly associated with glycogen. These display peak maxima at ~ 1150, 1075 and 1020 cm^{-1} and correspond to the C-O stretch, C-C stretch and C-O-H deformation modes respectively. A significant broadening of the OH stretching region is also apparent for all Type 1 spectra, which most likely reflects an increased degree of intermolecular hydrogen bonding in the cells. A large variation in the peak position of this band was noticeable between spectra and may have been influenced by contributions from atmospheric water vapour. However, great care was taken to eliminate such contributions by use of a purge ring that flushed the sample area with dry air. Therefore, it is more likely that this band shift is caused by an increased level of glycogen in the spectrum. The CH stretching region of the spectrum is also poorly resolved and only allows the antisymmetric methyl stretches to be identified, again likely to be caused by high levels of glycogen. When resolution was enhanced in this region, the relative peak intensities of the methyl and methylene stretches were observed to be very similar. This would suggest that these types of cells are not significantly contributed to by lipids, where it is more common to find an increased intensity for the methylene band [55]. The remainder of the spectrum is dominated by protein contributions with pronounced amide I and amide II bands found at 1648

and 1540 cm^{-1} respectively. The symmetric and antisymmetric bending modes associated with the methyl groups of protein side chains (1450 – 1350 cm^{-1}) could also be identified. However, when glycogen band intensities were observed to be larger than those recorded for the amide bands, the antisymmetric stretch became almost irresolvable, indicating glycogen absorbance in this region. Finally, in the majority of Type 1 spectra, an additional peak at 935 cm^{-1} was also observed. This band has only been documented in the literature by Wood *et al.* [33,56], who was unable to provide a peak assignment. However, this peak is likely to be attributed to glycogen after a similar band was observed in a study of pure glycogen [57].

The second cluster of spectra, which I have named Type 2, is clearly differentiated by the disappearance of the glycogen triplet. The distinct absence of glycogen within the spectrum provides a clearer region below 1300 cm^{-1} that allows both the symmetric and antisymmetric phosphate bands (PO_2^-) associated with nucleic acids to be identified (1080 and 1240 cm^{-1} respectively). However, both band intensities are relatively weak and are likely to reflect a small contribution to the spectra from the nuclei of the cells. There is a clear sharpening of the OH band and an appearance of a shoulder peak that can only be speculated as the N-H stretches of proteins. The CH stretching region is also much more resolved, with four definable bands observed in most spectra, attributed to the symmetric and antisymmetric methyl and methylene modes respectively. When comparing these observations with those seen previously in Type 1 spectra, it would indicate that glycogen normally contributes strongly in all these regions of the spectrum. The remainder of the spectrum is again dominated by protein bands very similar to those of Type 1 spectra, with only small differences found between the amide II / amide I band intensity ratio.

When directly comparing the spectral datasets collected from low risk and high risk patients, only a small difference in the percentage of Type 1 to Type 2 spectra were observed, with high risk patients displaying a small but increased number of Type 2 spectra. Similar spectral profiles were observed in early macroscopic studies of cervical smear cell pellets [28,29] that correlated the disappearance of glycogen bands with the onset of cervical dysplasia. However, the results from our study upon single healthy squamous epithelial cells would indicate that such assumptions, although provocative, may have reached incorrect conclusions. A much larger natural variation in the cellular composition of these cells was observed, with a majority of the spectra dominated by glycogen contributions. When directly comparing the two spectral profiles for Type 1 and Type 2 spectra, it is apparent that glycogen can mask spectral features in many regions of the spectrum, especially below 1200 cm^{-1}, rendering this region inadequate for diagnostic purposes. The original location of these cells within the squamous epithelium is also likely to have contributed to the spectral changes we have observed. Both our own parallel studies upon cervical tissue sections, and those of other groups [38,53], have correlated distinct spectral changes that couple healthy squamous cell maturation. Thus glycogen absent spectra presented in this study, although unlikely, may have been collected from cells incorrectly sampled from the basal layer of the squamous epithelium, where glycogen contributions are negligible. However, glycogen content within squamous cells has also been shown to differ greatly throughout the menstrual cycle and can be reduced in women taking monophasic oral contraception [46]. In conclusion, the collection of IR spectra from individual healthy squamous epithelial cells has proved that a distinctly larger amount of natural biochemical variation is

occurring within these cells than originally thought. Spectral features of bands below 1200 cm^{-1} are notably masked by contributions from glycogen rendering this region inadequate for diagnostic purposes alone. Therefore, additional spectral differences observed above this region for dysplastic or malignant cells would need to be used in combination with glycogen absence for an effective diagnostic marker.

3.3.4 Novel Experiments whereby FTIR Microscopic Maps were collected from Exfoliated Cervical Cells and have been analysed by Multivariate Imaging

After our early point mapping experiments upon single exfoliated cells it became apparent that the abundance and appearance of abnormal cells was often negligible and sparsely orientated around the sample spot. Several attempts were made to screen the unstained samples before spectroscopic analysis and thus identify abnormal cells. By use of a green light filter that extenuated the nuclei of the cells, a histopathologist microscopically scrutinised the samples and chose cells he thought appeared abnormal. However, after staining it became evident that these attempts had failed with only healthy squamous cells again being identified. Therefore, in our second phase of experiments we focused toward the collection of large infrared maps in an attempt to increase the cell number and sample area analysed. Regions upon the sample spot were randomly chosen and areas between 1000 – 5000 μm^2 were mapped, incorporating vast numbers of individual cells ($\leq$ 10,000). In a compromise to achieve good signal to noise spectra with reasonable collection times ($\leq$ 6hrs), 16

interferograms per pixel (25 x 25 μm sample area) were coadded over the spectral range 4000 – 720 cm^{-1}. If we take into consideration our earlier findings upon single exfoliated cells and parallel studies upon cervical tissue sections, it is likely similar methods of data collection upon smear samples will be necessary for spectroscopic diagnosis. This method enables the collection of individual spectra from cells that are free from contaminating artefacts associated with macroscopic studies, and could allow the spectroscopic classification of all cell types present in the sample.

In this section I will report results taken from one particularly interesting map collected from an abnormal smear. In contrast to early cervical smear experiments, this sample was collected from a patient who had displayed high grade dysplasia in previous screenings. A white light image of the prepared sample spot is displayed in Figure 37a, and a magnified image of the area that was mapped is shown in 37b. A small amount of clumping among cells is noticeable, but in general individual cells can be visualised. The infrared map collected from this area consisted of 6400 individual IR spectra and was collected from an area of 2000 x 2000 μm. After data collection was complete the sample was conventionally PAP stained, and the region analysed by IR diagnosed by a histopathologist. Stained images taken from the same area are shown in Figures 38 – 41, which further characterise the main cell types present. Glycogen rich and glycogen absent healthy squamous cells have been encircled by blue and yellow colours respectively. In contrast, squamous cells diagnosed as showing low grade and high grade characteristics of dysplasia have been encircled by the green and red colours respectively. Inflammatory cells or polymorths are also present in the sample and have been encircled by a black colour.

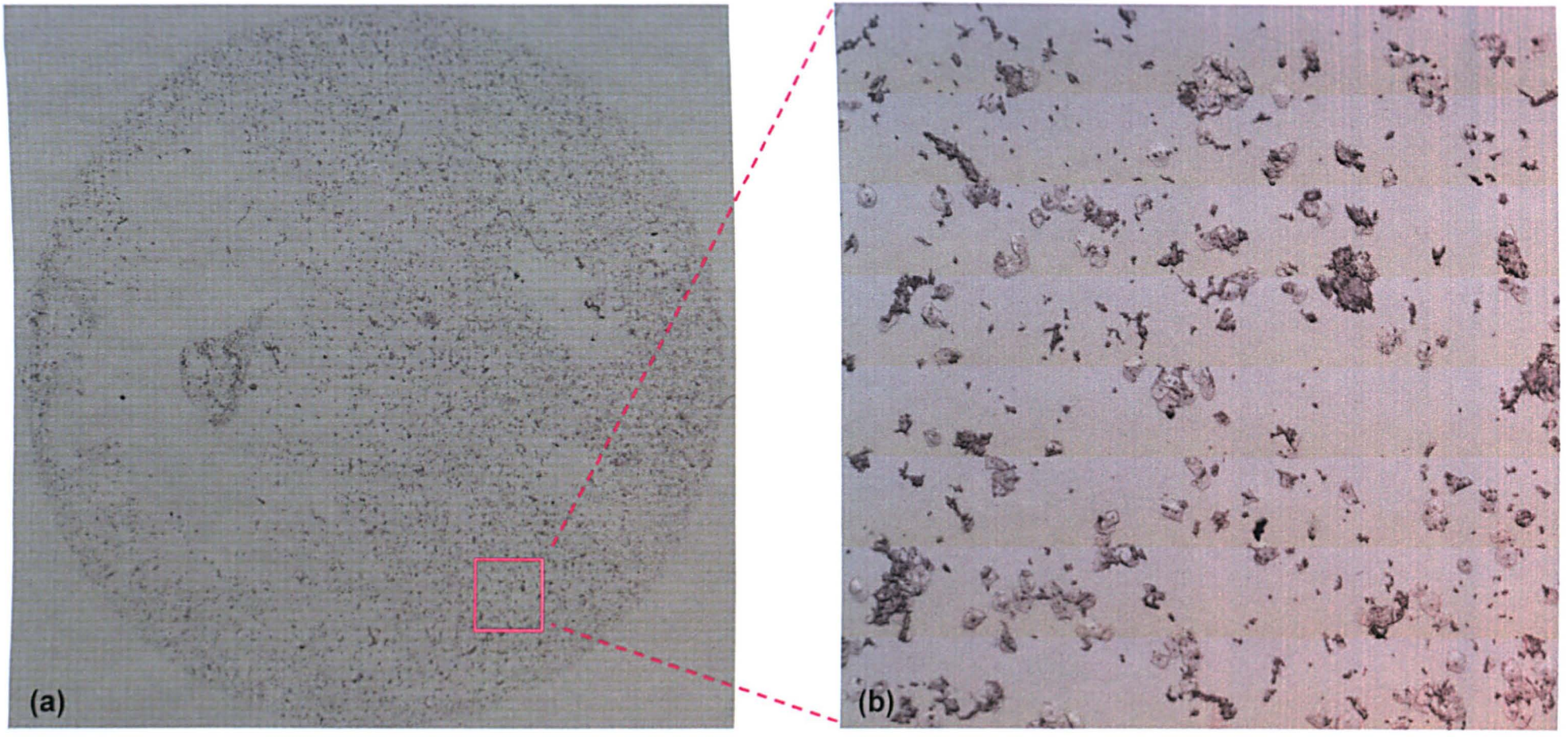

Figure 37: *Abnormal cervical smear displaying cells diagnosed as having high grade dysplastic changes. (a) White light image collected from the entire sample spot. (b) White light image collected from the IR mapped area.*

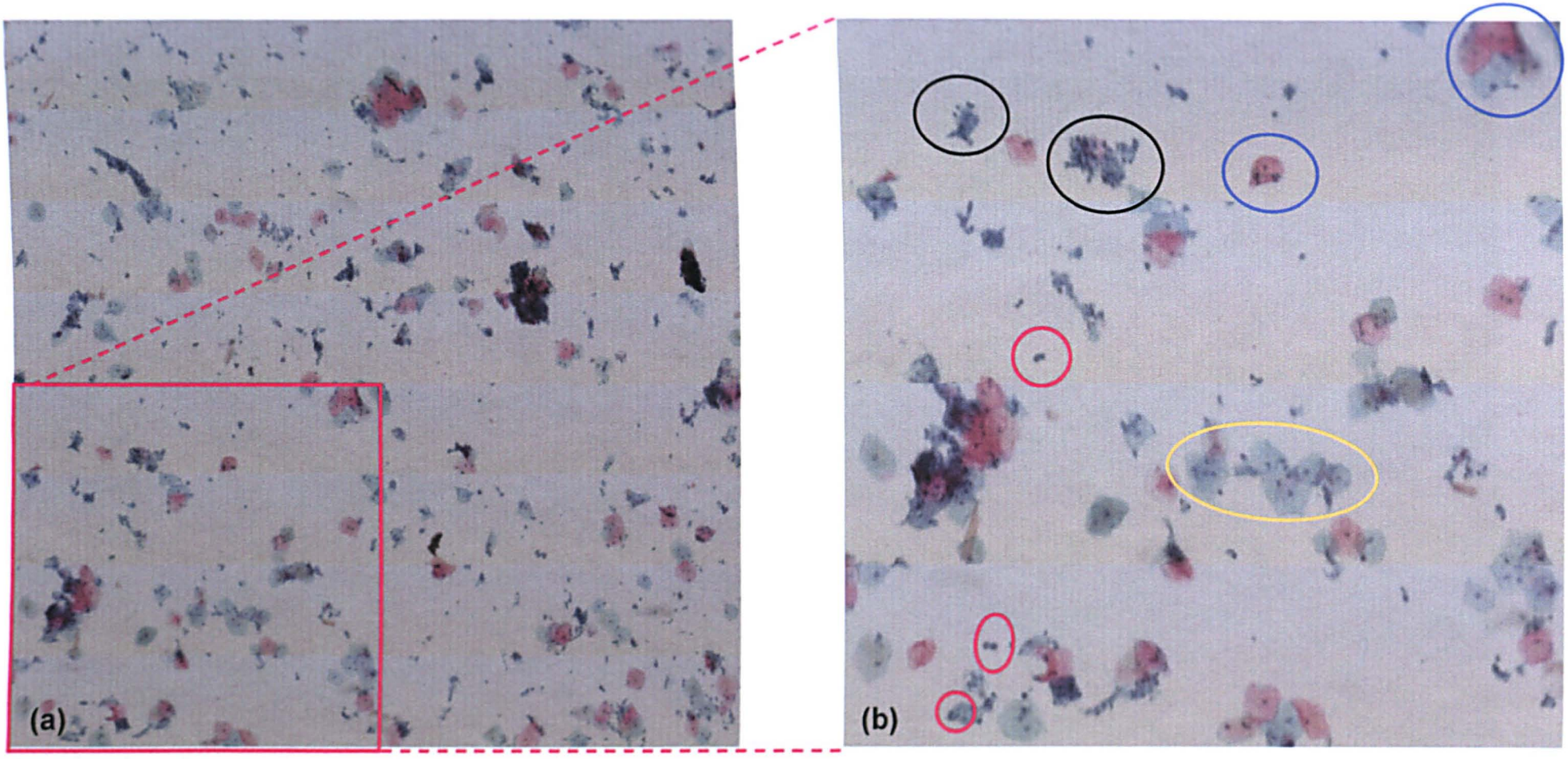

Figure 38: *Abnormal cervical smear displaying cells diagnosed as having high grade dysplastic changes. (a) PAP stained image collected from the IR mapped area. (b) Magnified PAP stained image taken from the bottom left region of the area examined via IR mapping. Healthy glycogen rich (blue), healthy glycogen absent (yellow), inflammatory (black) and high grade dysplastic (red) squamous cells can be visualised.*

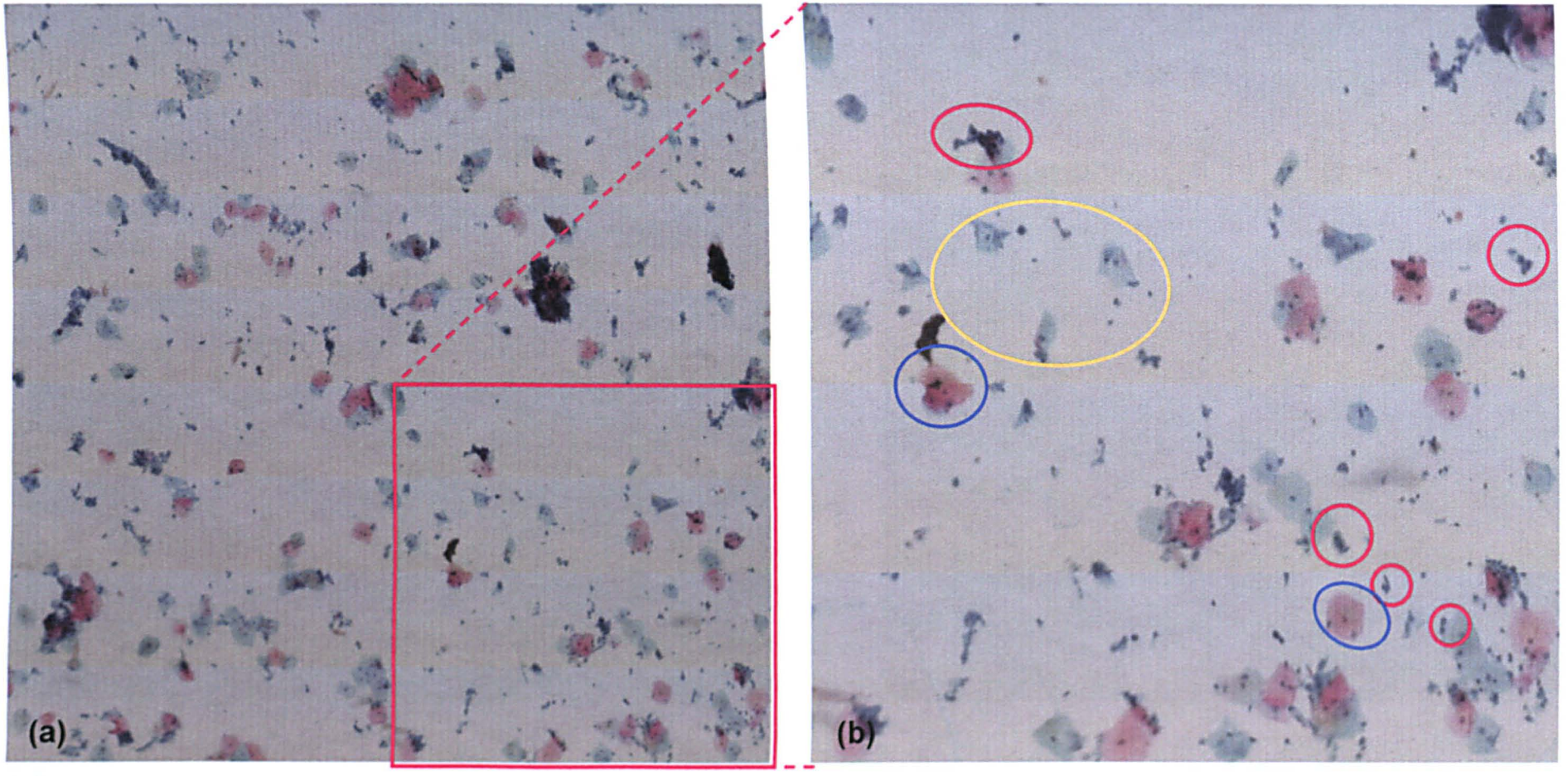

Figure 39: *Abnormal cervical smear displaying cells diagnosed as having high grade dysplastic changes. (a) PAP stained image collected from the IR mapped area. (b) Magnified PAP stained image taken from the bottom right region of the area examined via IR mapping. Healthy glycogen rich (blue), healthy glycogen absent (yellow) and high grade dysplastic (red) squamous cells can be visualised.*

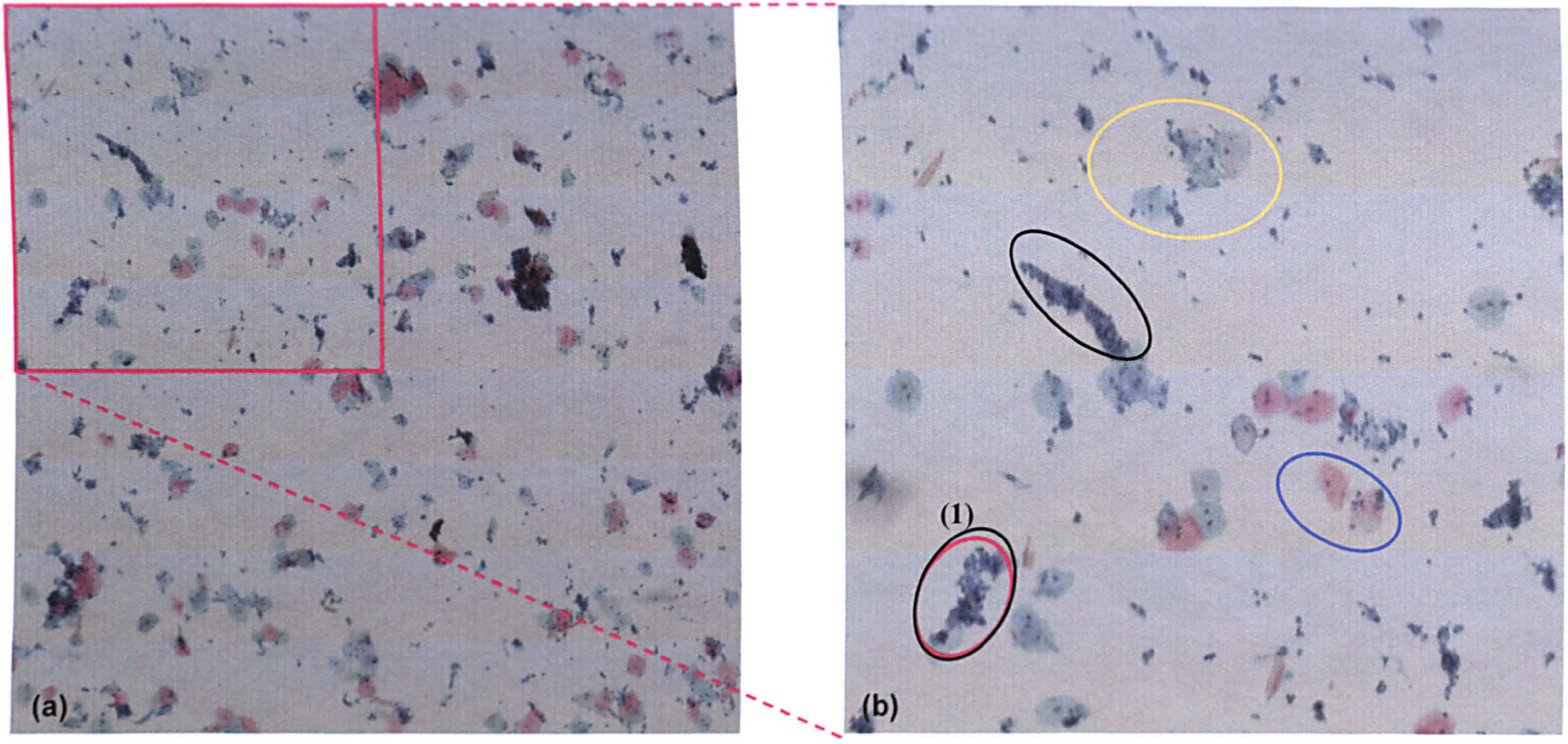

Figure 40: *Abnormal cervical smear displaying cells diagnosed as having high grade dysplastic changes. (a) PAP stained image collected from the IR mapped area. (b) Magnified Pap stained image taken from the top left region of the area examined via IR mapping. Healthy glycogen rich (blue), healthy glycogen absent (yellow), inflammatory (black) and high grade dysplastic (red) squamous cells can be visualised. The region labelled (1) displays a group of contrasting cell types that lie in close proximity or have been clumped. These include both inflammatory (black) and high grade dysplastic (red) squamous cells.*

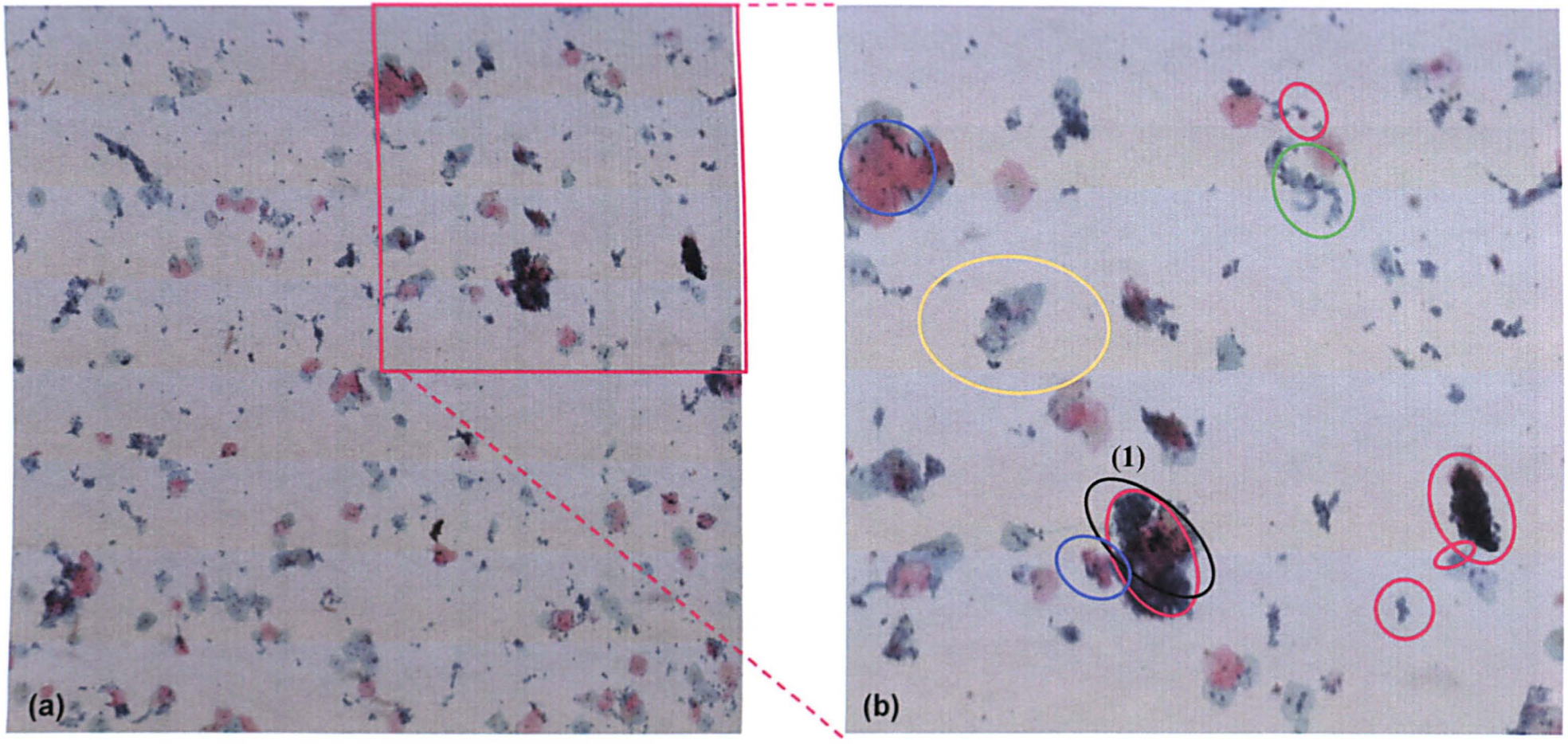

Figure 41: *Abnormal cervical smear displaying cells diagnosed as having high grade dysplastic changes. (a) PAP stained image collected from the IR mapped area. (b) Magnified PAP stained image taken from the bottom left region of the area examined via IR mapping. Healthy glycogen rich (blue), healthy glycogen absent (yellow), inflammatory (black), low grade dysplastic (green) and high grade dysplastic (red) squamous cells can be visualised. The region labelled (1) displays a group of contrasting cell types that lie in close proximity or have been clumped. These include both inflammatory (black) and high grade dysplastic (red) squamous cells.*

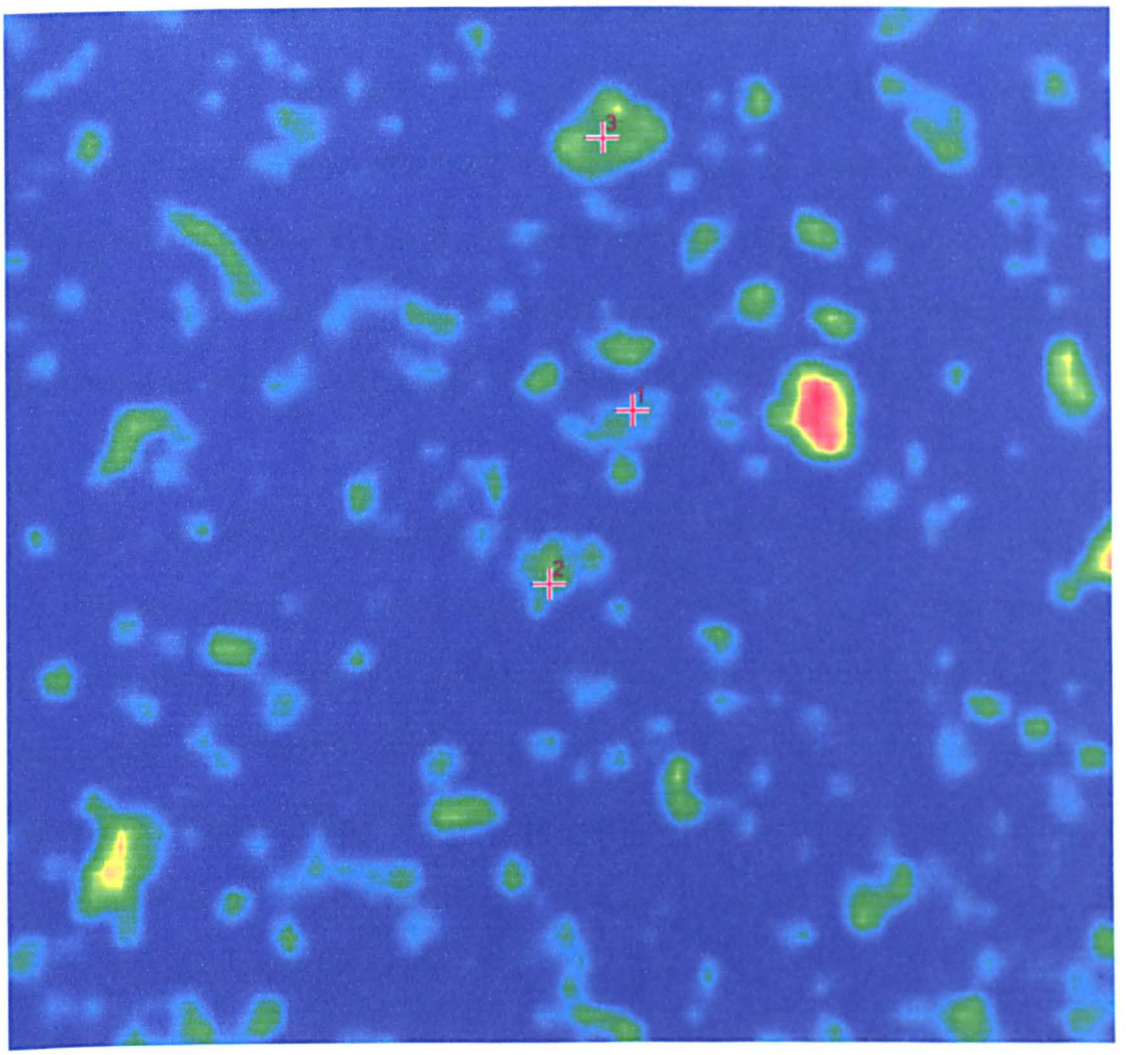

Figure 42: *Total absorbance image of the collected IR map (4000 – 720 cm^{-1}). The red coloured cross hairs labelled 1 – 3 indicate the original locations of the black, blue and red spectra in Figure 43 respectively.*

When initially scrutinising the data contained within the map, it became apparent that some spectra displayed a broad spectral feature centred at ~2000 cm^{-1} between the amide I and C-H stretching region. Three example spectra exhibiting this feature are displayed in Figure 43. A total absorbance image for the collected map is shown in Figure 42 and further allows the co-ordinates of the extracted spectra to be located. The red coloured cross hairs labelled 1 – 3 in Figure 42 indicate the original locations of the black, blue and red spectra in Figure 43 respectively. Similar baseline distorting features have been observed in single cell spectra collected from human

oral mucosa cells [58]. It is now believed that Mie-type scattering from tightly packed nuclei are responsible for these broad, undulating features, which are superimposed upon the spectral absorption features of the cells [58]. Scattering curves were calculated for spherical particles that mimicked the nucleus size of the cells examined. These were then subtracted from spectra that exhibited such undulating features to reveal reasonably straight baselines.

Since these distorting baseline features were not consistent for all spectra in our map, possibly associated with a change in nucleus size between cell spectra, correction via scattering curve subtraction is not feasible. If an incorrect subtraction was made it would most likely introduce another artefact into the spectrum. To avoid baseline distortions adversely affecting subsequent multivariate analysis, spectra were cut to only include data found between 1800 – 720 cm^{-1}, since this region of the spectrum did not appear as badly effected.

The spectral dataset was only subjected to PCA-FCM cluster analysis, since this technique had provided the best characterisation of cell types found within tissue section maps. A variety of data pre-processing routines were applied to assess their effects upon subsequent clustering and classification of cell spectra. Our first clustering experiments utilised our conventional data pre-processing technique, whereby all spectra are baseline corrected and subsequently vector normalised. However, on this occasion a 2 base point linear interpolation was applied to all spectra between 1800 and 720cm^{-1}. By requesting the analysis partition spectra into two clusters, it was hoped that background and cellular pixels contained in the map would be grouped into definable clusters. Unfortunately this was not the case and

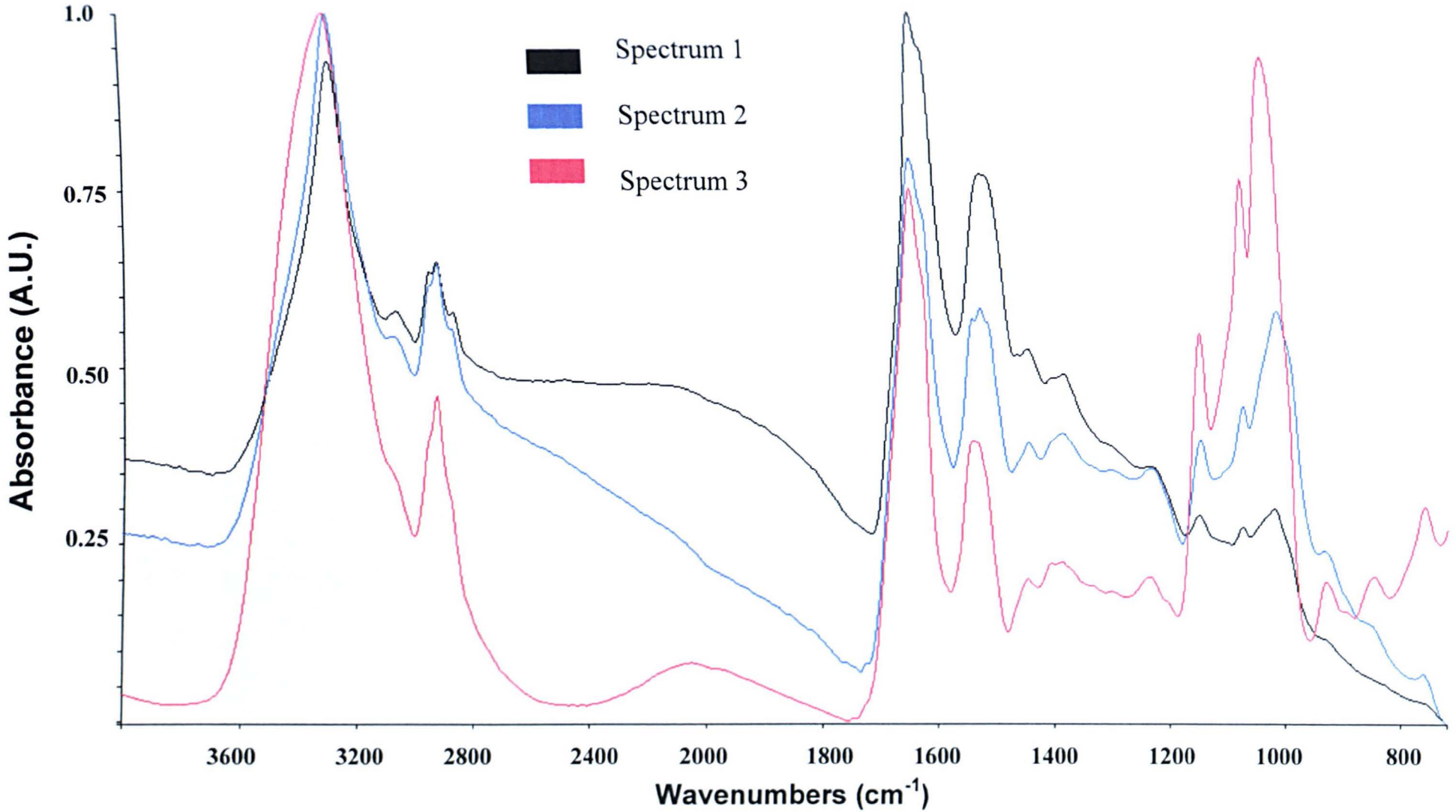

Figure 43: *Spectra extracted from collected map that display spurious baseline properties. Note the broad spectral feature centred at ~2000 cm^{-1}. The co-ordinates from which these spectra were extracted are indicated in Figure 42.*

cluster membership appeared confused. The cluster image constructed from this analysis is shown in Figure 44b. By directly comparing the cluster image to the total absorbance image in Figure 44a, we can clearly see that background and cellular spectra have not been partitioned into separate clusters as hoped. When the cluster number established by the analysis was increased, in the expectation that a multiple number of clusters may represent the background pixels, images only became more confused.

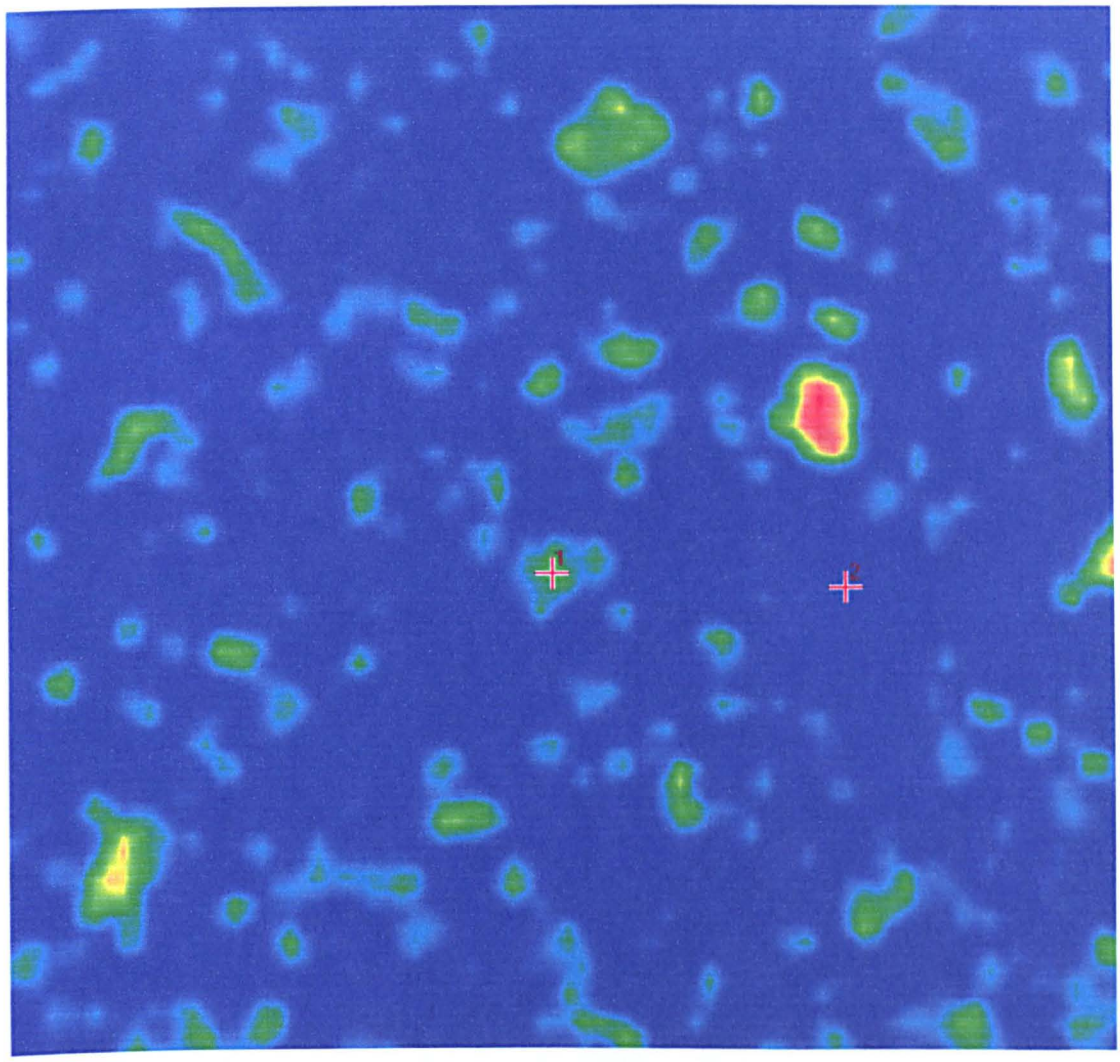

Figure 45: *Total absorbance image of the collected IR map (4000 – 720 cm^{-1}). The red coloured cross hairs labelled 1 – 2 indicate the original locations of the black and blue spectra in Figures 46 and 47 respectively.*

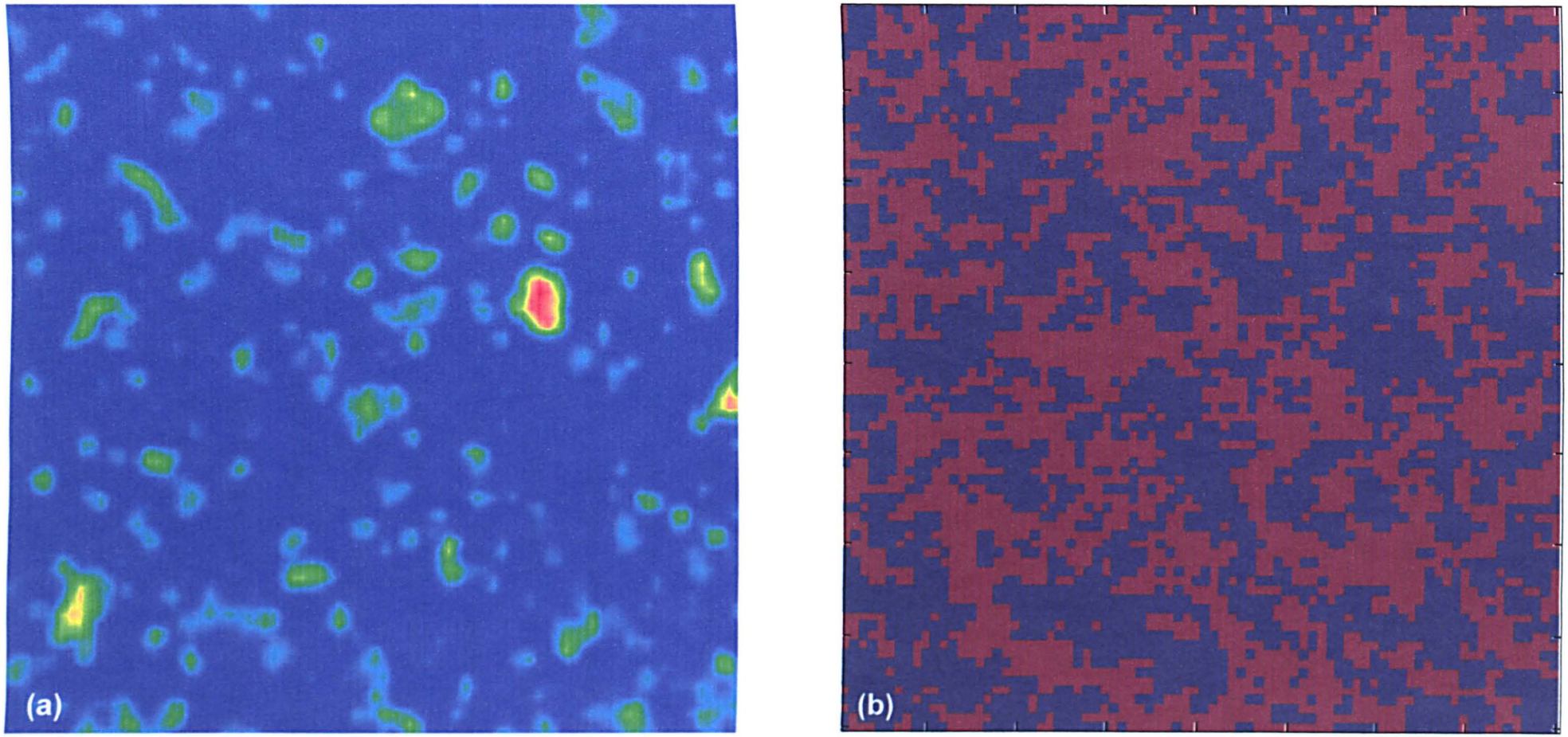

Figure 44: *PCA-FCM cluster imaging of abnormal cervical smear. (a) Total absorbance image of area IR mapped (4000 – 720 cm^{-1}). (b) False colour image constructed from a 2 cluster PCA-FCM analysis. Note that cluster membership does not mimic cellular presentation upon the slide. A large number of background spectra have been partitioned into the same cluster as cellular spectra.*

When investigating the data both before and after our data pre-processing routine, it became clear why insufficient separation of the background spectra was being made. Two example spectra that originated from a cellular and background pixel are shown both before and after data pre-processing in Figures 46 and 47 respectively. A total absorbance image for the collected map is shown in Figure 45 and further allows the co-ordinates of the extracted spectra to be located. The red coloured cross hairs labelled 1 – 2 in Figure 45 indicate the original locations of the black and blue spectra in Figures 46 and 47 respectively. When we examine the spectra before data pre-processing in Figure 46, we assume that clustering would be able to partition these types of spectra into separate groups since they are distinctly different. The cellular spectrum (black colour) exhibits a floating baseline with distinct absorption features characteristic of the cell, whereas the background spectrum (blue colour) displays a flat noisy baseline characteristic of conventional sample background subtraction. However, after our data-pre-processing routine (Figure 47) the background spectrum is distinctly different. The spectrum now lies at ~0.5 a.u. and has effectively been amplified by the normalisation process. Therefore features previously characteristic of noise have now been amplified to levels that could be incorrectly classified as peaks by the multivariate analysis. It now became clear that separation of background pixels must take place before baseline correction and normalisation procedures. This would enable only cellular spectra to be scrutinised by subsequent multivariate analyses. An FCM based filtering technique was thus developed to partition the background spectra. Before baseline correction and vector normalisation routines were employed, the raw spectra contained within the map were scrutinised by a 2 cluster FCM analysis. The resulting false colour image

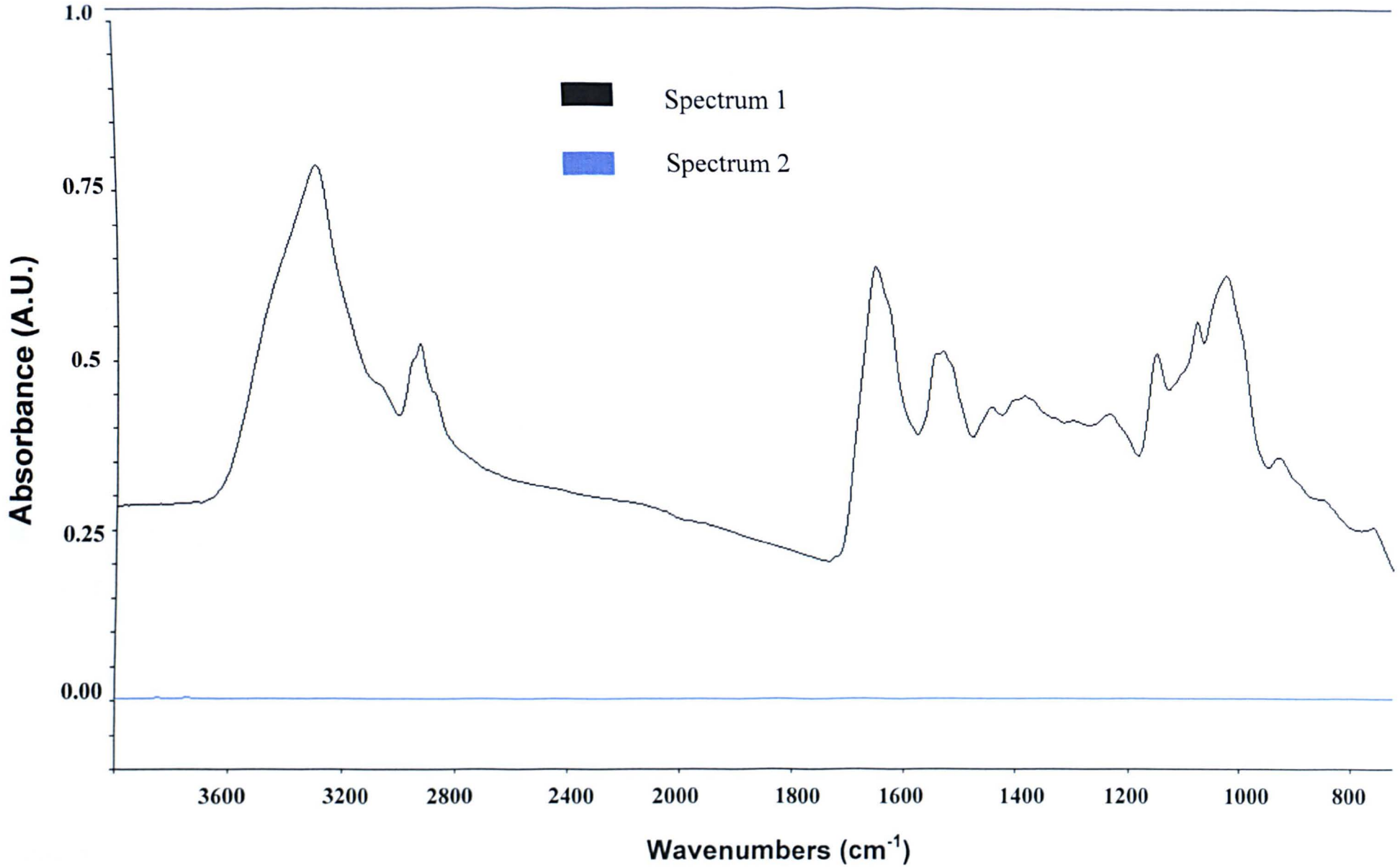

Figure 46: *A background and cellular spectrum extracted from the collected map. The co-ordinates from which these spectra were extracted are indicated in Figure 45.*

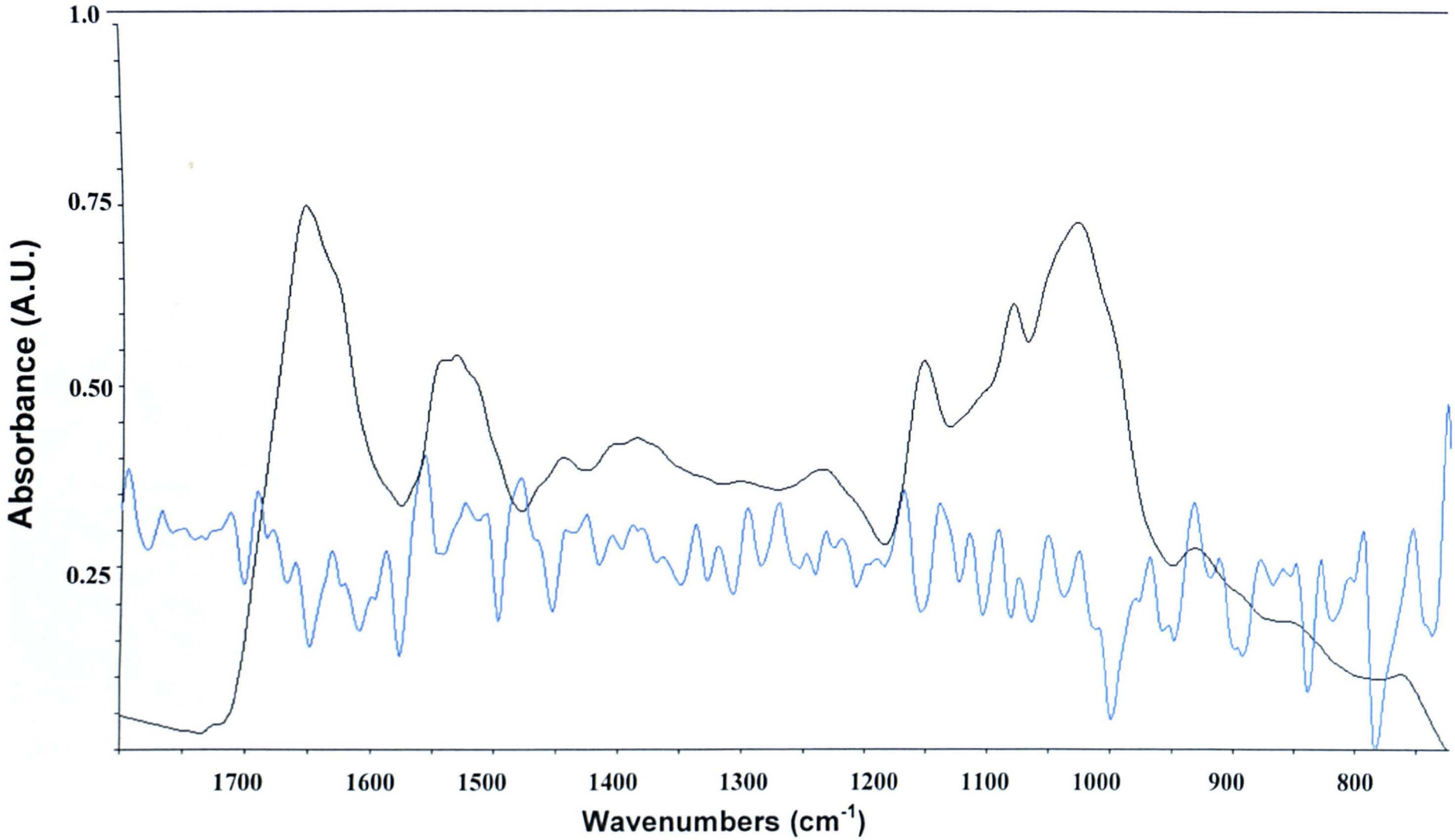

Figure 47: *A background and cellular spectrum extracted from the collected map after baseline correction and vector normalisation. The co-ordinates from which these spectra were extracted are indicated in Figure 45. Note the amplification of noise within the background spectrum that has rendered PCA-FCM cluster analysis less effective for the partitioning of these spectra.*

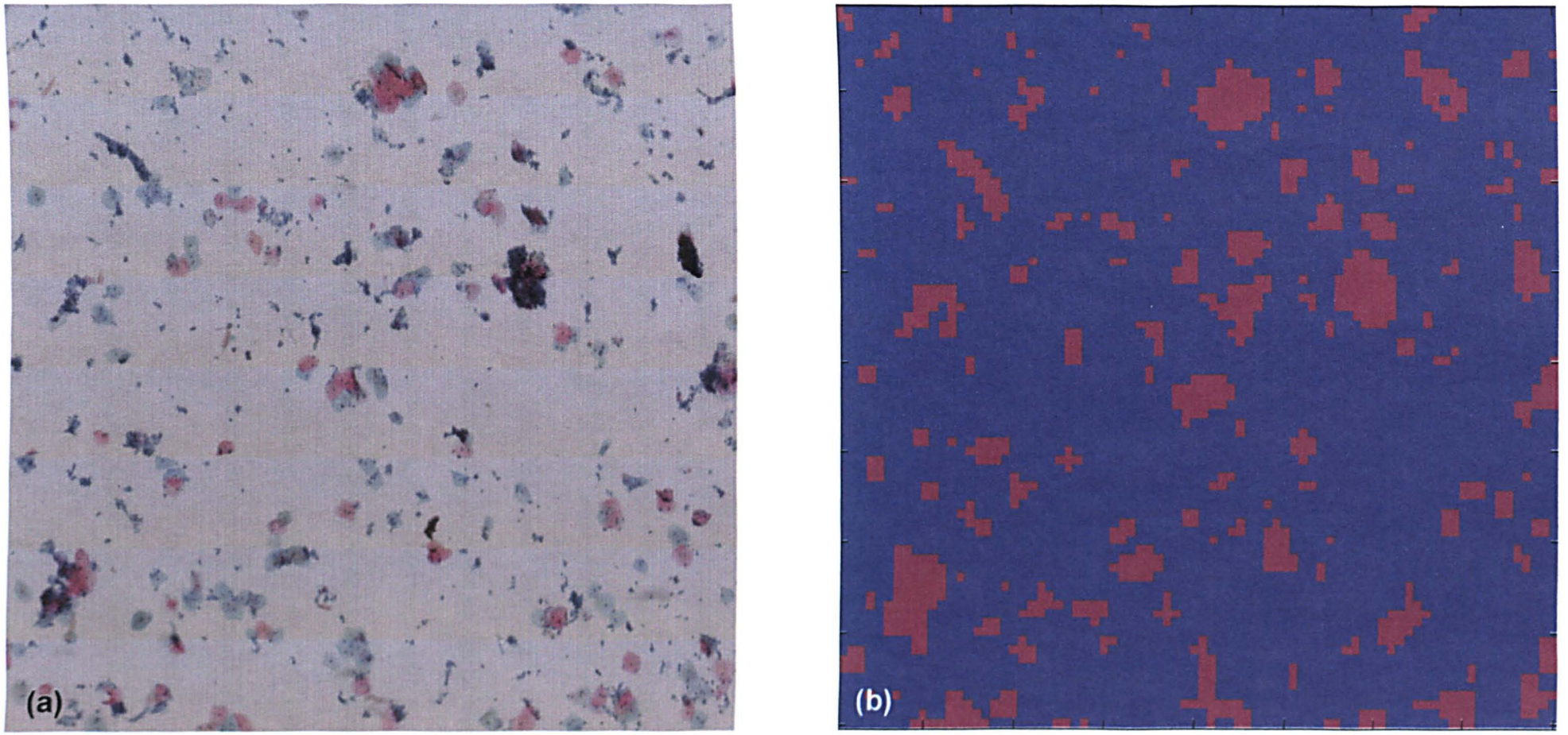

Figure 48: *A two cluster PCA-FCM analysis upon the raw spectra contained within the map. a) Pap stained image of mapped region. b) False colour image constructed from a 2 cluster PCA-FCM analysis of the raw spectra. Note the cellular spectra have been partitioned into the red cluster and the background spectra into the blue cluster.*

constructed from this analysis is displayed in Figure 48b. By direct comparison to the PAP stained image collected from the same region in Figure 48a, it shows this analysis was able to successfully partition the two types of spectra. Those spectra collected from points of cellular presence that displayed good signal to noise were partitioned into one group, and those that originated from background regions were partitioned into a separate group. This now enabled the background spectra to be filtered from any subsequent data pre-processing and PCA-FCM analysis. The cellular spectra were now processed via our conventional routines of baseline correction and vector normalisation before undergoing PCA-FCM cluster analysis. The cluster imaging results constructed from these analyses are displayed in Figures 49(b) – (h). When directly comparing these cluster images against the PAP stained image of the same region (Figure 49a), it appears the 4 cluster analysis provides some correlation to the cell types present. The orange cluster of spectra appears characteristic of healthy glycogen rich cells, whereas the cyan and maroon clusters seem to distinguish glycogen absent cells. However, a large amount of misclassification of polymorth spectra into these two clusters is also apparent. The cells diagnosed as displaying high grade dysplastic characteristics have alternatively been partitioned into the green cluster of spectra. But again a large amount of misclassification is noticeable with some polymorth spectra being partitioned into the same cluster. When cluster numbers were subjectively increased, the partitioning of spectra into more definable groups was not achieved and provided confusing results when correlated to histological diagnosis. However, after scrutinising the mean spectra calculated from the 4 cluster analysis, shown in Figure 50, it becomes clear that an additional distorting artefact is apparent in some spectra. Both the mean spectrum calculated for the cyan and red clusters display a distorted band shape.

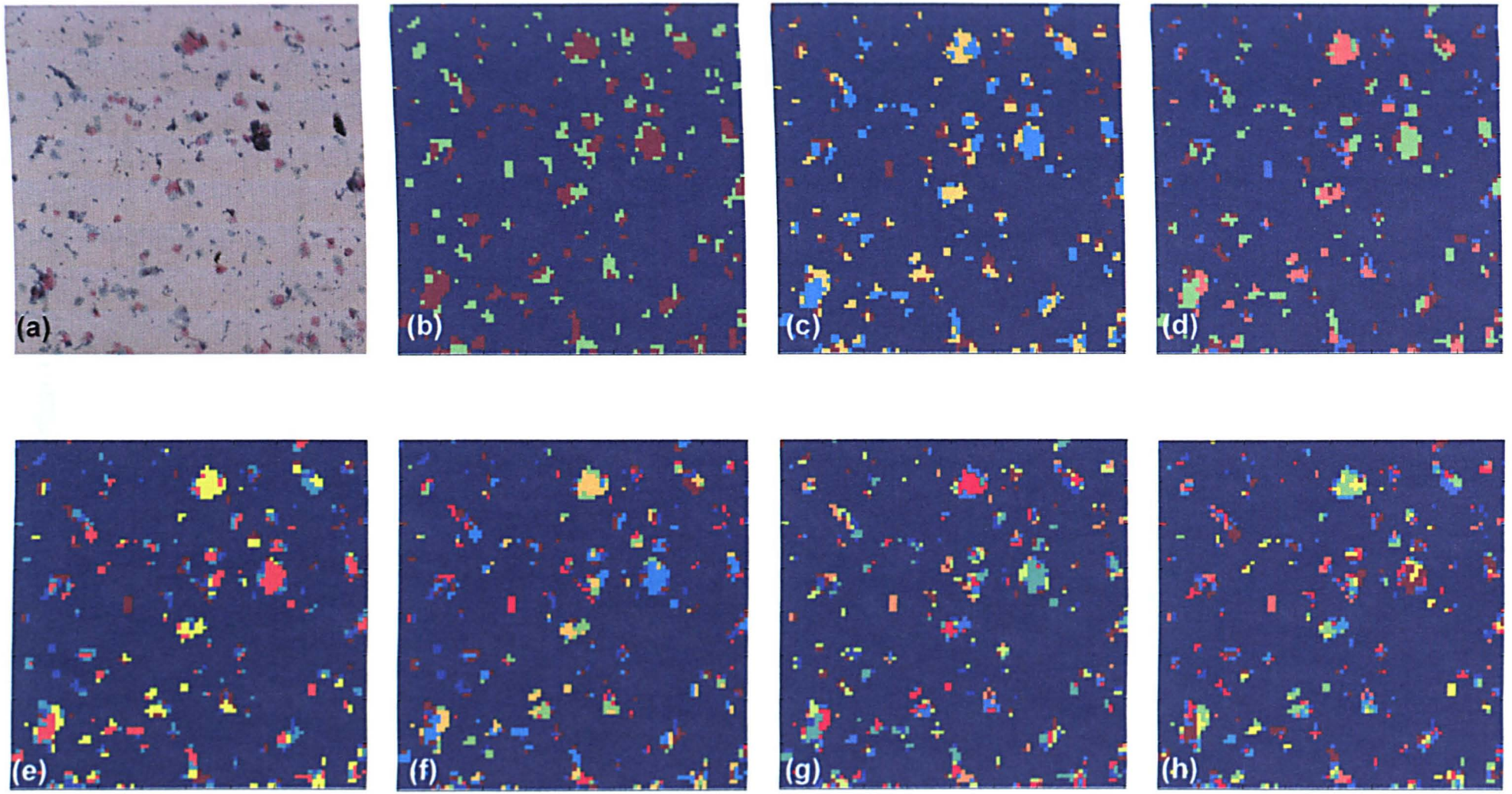

Figure 49: *IR imaging of cervical smear map via PCA-FCM Clustering. (a) PAP stained image of mapped area. (b) – (h) False colour images constructed using PCA-FCM Clustering Analysis results. Note the cluster numbers were subjectively increased from 2 – 8. Pixels with the same colour in each image are spectra that were partitioned into the same cluster. Additional data pre-processing included the initial filtering of background spectra via a 2 cluster FCM analysis. The remaining cellular spectra were then separately clustered using PCA-FCM analysis.*

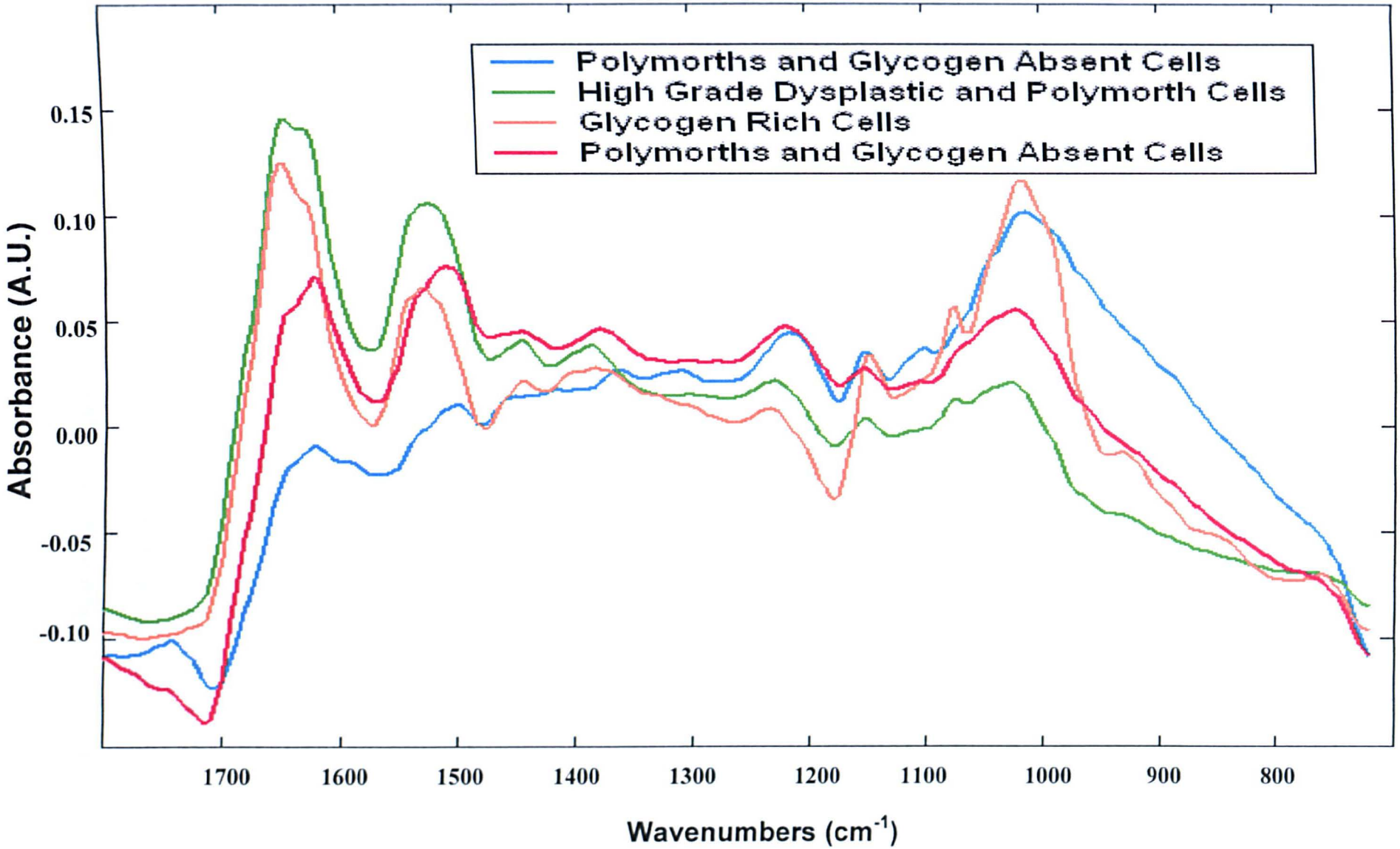

Figure 50: *4 Cluster PCA-FCM Analysis Results. Mean average spectra calculated from each cluster in the analysis.*

Peak maxima of some bands have been shifted toward lower wavenumber by up to 30 cm^{-1}. The amide I / amide II band intensity ratio also appears distorted, the bands now displaying relatively similar intensities uncharacteristic of tissue spectra. Similar dispersive band shapes have been identified previously for transflection spectra that were collected from the edges of tissue sections mounted on the same reflective substrates [59,60]. It is believed that the distorting artefact present in these spectra is caused by the superposition of dispersive and absorptive line shapes. The cause of this negatively contributing dispersive line shape was not discussed, but is believed to originate from rays of light that are far from the normal and thus strongly diffracted. These effects are apparent more so in spectra with very small absorbance, similar to those collected from individual cells. An algorithm to correct for these contaminations has been proposed by Diem and co-workers [60], and relies upon the transformation of a spectrum back into time (mirror displacement) domain by a complex reverse fast Fourier transform (FTT). This back transformation produces a "real" and "imaginary" interferogram. These are then separately forward transformed into the frequency domain to yield the pure reflective and absorptive components. A corrected spectrum with reduced artefact can thus be constructed by calculating a power spectrum from the reflective and absorptive components [60]. The correction algorithm can also be applied to undistorted spectra, which is particularly important since all spectra contained within a dataset must be uniformly pre-treated before multivariate analyses. We have collaborated with Diem and co-workers and used their correction algorithm for our work. This was further applied to our raw collected data. To demonstrate the effect of the dispersion algorithm, three cellular spectra have been extracted from our dataset. These include two spectra that were strongly contaminated by a reflection artefact, and one weakly

distorted spectrum. Both the raw and corrected spectrum for the three extracted spectra are displayed in Figures 52 and 53 respectively. A total absorbance image for the collected map is shown in Figure 51 and further allows the co-ordinates of the extracted spectra to be located. The red coloured cross hairs labelled 1 – 3 in Figure 51 indicate the original locations of the two strongly contaminated spectra (Spectrum 1 & 2) and the weakly distorted spectrum (Spectrum 3) respectively.

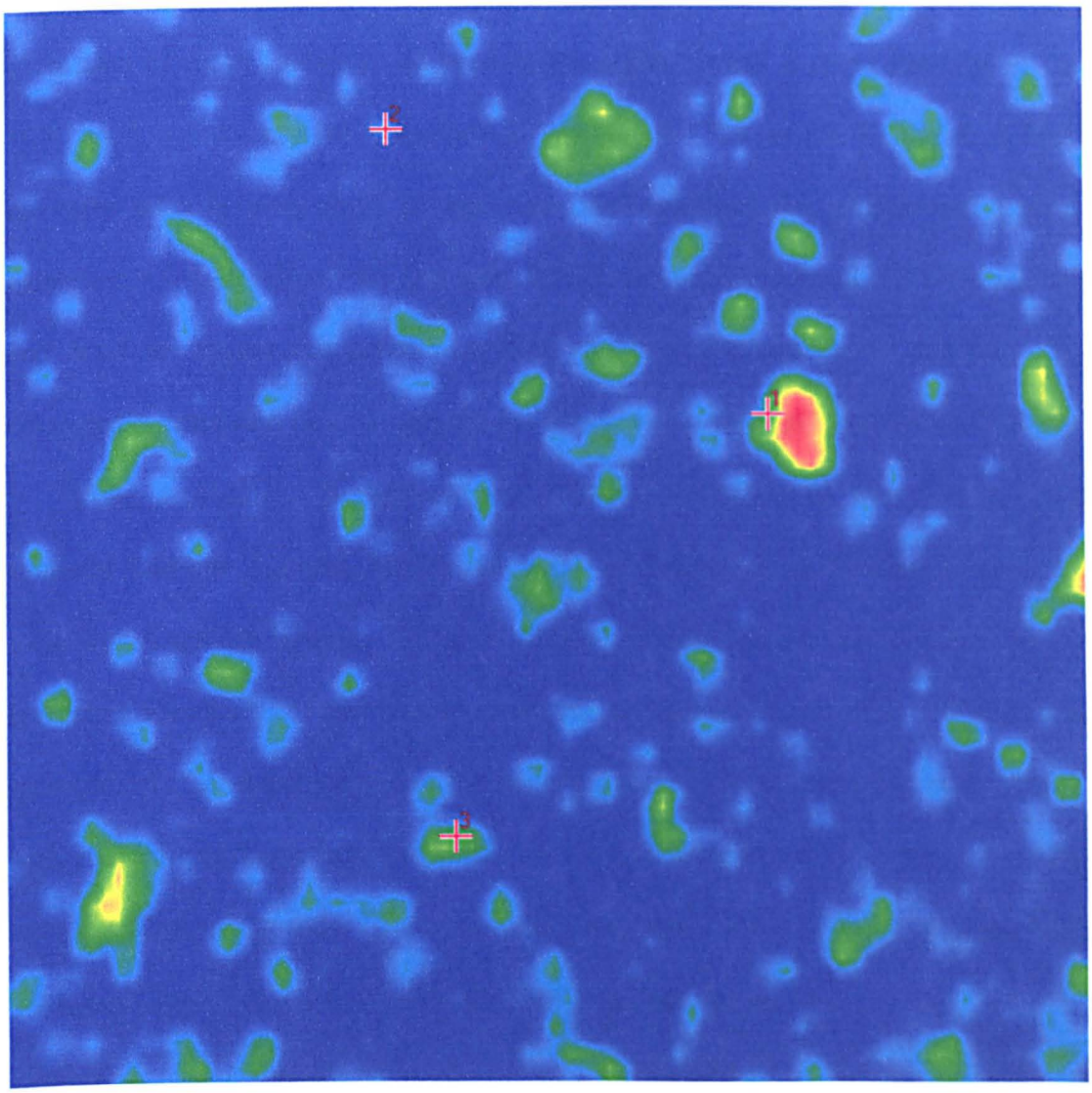

Figure 51: *Total absorbance image of the collected IR map (4000 – 720 cm^{-1}). The red coloured cross hairs labelled 1 – 3 indicate the original locations of spectra 1 – 3 in Figures 52 – 54 respectively.*

When examining the raw spectra in Figure 52, contamination from dispersion is noticeable with a clear shift in band shapes and intensities not observed in undistorted spectra. A large downward pointing feature at ~1700 cm^{-1} is also apparent and is thought to be due to the reflective component. Because of the reflective components negative intensity, the amide I band is shifted to a much lower wavenumber and a decrease in the intensity of this band is also observed. A similar feature is apparent in the weakly distorted spectrum shown in Figure 53. After correction of these contaminated spectra, the amide I peak is now observed at c.a. 1650 cm^{-1}, a more normal frequency for this band. The amide I / amide II band intensity ratio is also close to that normally observed for undistorted spectra. However, a broadening of the OH stretch region is apparent in all corrected spectra and does not appear comparable to those observed for undistorted spectra. The corrected spectra also display unusual baselines below 900 cm^{-1} and above 2000 cm^{-1}. When scrutinising further corrected spectra contained in the dataset, these spurious baseline features, which reach far up into high absorbance values, are inconsistent among spectra. A more unusual feature can be distinguished in the corrected spectrum of the weakly distorted spectrum (Figure 53). Within this spectrum, the band intensity ratio for the glycogen triplet of peaks has been altered, with peaks displaying maxima that are very close in intensity. This is never normally seen in undistorted spectra. But in general, the spectral features found between 1800 – 900 cm^{-1} appear closer to those observed for undistorted spectra.

To assess whether the use of this algorithm could improve subsequent multivariate analyses, it was applied to the entire spectral dataset. After correction, the spectra were cut to only include intensities recorded within the 1800 – 900 cm^{-1} region and

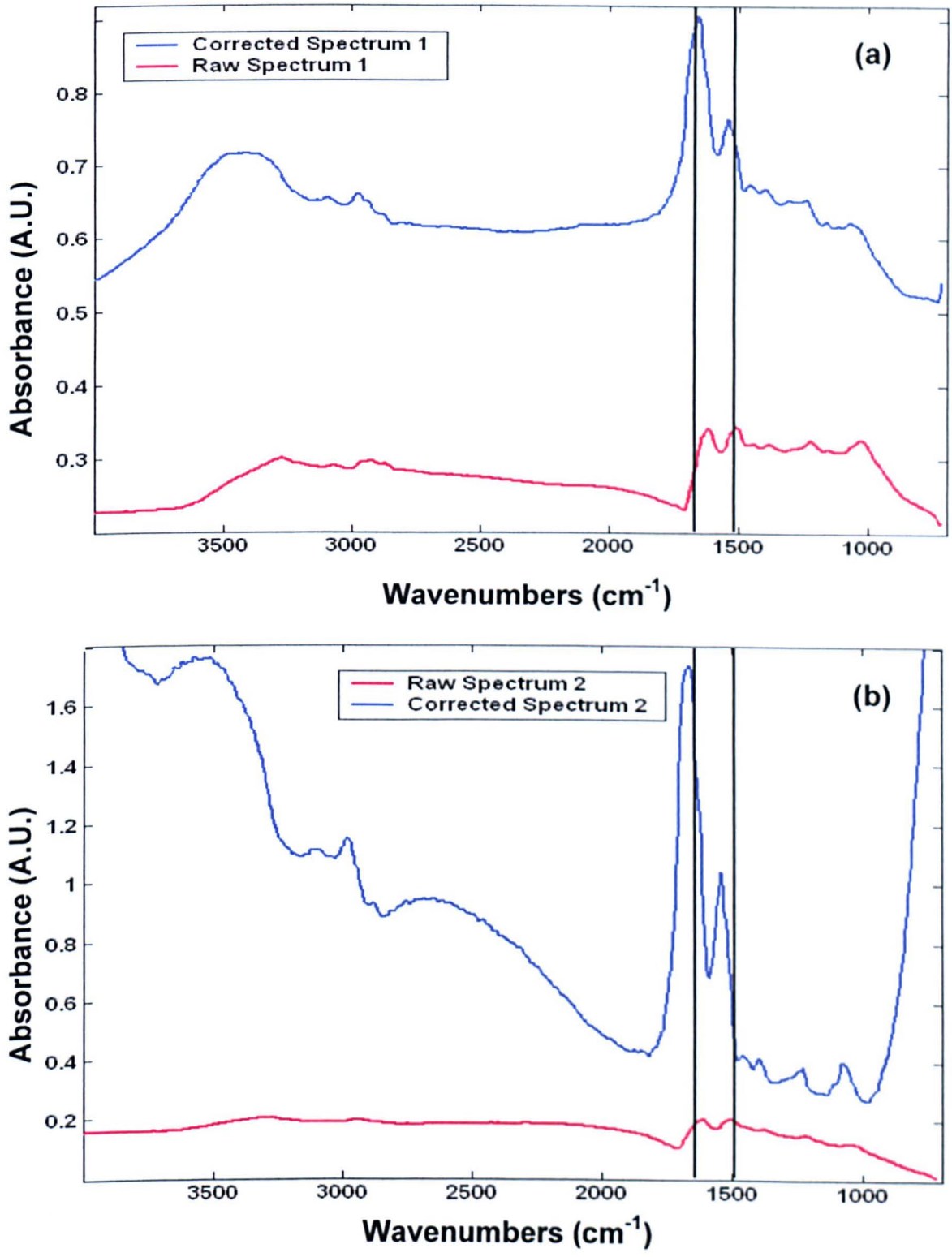

Figure 52: *Comparison of raw and corrected spectra. (a) Spectrum 1. (b) Spectrum 2. The raw spectra display strongly distorted band shapes. Note the shift of peak maxima toward lower wavenumbers by as much as 30 cm^{-1}. The amide I / amide II band intensity ratio also appears distorted. After correction for dispersion a shift in the band position and intensities is noticeable. However, it also apparent that the band shape above 2000 cm^{-1} and below 900 cm^{-1} is not comparable to that observed for undistorted spectra. The co-ordinates from which the spectra were extracted are indicated in Figure 51 by the cross hairs labelled 1&2 respectively.*

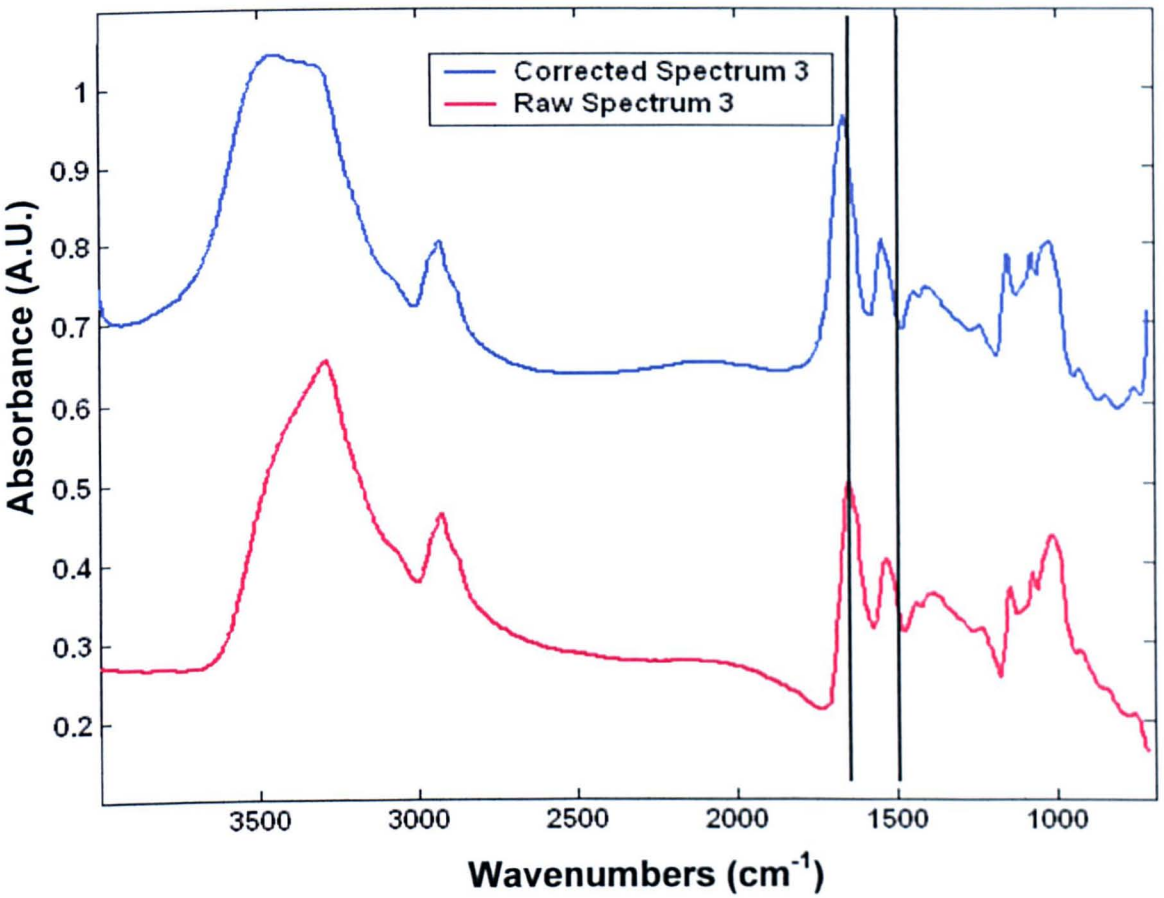

Figure 53: *Comparison of raw and corrected Spectrum 3. Note the small shift of peak maxima for the amide I band and the broadened OH stretch region for the corrected spectrum. The glycogen triplet of peaks also appears effected with a change in the peak intensity ratio noticeable. The co-ordinates from which the spectrum was extracted are indicated in Figure 51 by the cross hair labelled 3.*

baseline corrected using a 2 base point linear interpolation between the start and end data points. All spectra were then uniformly vector normalised before undergoing PCA-FCM cluster analysis. The cluster imaging results constructed from these analyses are displayed in Figures 54(b) – (h). When directly comparing these cluster images against the PAP stained image of the same region (Figure 54a), it appears on this occasion the 4 cluster analysis provides better correlation to the cell types present. The orange cluster of spectra appears characteristic of healthy glycogen rich cells, whereas the cyan cluster of spectra is more distinctive of glycogen absent cells. Some misclassification has been made with a number of polymorth spectra being

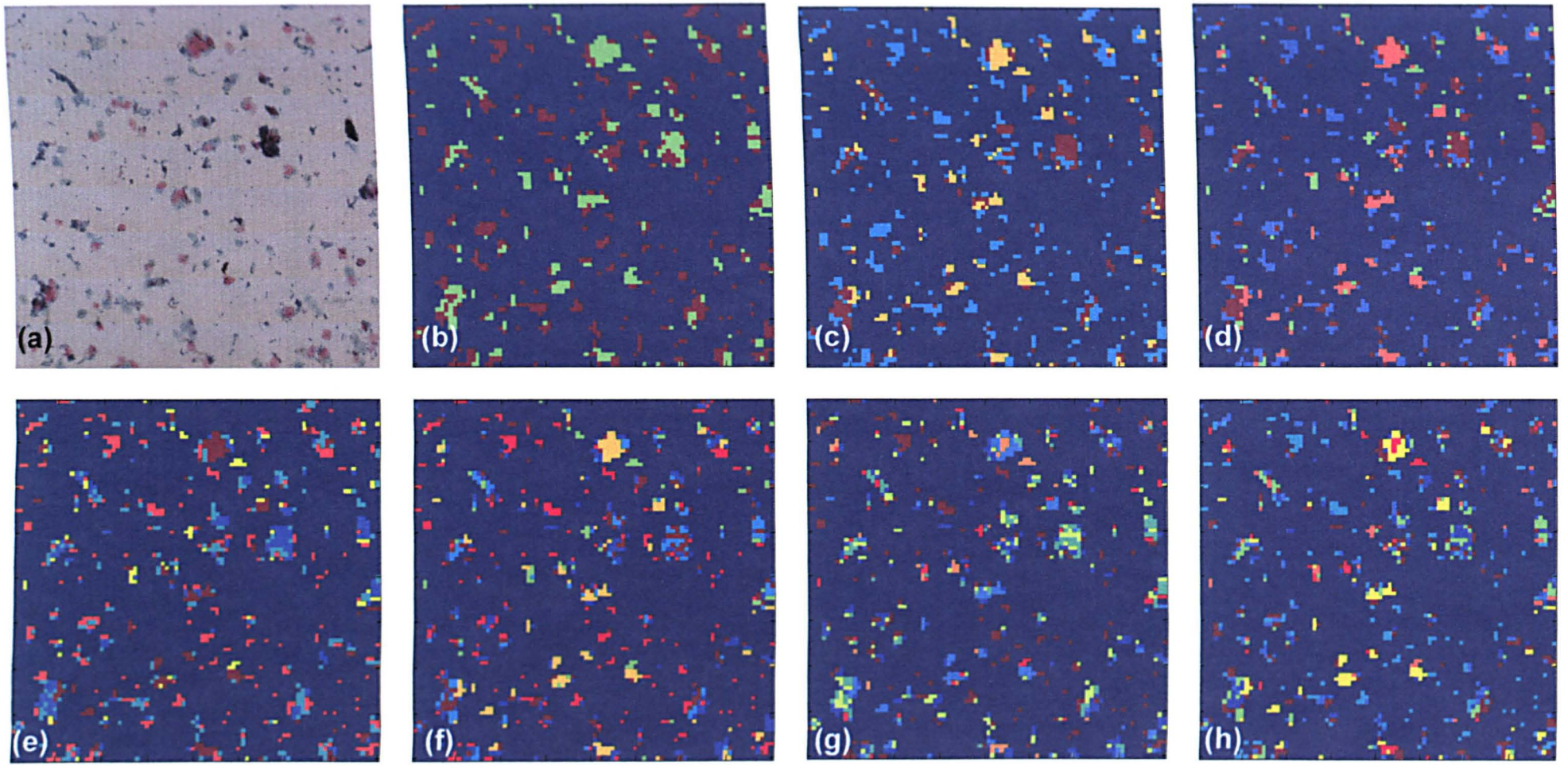

Figure 54: *IR imaging of cervical smear map via PCA-FCM Clustering. (a) PAP stained image of mapped area. (b) – (h) False colour images constructed using PCA-FCM Clustering Analysis results. Note the cluster numbers were subjectively increased from 2 – 8. Pixels with the same colour in each image are spectra that were partitioned into the same cluster. Additional data pre-processing included the application of a dispersion correction algorithm to all spectra and the initial filtering of background spectra via a 2 cluster FCM analysis. The remaining cellular spectra were then separately clustered using PCA-FCM analysis.*

partitioned into the cyan cluster for glycogen absent cells. However, when compared to the number of misclassifications observed previously without dispersion correction, this has been significantly reduced. The cells diagnosed as having high grade or low grade dysplastic characteristics were partitioned into the maroon cluster of spectra. Some misclassification of polymorth spectra into the same cluster is evident. Finally the green cluster of spectra appears to highlight polymorth cells. The mean average spectrum calculated for each cluster in the analysis is displayed in Figure 55. Similar spectral profiles are observed in tissue section mapping experiments as reported previously. The healthy squmaous cells are characterised by two main groups. Those that are rich with glycogen (orange cluster) display a triplet of peaks at c.a. 1150, 1075 and 1020 cm^{-1} that correspond to the C-O stretch, C-C stretch and C-O-H deformation modes respectively. In contrast, healthy squamous cells that are absent of glycogen (cyan cluster) display a more resolved region below 1200 cm^{-1}. This allows the symmetric phosphate (PO_2^-) band at 1080 cm^{-1} to be revealed characteristic of nucleic acids. However, the intensity of this band is relatively weak when compared to the amide I and amide II modes located at within the 1700 – 1500 cm^{-1} region. This difference is likely to reflect a large contribution from proteins in these types of cells. Abnormal cells diagnosed as being high grade or low grade dysplasia's have a similar spectral profile. But in contrast to healthy cells, the symmetric and antisymmetric phosphate bands located at 1080 and 1240 cm^{-1} are very pronounced and have a relatively high contribution to the spectrum when compared to the amide band intensities. This is likely to reflect a larger contribution from nucleic acids in abnormal squamous cells, a finding which correlates to previous work upon unhealthy tissue sections. The average spectrum

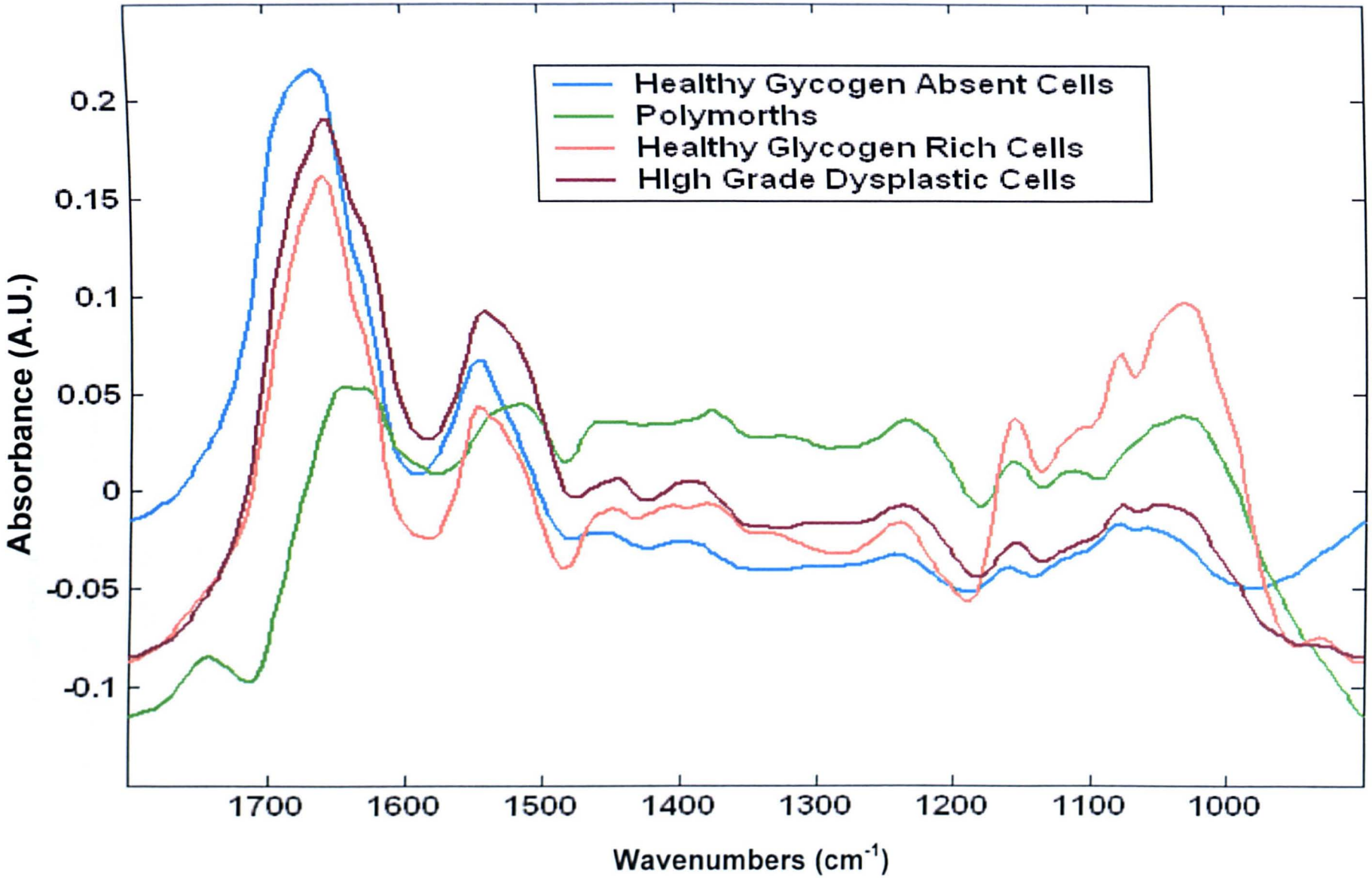

Figure 55: *4 Cluster PCA-FCM Analysis Results after dispersion correction. Mean average spectra calculated from each cluster in the analysis.*

for polymorths (green cluster) is hard to interpret since the spectrum still appears distorted. The amide I and amide II modes are again shifted to lower wavenumber and display similar band intensities, an uncharacteristic feature for cellular spectra. However, it appears as though there is a strong nucleic acid contribution to these spectra. If we consider that polymorth cells are c.a. 1 – 10 μm in size and are comprised almost entirely by nuclei, it is not surprising that large nucleic acid contributions are present. The distortion observed in these spectra from dispersive line shapes is also understandable considering a 25 x 25 μm spatial resolution was utilised. Unless a multiple number of polymorths were clumped together in a group, the sampling area would not be entirely filled and could thus allow stray light to contribute to the spectrum collected.

In conclusion, the spectroscopic diagnosis of exfoliated cervical cells remains a complex goal. The introduction of LBC techniques for sample preparation, which provide monolayer cellular presentation upon substrates, has enabled spectroscopic analysis of single cells. This step is vitally important since macroscopic spectra collected from exfoliated cervical cell pellets are prone to feature contaminating artefacts from non-diagnostic cells. However, care must be taken when preserving the exfoliated cells. If smear material is stored within common medical preservative solutions such as formalin or methanol, the protein structures within cells can become changed. This change in protein structure is manifested within spectra by a movement in the frequency or a splitting of the amide I band, which renders spectroscopic diagnosis impossible. From our method development experiments, we believe a 24% ethanol solution provides the best preservation of cells for spectroscopic applications, with little or no spectral artefacts present. Until relatively

recently, the collection of spectral maps from large spatial areas was very time consuming. This was achieved by scanning a sample in a raster pattern through the focal point of a single detector, using steps that were the same size as the x and y dimensions of the pixel element. However, the advent of highly sensitive linear array detector systems that are coupled to rapid stage motion has enabled the collection of much larger spectral maps within reasonable collection times. This type of data acquisition appears ideal for the examination of exfoliated cell smears, where cell populations are vast and diagnostic cells often sparsely orientated around the sample area. During our study, we opted to use a 25 x 25 μm pixel size to enable large sample areas to be examined within reasonable collection times. Although this data acquisition method permitted the collection of spectra from up to 10,000 cells, it appears it may have also introduced dispersion artefacts into the spectra. When the sample area is not entirely filled by cellular material it is possible stray light could contaminate the spectrum. Spectra collected from pixels that encompass very small cells or lied upon the edges of large cells or clumps appear to be most effected by this contamination. The application of a dispersion correction algorithm did improve the appearance of these distorted spectra within the 1800 – 900 cm^{-1} region. However, unusual artificial artefacts were introduced at other points of the spectrum, with a distinct broadening of the OH stretch region and unusual baselines evident. The use of a smaller sampling area (i.e. 6.25 x 6.25 μm) could reduce the number of these distorted spectra and allow very small cells to be scrutinised, which are often those displaying dysplasia. But this would in turn significantly reduce the size of the region that could presently be examined and increase the collection time required. The presence of a single cell exhibiting dysplasia provides an abnormal diagnosis by a histopathologist. It is therefore very important a significantly large cell population

is examined. At present, linear array detector systems are not capable of providing such sensitivity, but it is reasonable to believe future systems could achieve this. The reflective substrates utilised for these experiments may also be limiting, since dispersion artefacts observed in these transflection spectra have not been seen previously in our transmission experiments upon BaF_2 substrates. Thus exfoliated cell preparation upon transmission substrates, although expensive, may provide spectra that are not as strongly contaminated by dispersion artefacts. A system that utilises large BaF_2 blocks and include multiple sample spots is feasible. These could then be washed after spectroscopic diagnosis and re-used, since collected image data could be stored within medical records rather than sample slides.

Despite the contaminating features present within the spectra, the main types of cell could be identified via PCA-FCM cluster imaging. Glycogen rich, glycogen absent and highly dysplastic cells were identified via distinct spectral changes that characterise each cell type. But a high degree of misclassification is also apparent and presently limits the statistical confidence of cell diagnosis via spectroscopy. These misclassifications were caused in part by dispersion contamination to the collected spectra. But in addition, polymorth and dysplastic cells display very similar spectral profiles, both having large contributions from nucleic acids.

3.4 Conclusions

In this chapter we have used FTIR imaging to study cervical tissue sections and exfoliated single cells. To summarise the results I have:

- Demonstrated that frozen sectioning of cervical tissue specimens does not adversely affect the sample characteristics. This preparation method negates standard procedures more commonly employed that include paraffin embedment and subsequent de-paraffinization.

- Assessed a number of different liquid based techniques for the preparation of cervical smear material. The use of ethanol as a preservative appears to provide good cellular presentation upon reflective substrates without any substantial cell damage or change in the biochemical make up of the cells.

- Mounted tissue section samples upon BaF_2 substrates that enable transmission spectra to be collected. These were free from contaminating dispersion artefacts that are often observed using transflection sampling methodologies. Subsequent multivariate analyses could therefore utilise the full spectral range of the data and classify spectra according to spectral features that were characteristic of the sample alone.

- Applied a variety of unsupervised multivariate analysis techniques to the collected spectral datasets. A comprehensive and detailed comparison between techniques for tissue discrimination was therefore achieved. When correlating the results to the known histology of the samples, FCM clustering achieved the best tissue characterisation.

- Collected spectral datasets from an array of different cervical tissue sections that describe a number of different pathological states and tissue types. The

spectral characteristics that are descriptive for both healthy and unhealthy squamous and columnar epithelium are reported in detail. Diseased or abnormal cells exhibit distinctive spectral changes. Contributions from glycogen and glycoproteins are absent and replaced by more pronounced nucleic acid features below 1400 cm^{-1}. The protein absorption bands also display changes with cellular abnormality. It is therefore essential to collect spectra that are free from dispersion artefacts or be able to correct for these contaminations, as such spectral differences (distortion of the amide I and II absorption bands) would be detrimental for accurate tissue discrimination.

- Demonstrated that FTIR imaging can be used to examine and classify exfoliated cervical cells in conditions similar to that found in real cytology. However, the use of reflective substrates appears to introduce dispersive artefacts to the collected spectra. These contaminating features limit the statistical confidence for accurate spectroscopic diagnosis. Efforts were made to correct for these contaminations by use of a dispersion artefact reduction algorithm, which did substantially improve the spectral maps. But problematic artificial components were introduced into some spectra by an overcompensation of the correction algorithm. Further experiments are therefore required to determine the best parameters for data acquisition and to assess the spectral variations that accompany both healthy and abnormal smear material.

3.5 References

[1] D. M. Parkin, P. Pisani and J. Ferlay, *Int. J. Gynecol. Cancer*, 1999, **80**, 827.

[2] D. M. Parkin, F. Bray, J. Ferlay and P. Pisani, *CA Cancer J. Clin.*, 2005, **55**, 74.

[3] Taken from Cancer Research UK website, located at URL http://info.cancerresearchuk/org/cancerstats/types/cervix/incidence/#source5.

[4] Taken from Cancer Research UK website, located at URL http://info.cancerresearchuk/org/cancerstats/types/cervix.

[5] M. Quinn, P. Babb, J. Jones and E. Allen, *Brit. Med. J.*, 1999, **318**, 904.

[6] G. N. Papanicolaou and H. F. Traunt, *Am. J. Obstet. Gynecol.*, 1941, **42(2)**, 193.

[7] M. H. Schiffman, H. M. Bauer, R. N. Hoover, A. G. Glass, D. M. Cadell, B. B. Rush et al., *J. Natl. Cancer Inst.*, 1993, **85**, 958.

[8] E. L. Franco, E. Duarte-Franco and A. Ferenczy, *Can. Med. Assoc.J.*, 2001, **164**, 1017.

[9] G. D. Zielinski, P. J. F. Snijders, L. Rozendaal, F. J. Voorhorst, H. C. van der Linden, A. P. Runsik, et al., *Br. J. Cancer*, 2001, **85**, 398.

[10] K. Elfgren, M. Jacobs, J. M. M. Walboomers and C. J. L. M. Meijer, *Obstet. Gynecol.*, 2002, **100**, 965.

[11] P. D. Sasieni, J. Cuzick and E. Lynch-Farmery, *Br. J. Cancer*, 1996, **73**, 1001.

[12] J. D. Gay, L. D. Donaldson and J. R. Goellner, *Acta Cytol.*, 1985, **29**, 1043.

[13] Morbidity and Mortality Weekly Report, *Regulatory closure of cervical cytology laboratories: recommendations for a public health response*, 1997, U.S. Department of Health and Human Services, Atlanta.

[14] L. DiBonito, G. Falconieri, G. Tomasic, I. Colautti, D. Bonifacio and S. Dudine, *Cancer,* 1993, **72**, 3002.

[15] R. K. Dukor, *Handbook of Vibrational Spectroscopy*, 2002, volume 5, 3335. Edited by J. M. Chalmers and P. R. Griffith, Wiley and Sons, Chichester.

[16] J. Patnick, *European J. Cancer*, 2000, **36**, 2205.

[17] L. J. Mango and P. T. Valente, *Acta Cyto.*, 1998, **42**, 227.

[18] B. J. Fetterman, G. F. Pawlick, H. Koo, J. S. Hartinger, C. Gilbert and S. Connell, *Acta Cytol.*, 1999, **43**, 13.

[19] J. J. Baker, *Diagn. Cytopathol.*, 2002, **27**, 185.

[20] D. C. Wilbur, *Am. J. Clin. Pathol.*, 2002, **118**, 399.

[21] D. L. Wetzel and S. M. LeVine, *Biological applications of infrared microspectroscopy,* 2001, 1. Edited by H. U. Gremlich and B. Yan, Marcel Dekker, New York.

[22] H. H. Mantsch and D. Chapman, *Infrared spectroscopy of biomolecules*, 1996, Wiley-Liss, New York.

[23] M. Diem, *Analyst*, 2004, **129**, 880.

[24] M. Jackson and H. H. Mantsch, *Infrared Spectroscopy: Ex Vivo Tissue Analysis, Encyclopedia of Analytical Chemistry*, 2000, **1**, Wiley, Chichester.

[25] M. Diem, S. Boydston-White and L. Chiriboga, *Appl. Spectrosc.*, 1999, **53(4)**, 148A.

[26] M. Diem, *Vibr. Spectrosc.*, 2003, **32(1)**, 1.

[27] S. Boydston-White, M. Diem, M. Romeo, R. Mendelsohn and Y. Ozaki, *Vibr. Spectrosc.*, 2005, **38(1-2)**, 1.

[28] P. T. T. Wong, R. K. Wong and M. F. K. Fung, *Appl. Spectrosc., 1993,* **47(1)**, 1058.

[29] P. T. T. Wong, R. K. Wong, T. A. Caputo, T. A. Godwin and B. Rigas, *Proc. Natl. Acad. Sci.*, 1991, **88**, 10988.

[30] M. A. Cohenford and B. Rigas, *Proc. Nat. Acad. Sci.*, 1998, **95**, 15327.

[31] M. A. Cohenford, T. A. Godwin, F. Cahn, P. Bhandare, T. A. Caputo and B. Rigas, *Gynecol. Oncol.*, 1997, **66**, 59.

[32] M. A. Cohenford, P. S. Bhandore, B. Rigas and K. Krishman, *Mikrochim. Acta.*, 1997, **14**, 433.

[33] B. R. Wood, M. A. Quinn, F. R. Burden and D. McNaughton, *Biospectrosc.*, 1996, **2**, 143.

[34] B. R. Wood, M. A. Quinn, B. Tait, T. Hislop and M. Romeo, *Biospectrosc.*, 1998, **4**, 75.

[35] B. R. Wood, M. Q. Quinn and D. McNaughton, *Spectroscopy of Biological Molecules*, 1997, 445. Edited by P. Carmona, R. Navarro and A. Hernanz, Kluwer Academic Publishers, Dordrecht.

[36] B. R. Wood, M. Q. Quinn, B. Tait, M. Romeo and H. H. Mantsch, *Biospectrosc.*, 1998, **4**, 75.

[37] L. Chiriboga, P. Xie, H. Lee *et al.*, *Biospectrosc.*, 1998, **4**, 47.

[38] L. Chiriboga, P. Xie, V. Vigorita, D. Zarou, D. Zadim and M. Diem, *Biospectrosc.*, 1997, **4**, 55.

[39] L. Chiriboga, P. Xie, W. Zhang and M. Diem, *Biospectrosc.*, 1997, **3**, 253.

[40] L. Chiriboga, P. Xie, H. Yee, D. Zarou, D. Zakim and M. Diem. *Cell. Mol. Biol.*, **44**, 219.

[41] S. Boydston-White, T. Gopen, S. Houser, J. Bargonetti and M. Diem, *Biospectrosc.*, 1999, **5**, 219.

[42] L. Chiriboga, H. Yee and M. Diem, *Appl. Spectrosc.*, 2000, **54**, 1.

[43] L. Chiriboga, H. Yee and M. Diem, *Appl. Spectrosc.*, 2000, **54**, 480.

[44] M. Diem, L. Chiriboga and H. Yee, *Biopolymers*, 2000, **57(5)**, 282.

[45] M. Romeo, B. R. Wood, M. A. Quinn and D. McNaughton, *Vibr. Spectrosc.*, 2002, **28**, 167.

[46] M. Romeo, B. R. Wood and D. McNaughton, *Vibr. Spectrosc.*, 2002, **28**, 167.

[47] A. Stevens and J. Lowe, *Histology*, 1992, Glower Medical Publishing, London.

[48] Taken from the Cytyc Website, manufacturers of the ThinPrep instrument, located at URL http://www.cytyc.com/women/women_cervical_cancer_clinical.shtml

[49] M. Jackson and H. H. Mantsch, *Biomedical Applications of Spectroscopy*, 1996, 185. Edited by H. Clark, Wiley, New York.

[50] P. Lasch, W. Haensch, D. Naumann and M. Diem, *Biochimica et Biophysica Acta*, 2004, **1668**, 176.

[51] B. R. Wood and D. McNaughton, *Multichannel detectors and multivariate imaging: FPA Imaging and spectroscopy for monitoring chemical changes in tissue.* Edited by R. Bhargava and I. W. Levin, 2005, Blackwell Publishing, Oxford.

[52] M. F. K. Fung, M. Senterman, P. Eid, W. Faught, N. Z. Mikael and P. T. T. Wong, *Gynecol. Oncol.*, 1997, **66**, 15.

[53] B. R. Wood, L. Chiriboga, H. Yee, M. A. Quinn, D. McNaughton and M. Diem, *Gynecol. Oncol.*, 2004, **93**, 59.

[54] P. Lasch, W. Haensch, E. N. Lewis, L. H. Kidder and D. Naumann, *Appl. Spectrosc.*, 2002, **56**, 1.

[55] M. Mantsch and M. Jackson, *J. Mol. Struct.*, 1995, **347**, 187.

[56] M. Romeo, F. Burden, M. Quinn, B. Tait, B. Wood and D. McNaughton, *Cell. Mol. Biol.*, 1998, **44(1)**, 179.

[57] J. Kent, 4th Year MSci Project Report, Nottingham Univeristy, 2002.

[58] B. Mohlenhoff, M. Romeo, M. Diem and B. R. Wood, *Biophysical Journal*, 2005, **88**, 3635.

[59] M. Romeo and M. Diem, *Vibr. Spectroscoc.*, 2005, **38**, 115.

[60] M. Romeo and M. Diem, *Vibr. Spectrosc.*, 2005, **38**, 129.

Chapter 4

Experimental Section and Method Development

4.1 Sample Preparation

4.1.1 Tissue Sample Collection and Preparation

Human tissue specimens were collected from both the lymph node and cervix for spectroscopic analysis. The biopsy material was collected under the approval of the Research Ethics Committee at either Gloustershire Royal or Derby City Hospitals. Cervical samples were obtained by either cone biopsy or hysterectomy. The lymph node samples were collected during routine surgical resection for breast cancer. The lymph nodes examined in this study existed toward the end of the chain of nodes. This allowed conventional histological diagnosis, using the sentinel and immediately adjacent lymph nodes. The tissue specimens collected, both cervical and lymph node, were then mounted onto acetate paper and immediately snap frozen in liquid nitrogen to maintain their biochemical condition. Tissue sections from the specimen were then prepared using a freezing-microtome, producing 7μm thick sections which were suitable for spectroscopic analysis (Figure 1). These were then placed onto BaF_2 disc and stored in a cryovial ready for spectroscopic data collection. The remainder of the tissue specimen was then treated in the conventional way for histology, allowing a parallel tissue section to be cut for comparative analysis by a consultant histopathologist. A variety of different tissue sections were examined in this thesis,

and ranged from those diagnosed as being completely benign, to positive sections almost entirely infiltrated by malignant tissue.

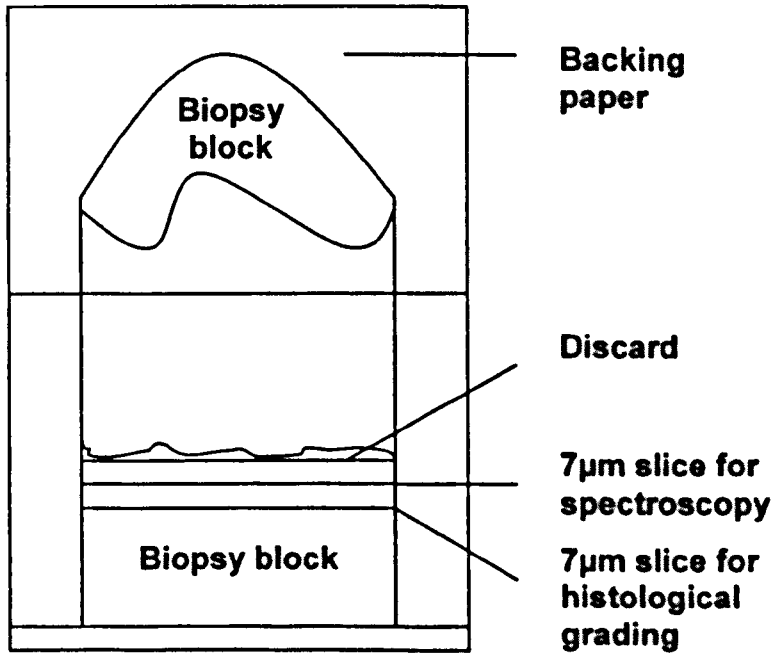

Figure 1. A schematic representation of the process used for frozen sectioning. Biopsy material was placed onto acetate paper and immediately snap frozen. By use of a freezing microtome, 7µm thick tissue sections were then cut for subsequent spectroscopic or histological analysis.

4.1.2 Exfoliated Cell Sample Preparation

Methods for preparing exfoliated cells for cytological screening have come under great scrutiny over the past decade. The standard method introduced by the National Health Service Cervical Screening Program (NHSCSP) utilised the Papanicolaou (PAP) smear test [1 – 2]. A pap smear was carried out by a GP or nurse at a primary care or community clinic. Cervical cells were collected using a wooden disposable

spatula device, spread across a glass slide, and subsequently fixed. The slides produced were then transported to a hospital laboratory to be stained for examination by a cytologist. At the time of our initial experiments, this was still the method of choice used by all hospital laboratories in the UK. However, this procedure presented several problems to both the cytologist and spectroscopist. The samples created often contained a number of unwanted characteristics making them insufficient for effective diagnosis. They displayed varying thickness across the slide with large areas of cell clumping and the formation of debris from cells that were damaged or broken apart. Contamination of samples with blood and inflammatory cells were also apparent in some cases, again making them hard to interpret by a cytologist. In terms of spectroscopic analysis, locating and scrutinising an individual cell by infrared microscopy, which could later be diagnosed, was often hard to achieve. To overcome this problem a different method named Liquid Based Cytology (LBC) was investigated and has gained large support in subsequent literature [3,4]. The technique was initially introduced in the Untied States in 1991 [5], and enabled a monolayer coverage of cells to be achieved when analysing fine needle biopsy and body fluid material. After its initial success it finally gained approval for clinical use in PAP smear analysis during 1996 in the USA [6]. A distinct advantage of the LBC method is the ability to create more homogenous samples that enable a larger proportion of the sample to be analysed. This in turn allows multiple tests to be carried out on an individual sample [7]. Most importantly, pilot studies have shown the technique can allow increased sensitivity and specificity for the detection of pre-cancerous lesions [7-9].

Current methods that use LBC technology include: (i) ThinPrep; (ii) SurePath; (iii) Cytoscreen and (iv) Lebonard Easy Prep. ThinPrep and SurePath LBC systems were examined in this Thesis, since these techniques represent the two leading methods presently being adopted by local medical councils around the UK and United States. Both systems enable monolayer cell coverage upon glass slides, but utilise different techniques to achieve this. At present neither system has been identified as the leading or preferred technique for sample preparation with many clinics in the USA providing both for patient choice.

4.1.2.1 ThinPrep Sample Preparation

This method involves the use of apparatus developed in the USA that provide a semi-automated (T2000) or fully automated (T3000) slide processor system (Cytyc Corporation, USA) as shown in Figure 2.

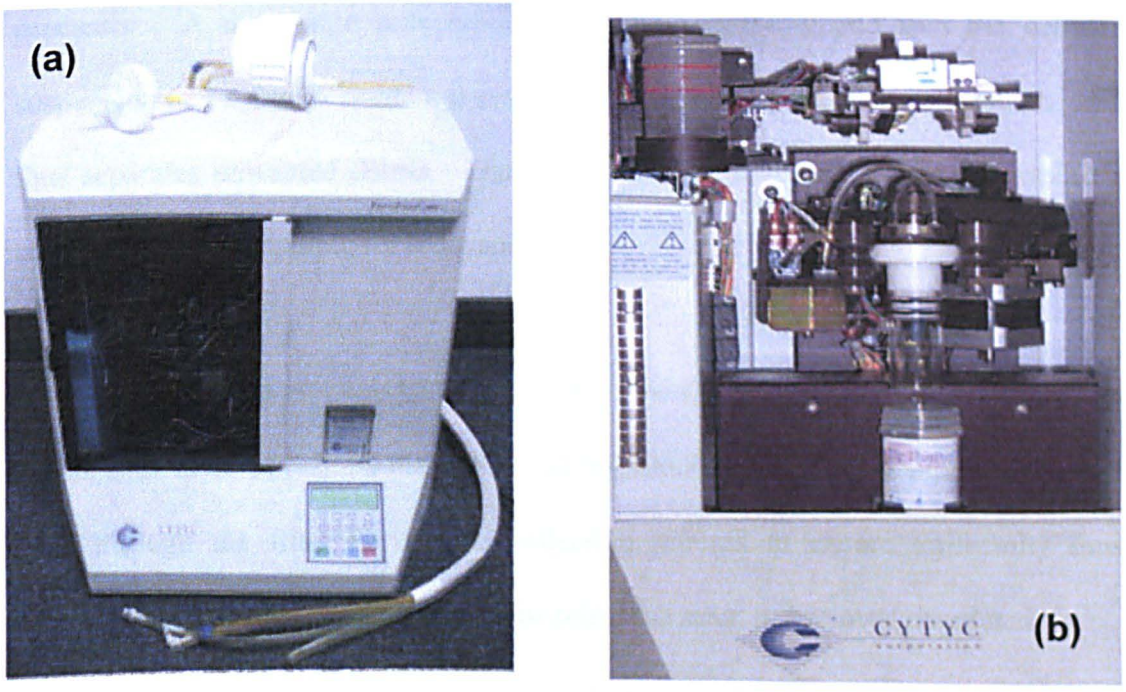

***Figure 2**: Pictures of the apparatus used for thin prep work. a) Thin Prep 2000 Slide Processor. b) Internal features of the thin prep instrument. Taken from ref [10].*

The method can be characterised by the following 5 steps:

Step 1: A gynaecological sample is collected using a broom-type or cytobrush/spatula cervical sampling device (for example, Cervex® Rovers Diagnostic Devices, USA). The central bristles of the device are inserted far enough into the cervical space enabling cells from the endocervix to be obtained. The side bristles sweep cells from the ectocervix and transformation zone.

Step 2: Instead of smearing the exfoliated cells onto a slide the sampling device is rinsed into a transport vial containing PreservCyt®, a preservative 54% methanol based liquid that additionally lyses any red blood cells present.

Step 3: At the laboratory, the vial is then placed into the ThinPrep 2000 slide processor. A disposable polycarbonate filter is gently dipped into the cellular suspension and spins to create a current that breaks up blood, disperses mucus, and thus separates unwanted debris. The sample is then thoroughly mixed to create a more homogenous sample for analysis.

Step 4: A negative pressure pulse is used to draw fluid through the filter and collect a thin, even layer of cellular material. The instrument constantly monitors the rate of flow through the filter during the collection process to ensure uniformity thus preventing the cellular presentation from being too scant or too over populated.

Step 5: A modest vacuum removes the excess liquid, inverts the filter and effectively stamps the collected cells onto a slide. The sample spot created is circular in shape (ca. 1 cm in radius). The slides produced can then be stained and evaluated using methodology similar to a conventional PAP smear.

4.1.2.2 SurePath Sample Preparation

The SurePath method for slide preparation is greatly more labour intensive than other LBC techniques and involves several different phases that have been collectively termed the PrepStain system. Unlike vacuum filtration that separates cells based upon there size, this technique uses a cell enrichment process. The main steps involved in this preparation are listed below:

Step 1: Again gynaecological specimens are collected in the recommended manor using a broom-type cervical sampling device. However, the PrepStain method dictates that the head of the brush is placed into a vial containing SurePath preservative, a buffered 24% ethanol solution.

Step 2: Sample vials are vortexed for 15 seconds at 3000rpm to allow randomisation of specimens. Shearing forces of the vortex free cells and cell clusters from the specimen collection device and partially disaggregate cell clusters.

Step 3: By use of the instrument PrepMate®, the sample is mixed and subsequently removed from the preservative vial by a syringe that then layers the specimen onto a density reagent held within a centrifuge tube (Figure 3). The reagent used in the

PrepStain method is a polysaccharide solution with sodium azide added as a preservative. The cell suspension is then centrifuged through the density reagent for 2 minutes. Small particles and debris which are trapped above the interface between the supernatant preservative fluid and the density reagent are removed to enrich the clinical materials in the sample.

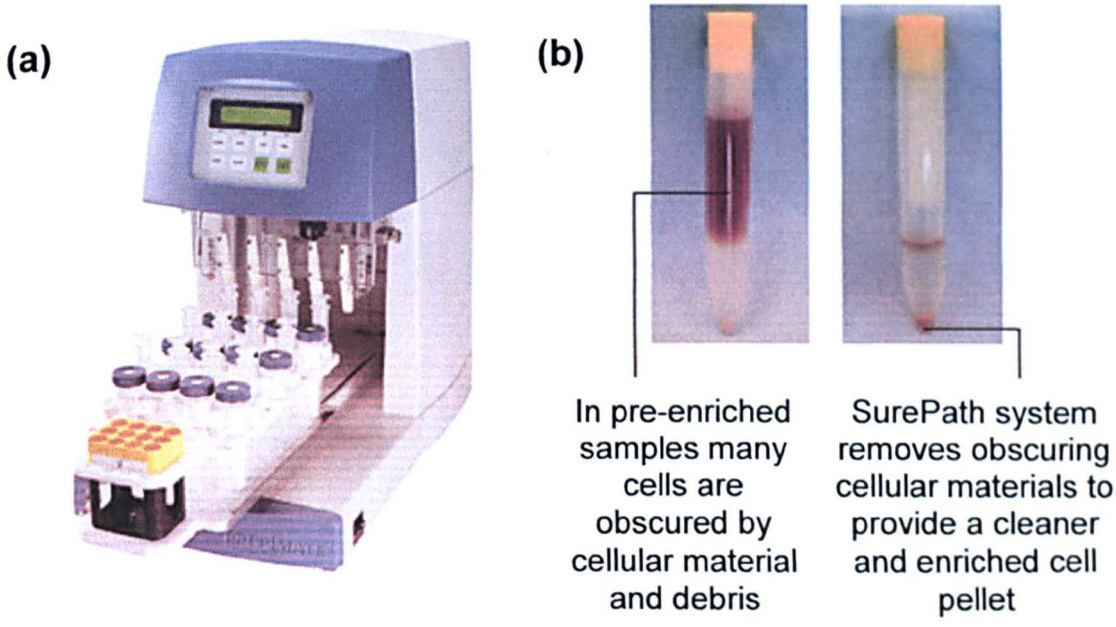

***Figure 3**: Pictures of the apparatus used for SurePath work. a) PrepMate instrument used to mix and layer the sample onto a density reagent. b) Example of cervical smear material before and after enrichment process. Taken from ref [11].*

Step 4: A second centrifugation for 10 minutes concentrates the diagnostic cellular material at the bottom of the tube. The remaining density reagent is then decanted, leaving the resulting enriched pellet of cellular material inside the centrifuge tube. The sample is then allowed to vortex ready for slide preparation.

Step 5: Next the PrepStain Slide Processor® is utilised, which performs the automated sample transfer and staining steps for the thin-layer preparation of cytologic materials on a microscopic slide (Figure 4). The instrument utilises a robotic arm and disposable tip assembly for aspirating and pipetting samples.

However, before this process can start, the microscopic slides are coated with a film of high molecular weight cationic solution. The resulting positive charge allows adhesion of diagnostic cytological materials to the slide throughout the slide preparation process.

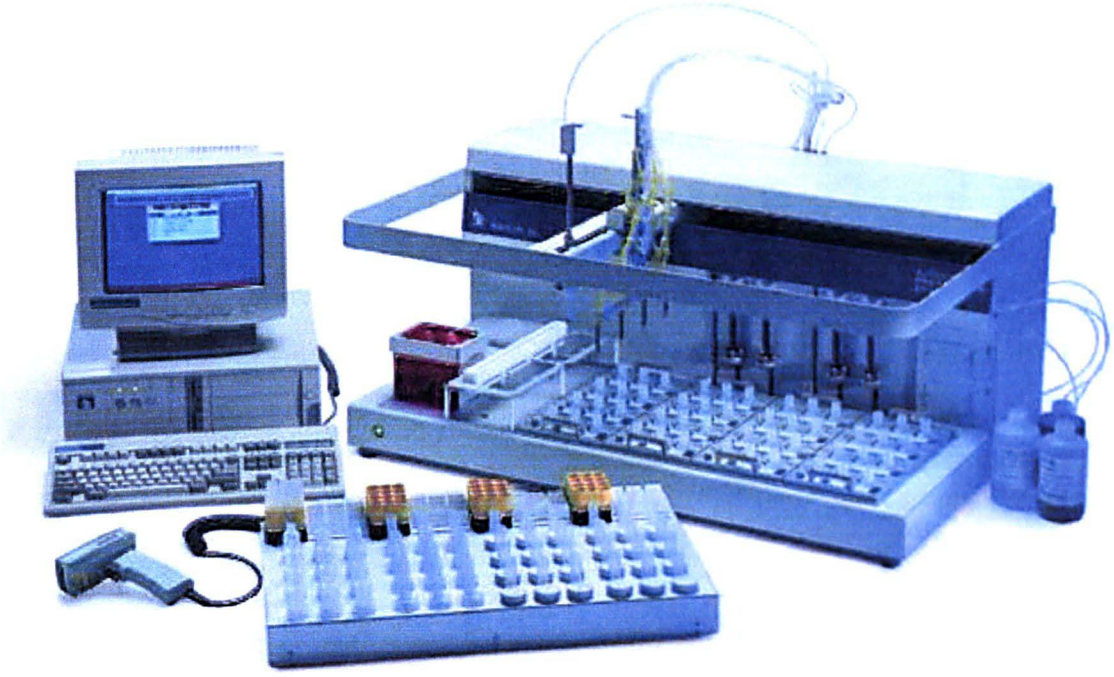

***Figure 4**: A picture of the apparatus used for SurePath work showing the PrepStain slide processor. Taken from ref [11].*

Step 6: The instrument initially resuspends the pelleted cell sample in 1000 µl of buffered deionised water and then mixes the resulting solution by flushing it in and out of the tube 8 times.

Step 7: Next, the PrepStain instrument aspirates 200 µl of the sample from the centrifuge tube and injects this into a settling chamber previously placed onto the coated microscopic slide. This allows the sample aliquot to deposit within a defined sample area of 13mm and avoids sample cross contamination (Figure 5). The tip is subsequently washed with 600 µl of buffered deionised water allowing the tube and remaining specimen to be discarded, or retained for adjunctive testing.

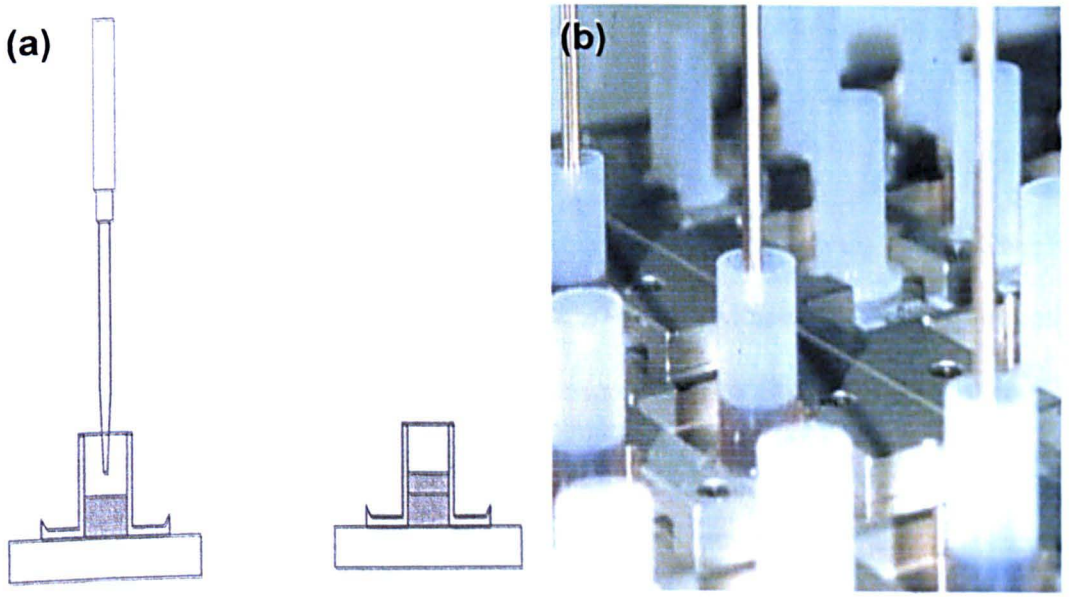

***Figure 5**: PrepStain slide preparation method. a) Schematic displaying the transfer of a sample into a settling chamber. The chamber holds the sample above the slide within a defined area enabling cellular transfer. b) Picture of the same process. Taken from ref [11].*

Step 8: The PrepStain instrument then adds a 600 µl alcohol wash to the sample and evacuates all remaining fluids. The sample is then allowed to dry for approximately 60 seconds.

Step 9: The last part of the automated process is a sequence of stain and rinse cycles. Stain and rinse cycles are essentially the same, all that varies from cycle to cycle is the reagent used and the duration of the pause

4.1.3 Liquid Based Cytology Method Development

The preparation of cervical smear material onto reflective slide for IR analysis has undergone three main routes during this study as the Thesis progressed. Two

methods of spectroscopic data collection were used and are described in chronological order. The choice of method was greatly dependent upon instrument availability during the project.

The first batch of cervical samples was prepared following the ThinPrep methodology. Cervical smear material was deposited into their recommended preservative solution PreservCyt®, and then transported to Derby Infirmary Hospital where they were prepared onto reflective slide using the ThinPrep instrument. All IR microspectral analysis of these samples was carried out utilising the IR beamline located at the Daresbury SRS laboratory. IR spectra were collected in a point by point fashion from individual cells that were clearly discernable and thus diagnosable by a cytologist. To take advantage of the increased signal to noise available with synchrotron sources, spectra were collected from spatial regions within individual cells that incorporated both the cytoplasm and nucleus. Large cells were examined using an aperture of 15 μm, and small cells with an aperture of 10 μm. Infrared spectra were collected in reflectance mode with a spectral resolution of 8 cm^{-1}. Dependent upon signal intensity, either 512 or 1024 spectra were coadded over the range 4000 – 650 cm^{-1}. Appropriate background spectra were collected in areas off the sample to ratio against the single beam spectra produced.

The second batch of cervical samples were prepared using the ThinPrep methodology. However, in this set of experiments the cervical smear material was deposited into a vial containing a 70% ethanol solution to preserve the cells. Further support for ethanol as a general solution for smear material was reported in the literature by Wood *et al.* [12], who compared ethanol with saline. They concluded

that the structural integrity of cells was preserved, the ethanol serving to rapidly dehydrate cellular material without the deposition of trace molecules, which could mask spectra. It also removed the need for fixing agents, which could distort spectra and also proved a good, inexpensive preservative of cells. Due to the use of this alternative solution, the ThinPrep instrument at Nottingham City Hospital was utilised as it was not limited by any solution protocol restrictions. This solution was chosen under recommendation from previous work undertaken by Ms Jodi Kent [13]. In this study, cervical tumour cells were cultured and prepared onto slide by the ThinPrep instrument using a variety of alcohol, saline, phosphate buffered saline (PBS), and PreservCyt® solutions. Initially, it was thought that infrared microspectroscopy would be used to establish any spectral distortions that may be occurring due to the solution type used. However, due to the inherently immature nature of tumour cells, the cell sizes ranged from 8-10 μm making IR spectra unattainable using a conventional FTIR instrument at that time. The synchrotron source at Daresbury laboratory had the capability to examine the cells but at the time of study showed poor signal to noise during allotted data collection runs due to beam instability. Conclusions were therefore drawn from visual analysis that assessed cell deposition. The slides prepared using this method were examined by use of the Spotlight Spectrum Imager.

A final small number of cervical smear slides were created by use of the SurePath methodology. In collaboration with researchers at PathLore, the instrument manufacturers, multiple slides were created from one collected sample at their Nottingham laboratory. To assess whether the high molecular weight cationic coating would adversely affect the spectra produced, slides were prepared both with

and without this proposed coating. Cells deposited onto these slides were again scrutinised by IR using the Spotlight Spectrum Imager, following the data collection criteria of previous experiments.

4.1.3.1 ThinPrep Preparation utilising PreservCyt® Preservative Solution

The application of LBC techniques for cervical smear sample preparation does undoubtedly give marked improvement in specimen quality and presentation by reducing blood, mucous, inflammation and other obscuring artefacts. A marked improvement in the preservation of cells was noticeable utilising the ThinPrep method and the PreservCyt® solution. Cell overlap and clumping normally encountered with the conventional PAP smear method is greatly minimised allowing individual cells to be visualised and examined with ease (Figure 6).

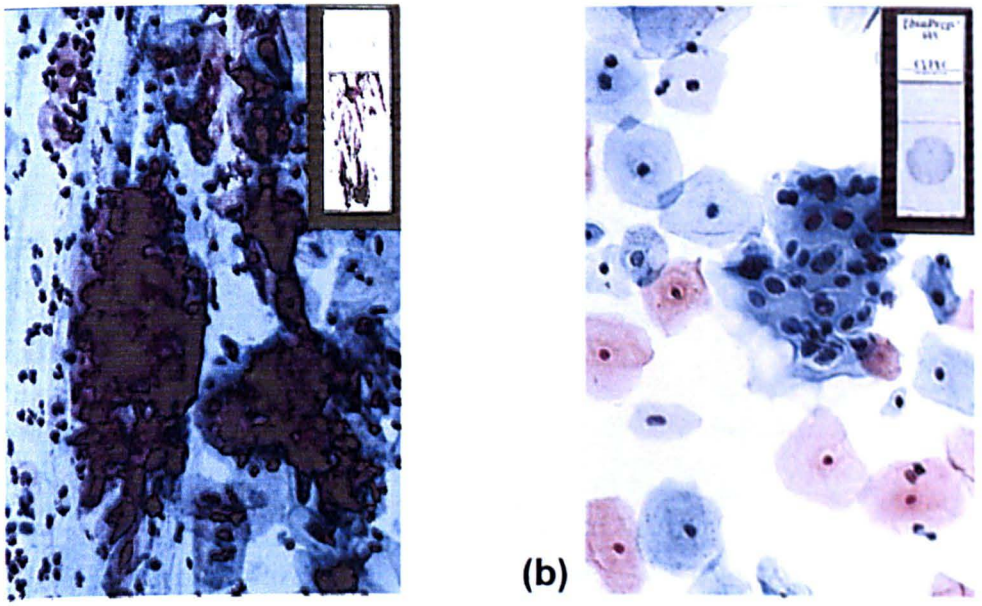

***Figure 6**: a) Conventional PAP smear preparation. b) ThinPrep preparation Taken from ref [10].*

However, after close examination of the data collected from individual cells prepared in this manor, a major artefact began to manifest itself in the spectra produced.

Figure 7 displays several overlaid spectra collected from single cells diagnosed as being normal in nature by a cytologist. It can clearly be observed that the amide I peak sensitive to protein concentration and structure within cells [14, 15], exhibited either a doublet or shouldered peak with maxima at 1646 and 1626 cm^{-1} respectively.

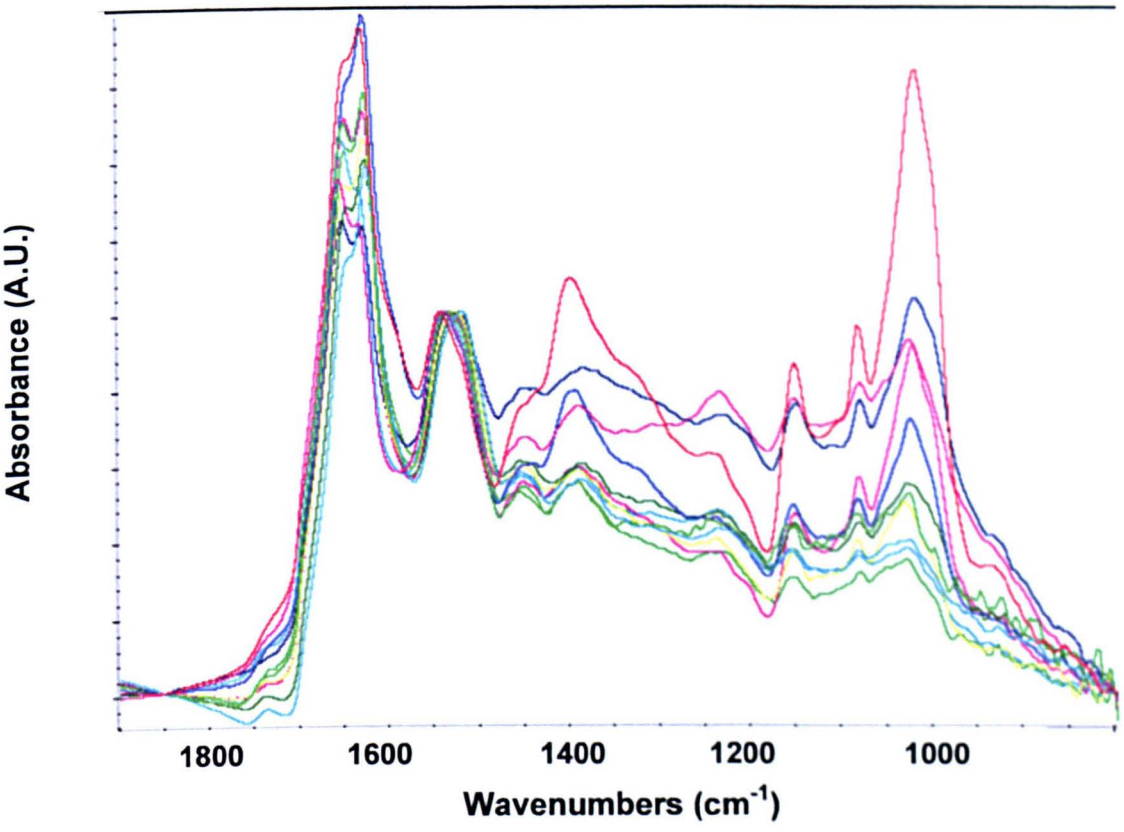

Figure 7: *Multiple overlaid IR spectra collected from individual cervical cells diagnosed as being healthy in nature. These cells were prepared using PreservCyt solution as a preservative. Note the splitting of the amide I absorption band.*

All samples examined had originated from patients displaying low-grade disease and were subsequently attending follow up clinics. Therefore, initial thoughts were that this splitting of the amide band could be a precursor for cancerous change, the protein side chains in effect changing from an α-helical to a β-sheet configuration with the onset of disease. However, this hypothesis was not verified in previous work examining cervical tissue sections (see sections 3.3.1 & 3.3.2). A variety of different tissue sections with contrasting diagnoses had been examined, and all

displayed a single peak in the amide I region found at 1646cm^{-1}, diagnostic of an α-helical configuration of the protein side chains. It was consequently concluded that this observed splitting of the amide I band was an artefact introduced into the spectra by the PreservCyt® solution. Unfortunately no common pattern could be easily found for this splitting, and the maximum intensity alternated randomly at either 1626 or 1646 cm^{-1}. This possibly suggested that the change we are observing is related to the amount of time the cells are exposed to the solution, i.e. the longer they are exposed, the greater likelihood they will change protein structures. The size or maturity of each cell could also quite easily have an effect upon the speed to which they are changed by the solution. Unfortunately possible control experiments to verify this theory were not possible due to instrument and time constraints.

4.1.3.2 ThinPrep Preparation utilising 70% Ethanol as a Preservative Solution

Cervical smear material was prepared onto reflective slides using this method and again this approach displayed a marked improvement upon conventional smear presentation. The amount of cell clumping found upon the slides was reduced, but did not show the same efficiency obtained using the PreservCyt® solution. Cell damage was also observed in some of the prepared slides, especially around the surrounding edges of the sample spot. Closer examination at high magnification revealed that the cytoplasm of the cells was effectively being stripped away leaving individual nuclei scattering around the sample. This effect is likely to be due to the ethanol solution dehydrating the cells to such an extent that they have become extremely brittle and easily fractured during the vacuum filtration process. A

mucosal deposit surrounding the cells was also apparent in a small number of cases, and is due to ethanol's inability to effectively separate the diagnostic cells when mucus levels are high within samples. However, in the overwhelming majority of the sample area, monolayer coverage was evident, ideal for IR microscopy of individual cells. Figure 8 displays a number of overlaid spectra collected from individual cells. It can be seen from the spectra, that the splitting of the amide I band observed when using the PreservCyt® solution, is no longer apparent. This preparation does not appear to introduce an artefact to the spectra and it is hoped that this would simplify any future multivariate analyses.

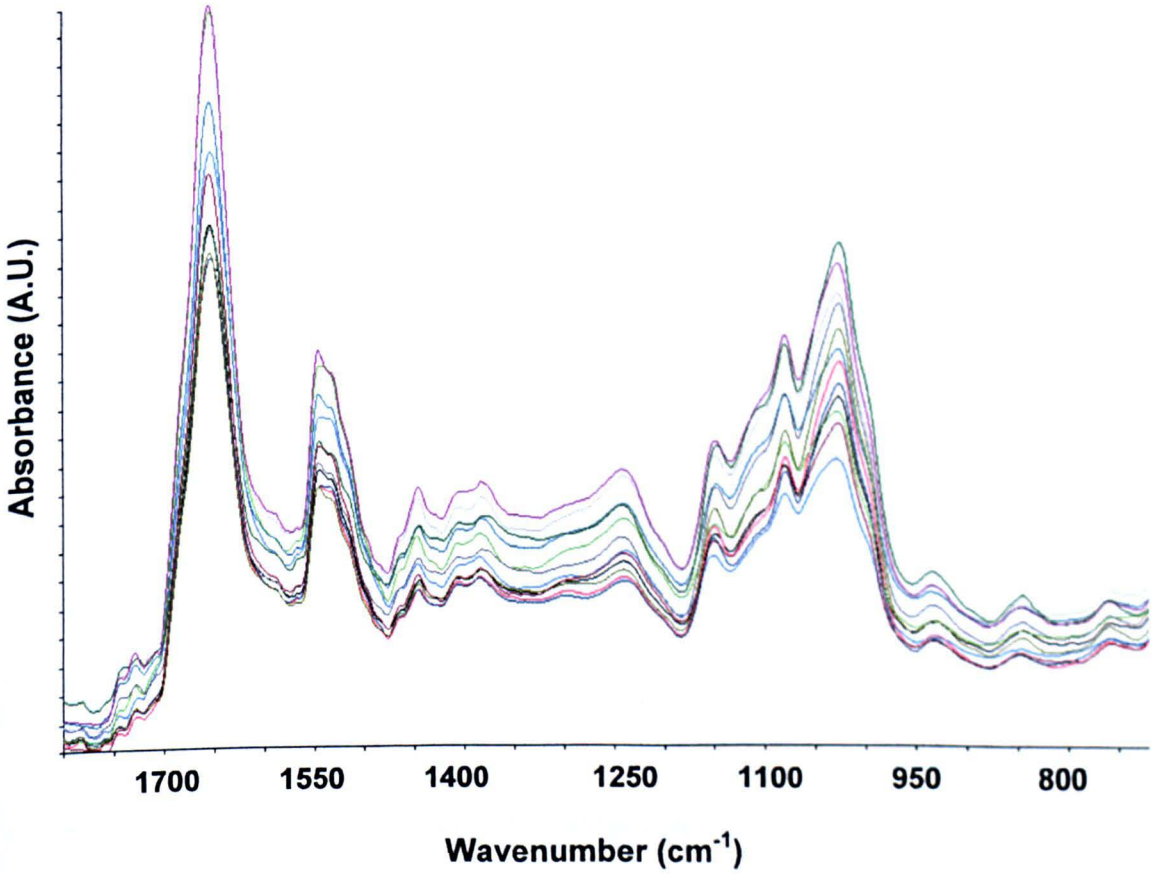

Figure 8: *Multiple overlaid IR spectra collected from individual cervical cells diagnosed as being healthy in nature. These cells were prepared using 70% Ethanol solution as a preservative.*

4.1.3.3 Surepath Preparation

This type of LBC preparation was only tested on one collected smear sample, but showed some distinct advantages to the ThinPrep methodology. The proprietary cell enrichment process resulted in the obscuring cellular material and debris from blood, mucus, and inflammatory cells being significantly reduced. The density gradient centrifugation process separates cellular material from debris using size, shape, weight and density, unlike vacuum filtration which separates based on size only. A significant increase in the number of cells on the slides was noticeable. Cell aggregation was negligible and good monolayer coverage was observed. This is again ideal for IR microscopy of individual cells. This technique also allows easy manipulation of the cell population found within the sample spot. By suspending the cell pellet produced after centrifugation in different amounts of deionised water, the number of cells can be tailored to best suit experimental requirements, whether those be a sparse or packed cell coverage. The creation of a slide with no cationic coating adversely affected the transfer of cells, resulting in only a small number of cells being deposited onto the slide. The application of such a coating would appear to be a necessity for cellular fixation onto the microscope slides. The IR analysis of individual cells collected from slides prepared in this manor showed no obscuring artefacts that could hinder subsequent multivariate analyses.

4.1.3.4 Liquid Based Cytology Method Development Conclusion

The application of liquid based techniques enabled cervical smear material to be presented in a manor ideal for IR microscopic analysis of individual cells.

Contamination from non-diagnostic cells which could obscure and hinder IR analysis is dramatically reduced. Cells are presented on the microscope slides with monolayer coverage permitting individual cells to be easily located and scrutinised. However, the solutions used to preserve the exfoliated cellular material can introduce changes in the biochemical makeup of the cells. These unwanted changes can then manifest themselves as artefacts in the spectra that can ultimately complicate any subsequent multivariate analysis. From the two types of alcohol solution used, ethanol gave the best results in that cells were adequately preserved and produced no artefacts in the spectra. However, the 24% Ethanol solution utilised in the SurePath method gave the best overall results, preserving the cells without any noticeable cell damage. Although both techniques showed marked improvements upon conventional smear preparation, the SurePath system gave the best presentation of cells with high cellularity. The ability to tailor the cell population created on the slide, coupled with the use of a density gradient to more effectively separate non-diagnostic cells, would make the technique a preferred choice for spectroscopic applications.

4.2 Instrumental

Infrared spectroscopy has proved over the past decades to be an extremely potent analytical tool for the analysis of biologically active materials [16 – 20]. When biological molecules are exposed to radiation in the mid-infrared region of the electromagnetic spectrum (400 – 4000 cm^{-1}), they exhibit characteristic absorptions from the excitation and vibration of bonds within the molecules. Creation of reference spectra for biochemical constituents such as proteins, lipids and nucleic

acids can be useful when assigning peaks, but the mixture of many different biomolecules within a cell will ultimately lead to very complex convoluting spectra. Therefore it is more useful to detect subtle changes in peaks and their positions rather than assigning a spectral feature to a particular cell constituent or cell type. Thus spectra that provide high signal to noise are essential to characterise these very small biochemical changes that occur between different cell types. Animal cells can also range in size between 5 – 50 μm, making the more conventional macroscopic study of samples insufficient for detailed and unpolluted characterisation of individual cells and their constituents. FTIR microspectrometry however, i.e. the coupling of microscopy and FTIR spectrometry, is proving a potent new technique that can provide spatially resolved spectroscopic information from very minute quantities of microscopic structures within a sample [20,21]. Such techniques have therefore been adopted in this study to examine both single exfoliated cells and tissue specimens. Three types of instrumentation have been utilised, including the Nicolet Continuum, Perkin Elmer Spotlight Imager and a Nicolet Nic-Plan microscope that was coupled to a synchrotron source. These separate instruments will be fully described in the following sections; however basic FTIR theory will not be discussed.

4.2.1 Nicolet Continuum FTIR Microspectrometer

The apparatus is comprised of a Nicolet Nexus 870 FTIR spectrometer (Nicolet Instruments, Inc. Madison, USA), fitted with a KBr beamsplitter. The spectrometer is additionally coupled to a Nicolet Continuum microscope that comprises its own liquid nitrogen cooled mercury-cadmium-telluride (MCT) single element detector (100 μm). The microscope uses the same apertures and optics for both the infrared

and visible light, where the visible optical train is parfocal and collinear with the IR radiation [22]. The system utilises a silicon carbide 'Globar' source, which is heated to ~1500K, providing incident radiation to the interferometer that shows similar emission characteristics across the frequency range. The signal from the interferometer is modulated and channelled to the sample by a series of mirrors and optics. To reduce the effect of stray light distorting the spectra, and the problems of diffraction causing optical aberrations, a single 'Reflex^{TM}' aperture is used. This set up allows the radiation to be directed by mirrors onto the sample and back through the same aperture before reaching the detector.

The MCT detector works on the principle that the absorption of IR photons by the photoconductive detector element, will give rise to the promotion of electrons from the valence band of the material, to the conduction band, resulting in the flow of current when a voltage is applied. Therefore it can be used proportionally as a measure of signal intensity. The detector is operated at very low temperatures, ~77K, and thus requires cooling by use of liquid nitrogen. The remaining components are all electrical in nature, an amplifier, analogue to digital converter and a computer that allows the processing of the signal to create a spectrum. The main limiting factor of this apparatus is the problem of diffraction, where the aperture dimensions are limited. The diffraction limited spatial resolution d is defined by the equation:

$$D = \frac{0.61\,\lambda}{N.A.} \qquad \textbf{(Equation 1)}$$

This is where λ is the wavelength of the incident radiation and N.A. is the limiting numerical aperture. Taking into account that the mid infrared radiation exists between the wavelengths of 25 – 2.5 μm, d is approximated to 10 μm. Because the radiation is now being passed through small apertures the Jaquinot advantage is reduced, and due to significant scattering of light, the smallest spatial resolution that can be studied with a reasonable signal to noise is 20 μm.

4.2.2 Perkin Elmer Spotlight Imager

The Perkin Elmer Spotlight Imager (Perkin-Elmer Corp., Sheldon, Connecticut) is also a FTIR microscope, similar to the Nicolet Continuum instrument, but alternatively comprises a dual set of detectors. The microscope is equipped with both a 100μm single element (MBMCT) detector and a NBMCT array detector. When operated in array mode, the system utilises a 16 x 1 element (400 μm x 25 μm) linear array of small area narrow band (4000 – 720 cm^{-1}) detectors that provide significant reduction in detector noise and thus improved signal to noise. Each detector has its own isolated gold connection used to perform their own signal processing, where all 16 channels are continuously sampled. In comparison, the Spotlight provides the IR radiation from below the sample when in transmission mode, but like the Continuum from above the sample in reflection mode. This is simply carried out by the movement of a mirror within the instrument that sends the IR radiation a different route. The optical axis of the instrument is set on the single element detector. However, the array is also well illuminated. When collecting data in array mode, the electronic stage is moved to compensate for the separation of the array from the optical axis, and then moves in 25 μm steps below the array.

Therefore the array is effectively sweeping the sample at a speed defined by the spectral resolution and the number of scans per pixel. A direct link between the interferometer and the electronic stage only allows the stage to move when the interferometer has reached its end point. A map of the sample area is then built up by raster scanning across the sample in both the X and Y planes. The array can also be used to examine the sample with a 6.25 μm pixel size. This is achieved by use of a Z fold tube that dips a 4X magnification mirror into and out of the beam. Unfortunately, more detailed information on the specific design of the array has not been made available, but the obvious advantages of this instrument are clear. It has the capability of scanning 16 different spatial areas at once, enabling large samples areas to be examined rapidly with signal to noise levels that are effective for sample characterisation at the microscopic level.

4.2.3 FT-IR Microspectroscopy utilising a Synchrotron Radiation Source

The dimensions of animal cells are typically comparable to the minimum resolvable distance of IR microscopes caused by the limitations of diffraction, c.a. 10 μm. However, a Synchrotron Radiation Source (SRC) can enable the collection of IR spectra at these spatial sizes with significantly higher signal to noise, where it is estimated that the brightness of the IR radiation produced can be up to 1000 times more intense [23]. In this study, the IR beamline located at the UK SRS laboratory in Daresbury was utilised, where they coupled their synchrotron source to a Nicolet Nic-Plan FTIR microscope. This instrument is very similar to the Nicolet Continuum previously described, but is configured slightly differently, where two apertures are used for transmission data collection. These both require focusing,

where the first dictates the area illuminated and the second captures the transmitted radiation, filtering out any unwanted light that may cause optical aberrations. Since the beam is highly collimated (approximately 20 x 30 μm) and hence more brilliant than a conventional source, the signal to noise that can be achieved is significantly higher.

4.3 FTIR Microspectral Data Collection

Both single point and mapping methods of data collection were utilised in this study of human tissues and cells. Tissue specimens were cut into thin sections and mounted onto IR transmissive BaF_2 discs suitable for spectroscopic analysis. These types of sample were analysed solely by the collection of transmission – absorption spectra using mapping techniques. Exfoliated single cells were alternatively prepared onto 'low e' substrates. These consist of a glass slide with a thin silver coating and a transparent overcoat to protect the silver layer. This type of substrate is completely reflective in the mid-infrared spectral region. An IR beam passing through a thin sample is reflected by the silver layer and subsequently passed back through the sample, thus experiencing twice the attenuation of a single pass. Additional advantages of these substrates are that they are also close to transparency in the visible, and can thus be examined via conventional light microscopy. Individual cells prepared onto these substrates were analysed by the collection of reflection – absorption spectra using both single point and mapping methods.

4.3.1 Tissue Section Analysis

Prior to IR analysis, tissue sections were removed from the cryovial and passively warmed to room temperature. The circular BaF_2 discs were then fixed into specially designed steel sample holders, similar in shape to a conventional glass slide, and positioned onto the sample stage of the infrared microscope utilised at that point in the study. A visual image was then acquired from the entire tissue section via a charge coupled device (CCD) camera that was referenced against a scribed mark previously etched onto the barium fluoride disc. Magnified visual images are collected under white light LED illumination, and subsequently quilted together to create a mosaic picture that is of arbitrary size and aspect ratio. By use of the adjacent and diagnosed H&E stained tissue section, areas of interest were then located upon the unstained sample that incorporated several different tissue types and effectively characterised the morphological infrastructure of the tissue section being examined. Large infrared maps were then obtained from these sites of particular interest at high resolution, and from the entire tissue section at a lower resolution if time permitted. To gain high signal to noise spectra necessary for effective multivariate analysis the spatial resolution and thus pixel size in our multivariate images was determined by the capabilities of each instrument. All spectra were collected in transmission mode with a spectral resolution of 8 cm^{-1}, over the spectral range 4000 – 720 cm^{-1}. The remaining instrumental parameters used for each instrument are listed in Table 1. It must be noted that the spatial resolution of the Perkin-Elmer Spotlight Imager when collecting at a pixel size of 6.25 μm is actually c.a. 12 x 12 μm, being limited by the diffraction limit [24,25]. The instrument is actually collecting data from a 12μm spatial area, but is being stepped by 6.25 μm,

Instrument	Resolution	Sampled Area	Number of Scans
Continuum	High	25μm x 25μm	1024
	Low	50μm x 50μm	512
Spotlight	High	6.25μm x 6.25μm	16
	Low	25μm x 25μm	8
Synchrotron	High	15μm x 15μm	512
	Low	25μm x 25μm	256

Table1: *Spectral data collection variables for each instrument utilised*

effectively over sampling by 2 fold. To reduce spectral contributions from atmospheric carbon dioxide and water vapour, the microscope was additionally purged with dry air which included a purge ring that surrounded the sample. The commercially available Spotlight instrument was not sufficiently purged, and a specially designed Perspex box to surround the entire microscope and sample stage was constructed to address this. Spectra collected were fast Fourier transformed using strong apodization to yield single beam spectra. An appropriate background spectrum was additionally collected off the sample to ratio against the single beam spectra. These ratioed spectra were then converted to absorbance, with each spectrum containing 821 data points (4 cm^{-1} data point interval). Acquisition time varied between several minutes to several hours dependent upon sample size.

4.3.2 Exfoliated Single Cell Analysis

Prepared slides were placed onto the sample stage of the microscope and a visual image taken from the entire sample spot created by liquid based cytology. Locations upon the sample spot where cell deposition was ideal, i.e. individual cells were

clearly definable, were located and an additional visual image taken at high magnification. Both images were reference against an etched mark on the slide to enable the accurate relocation of areas of interest and individual cells that were to be examined. Reflectance – absorbance spectra were then collected from individual cells using both point and mapping methods of data collection. Due to the high signal to noise requirements necessary for such data collection, spectra being highly contributed to by dispersive artefacts, only the synchrotron source and Spotlight instrument was utilised for these studies. Time at the synchrotron source focused toward the collection of spectra via point mode using a number of aperture sizes that varied between 10-25 μm to accommodate for the different cell shapes and sizes that were encountered. Effort was made to incorporate both the nucleus and cytoplasm in each case. A spectrum was then taken by co-adding 512 interferograms at 8 cm^{-1} spectral resolution. Experiments undertaken using the Spotlight instrument were aimed to examine as large a sample area as possible and thus a significantly higher number of cells. This objective was administered as in early studies in had become apparent that the number of abnormal cells on a slide can often be significantly low and sparsely orientated on the sample spot. Large square infrared maps were therefore collected from sample areas ranging from 1000 – 5000 μm in size. In these experiments, 16 interferograms per pixel (25 x 25 μm sample area) were coadded over the spectral range 4000-720 cm^{-1}. All spectra collected by either of these techniques were fast Fourier transformed using strong apodization to yield single beam spectra. An appropriate background spectrum was collected off the sample spot to ratio against the single beam spectra. These ratioed spectra were then converted to absorbance. Acquisition time varied between several minutes to several hours

4.4 FTIR Spectral Data Processing

All infrared micro-spectral data was uniformly pre-treated before undergoing further multivariate analysis. Small possible contributions to the spectra from atmospheric water vapour and carbon dioxide were removed by atmospheric correction algorithms integrated into both the Perkin Elmer Spotlight and Nicolet software. IR spectra collected from human tissues and cells may sometimes display sloping or curved baselines. These distorting affects may arise due to a variety of reasons. Contamination to the spectra may manifest itself through the superposition of dispersive and absorptive line shapes cause by the collection of unwanted stray light [26,27]. Another reported cause of distorted line shapes is the effect of Mie scattering from the nucleus of cells [28]. The nuclei of non-proliferating human cells contain tightly bound DNA and RNA strands reportedly making them almost opaque to IR radiation [28-30]. Applying a 6 base point linear interpolation to all spectra reduced the effects. Baseline points used in this process were located at 4000, 3744, 2200, 1836, 876 and 720 cm^{-1} respectively. Finally, to negate intensity differences caused by irregularities in sample thickness and cell density, spectra were uniformly normalised. This was achieved by scaling spectra such that the sum squared deviation over the indicated wavelengths (4000-720 cm^{-1}) equals unity (also known as vector normalisation). All data processing and subsequent multivariate analyses were performed using algorithms that operate on top of MATLAB version 6.5, release 13.0.1 (Mathworks, Natick, MA, USA).

4.5 Chemometrics

4.5.1 Introduction

Whether the aim of a spectrocopist is to determine a samples composition or identify a single species that may be present, the use of statistics to help extract the desired information from the experimental data is now common practice in all modern analytical techniques. In this study, infrared spectroscopic data has been collected from human tissues that produce very complex vibrational signatures. The interpretation of this type of data is not always straight forward and ultimately relies upon the detailed understanding of the tissue constituents. For the overwhelming majority of human tissues, the IR spectrum produced can be directly approximated to the summation of lipids, proteins and nucleic acids, the basic building blocks for all animal cells. These species give multiple and broad absorptions across the mid-infrared region (4000 – 1000 cm^{-1}), making the assignment of individual chromophores to specific tissue components an empirical process. To remove this subjectivity, statistical techniques can be applied to help interpret the data produced. Univariate type analyses commonly used in analytical spectroscopy are no longer sufficient in this scenario due to the inherent complexity of tissue spectra. However, another branch of chemometrics termed multivariate analysis can enable the manipulation and investigation of data that contains multiple variables, such as an IR spectrum.

Multivariate analyses can be separated into two main types, those that are supervised or unsupervised. Supervised pattern recognition methods utilise information that is already available for the sample you are examining. Thus a training set holding

information characteristic for each parent class can be utilised to identify a discriminant function by which new unlabelled data can be recognised and further classified into one of the parent classes. Unsupervised pattern recognition methods alternatively use no previous knowledge of the sample analysed and search for similarities within the data to characterise them. These types of unsupervised analysis have therefore become increasingly used in analytical spectroscopy because of this advantage. Underlying patterns hidden within extensively large and complex datasets can be identified that were previously undetectable using univariate or bivariate type analyses alone. Early experiments that utilised such methods for the analysis of IR spectra collected from animal cells were those undertaken by Naumann [31]. He utilised Hierarchical Cluster Analysis (HCA) to classify spectra that were collected from bacteria and was able to group them according to their bacterial strain. As a consequence, methods for rapid bacteria identification using IR spectra are presently being developed. In light of the reported sensitivity and successful application of such techniques, the overwhelming majority of spectroscopic data collected from biological material have been analysed using these types of multivariate analysis.

To present a multitude of unsupervised methods have been utilised for spectroscopic data analysis of human tissues. These techniques have included Principal Component Analysis (PCA) [32 – 35], Hierarchical Clustering Analysis (HCA) [26, 36-41], K-Means and Fuzzy C-Means Clustering (KM, FCM) [42 – 45] and Simulated Annealing Fuzzy C-Means Clustering (SAFCM) [46]. These studies indicated that each multivariate technique could to a degree, be applied to disease diagnosis using spectroscopic data. However, there is still a general lack of

comparative tests between techniques. A full spectrum of unsupervised multivariate techniques have therefore been applied to spectral datasets collected from human tissues in this study. A comprehensive and detailed comparison between alternative techniques for tissue discrimination could therefore be achieved. These studies also include the application of a newly developed PCA-FCM Clustering hybrid and a novel PCA-FCM merge method algorithm that automatically defines the best amount of clusters that describe a dataset. All multivariate analyses utilised in this study are described in greater detail in the remainder of this chapter.

4.5.2 Principal Component Analysis (PCA)

Principal component analysis (PCA) is a technique widely used for scrutinising multivariate type data [47]. The aims of applying such a technique to very large and complex datasets are essentially two fold. Initially, the analysis involves the rotation and subsequent transformation of the original, n, axes that describe the variables found within the dataset. This process is carried out in such a way that the newly created axes now lie upon paths that describe the maximum variance within the dataset and are orthogonal or uncorrelated to each other. As a consequence, each additional axes or 'principal component' created will account for less and less variability. It is now typically the case that the number of principal components, p, that are required to describe the majority of data variance is less than n. This type of analysis can therefore dramatically reduce the dimensionality of a dataset. After principal components are calculated the analysis can now reveal those variables, or combinations of variables that best describe patterns found in the data.

In this section, the use of PCA to detect underlying patterns in spectroscopic data will be discussed utilising the simplified dataset shown in Table 2. Infrared spectra were collected from 6 different tissue samples and further normalised to the most intense peak, removing possible tissue thickness effects. The resulting spectra were subsequently reduced to 4 discrete variables by extracting the response values found at wavelengths 1700, 1650, 1600, 1550cm^{-1} for each spectrum respectively. To aid visualisation of the resulting reduced spectra the data has been plotted in Figure 9. The resultant 6 x 4 response matrix was then subjected to Principal Component Analysis.

Wavelength (cm-1)	1	2	3	4	5	6
1700	0.00	0.01	0.00	0.00	0.00	0.00
1650	0.50	0.52	0.51	0.02	0.03	0.01
1600	0.05	0.02	0.01	0.50	0.48	0.52
1550	0.00	0.00	0.00	0.10	0.15	0.13

Table 2: *Extracted absorbance values of 6 different IR spectra at wavelengths 1700, 1650, 1600 and 1550cm^{-1} respectively.*

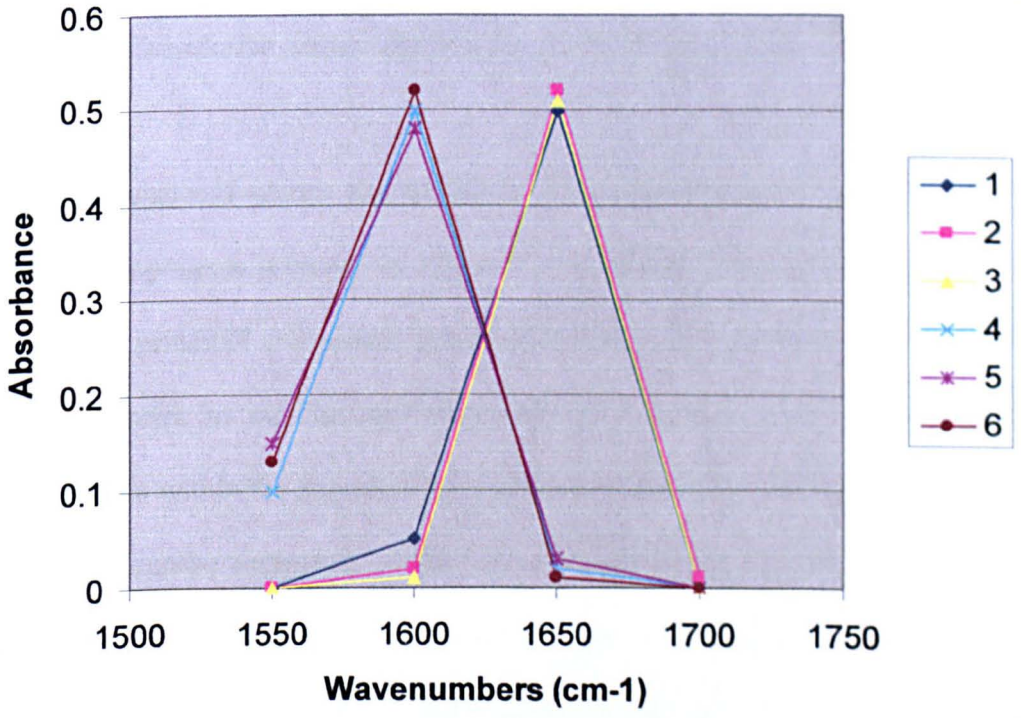

Figure 9: *Simplified example dataset plotted across the wavelengths extracted.*

The scree plot shown in Figure 10 describes the amount of variance each successive principal component is accountable for. The calculated first principal component now accounts for over 99% of the total variance from the original spectral dataset.

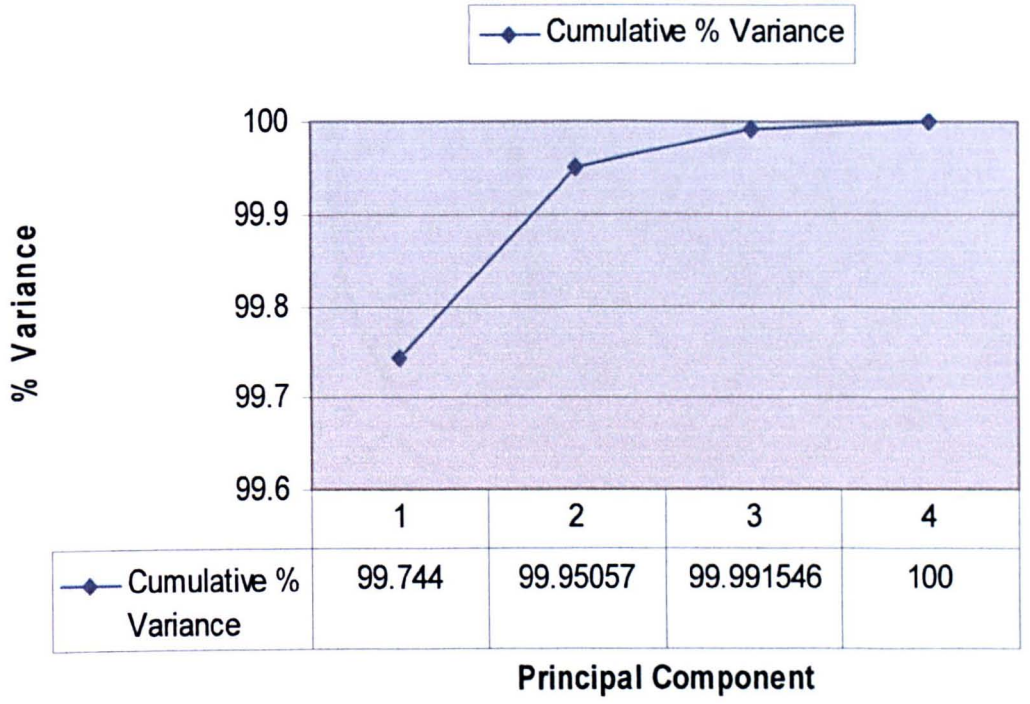

Figure 10: *Cumulative percentage plot for calculated principal components*

A three dimensional scatter plot of the tissue spectra projected onto the first three principal components is shown in Figure 11. Studying this plot it is clearly evident that the first principal component gives clear separation between two sets of points, allowing spectra to be clustered into two main groups. To help identify the characteristics within the data that cause this separation, the loadings of the principal components can be examined. These can be described as the principal component axes, or eigenvectors, as functions of wavelength. More simply, they highlight the weights that are given to each spectral data point in each of the original spectra. Figure 12 displays the loading plot for the first principal component in our analysis.

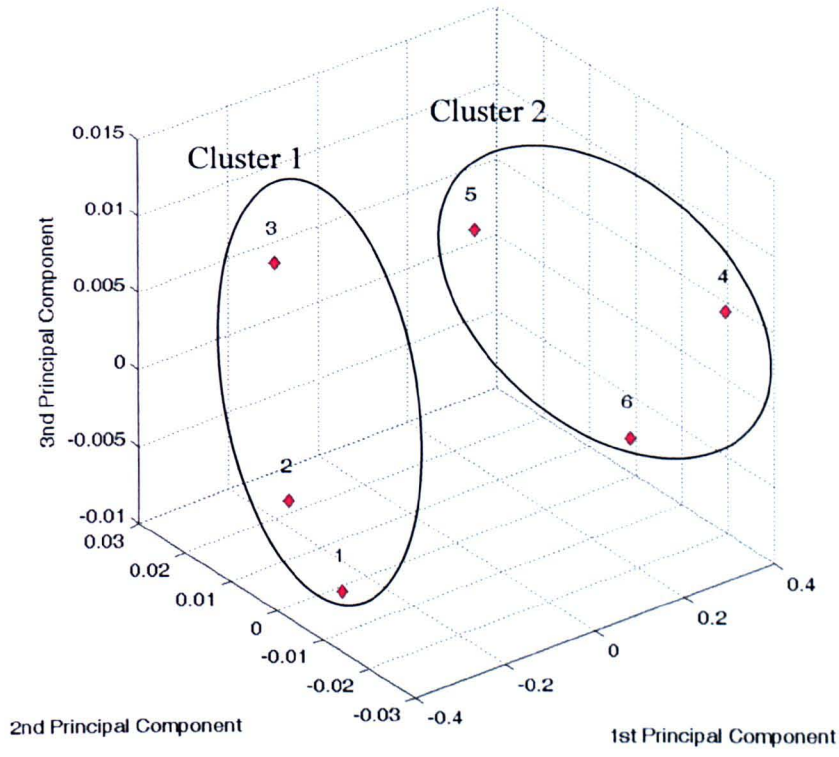

Figure 11: *A three-dimensional scatter plot of the reduced tissue section spectra projected onto the first three principal components. The numbers indicate the spectrum to which that data point belongs to. The spectra form two main clusters.*

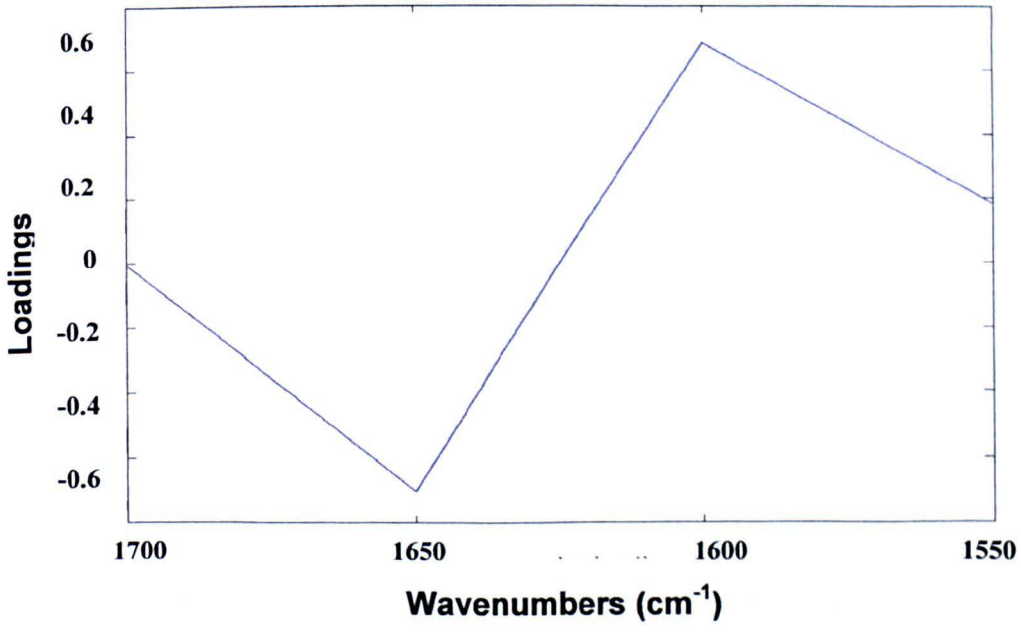

Figure 12: *Loadings plot for the first principal component*

Separation between samples can normally be identified in a loadings plot by the appearance of a strong positive or strong negative weighting at a point in the spectrum where the difference is occurring. Examining the loadings plot in our example, strong negative and positive weighting occurs at 1650 and 1600 cm^{-1} respectively. This clearly indicates that the spectra are being clustered dependent upon the position of the band in this region of the spectrum.

Principal component analysis can also be utilised for imaging purposes. From a statistical point of view, spectroscopic data with familiar features will have high correlations with each other, and vice versa. Therefore, an image can be constructed for each PC by applying a false colour weighting to each spectrum contained within a spectroscopic map or image. By use of this colour ranking, pixels on the created image will now reflect the intensity or correlation of each spectrum to that PC. These PC images now enable the identification of regions on a sample that are best described by that component, providing contrast between different spatial areas.

4.5.3 Multivariate Curve Resolution (MCR)

The complexity of FTIR datasets has lead to the development and use of a number of data resolution/unmixing methods called self-modelling curve resolution. One such method, termed Multivariate Curve Resolution (MCR), has been successfully applied in the analysis of spectroscopic image data [48-50]. The aim of this type of analysis is to separate the total spectral response from a large dataset into two factor modes, one that describes the pure spectra and the other the pure intensities. In more practical terms, these modes correspond to the real spectra of the components present

within the data and the uncalibrated concentration of each component for each spectrum. In practice this is achieved by applying Principal Factor Analysis (PFA), which outputs a number of factors and scores characteristic of the data. These are then transformed into pure spectral components and concentration scores by means of a constrained least-squares minimisation (LS) process. A non-negativity constraint was used to create factors with all positive attributes, given that a negative spectroscopic band would make no physical sense. The MCR solution can therefore enable the construction of false colour intensity images for each component factor produced in the analysis. The choice of factor number that best describes a dataset is another complex issue and can be identified by the use of indicators established in work by Malinowski [51,52]. The complete algorithm used for MCR analysis is described pictorially in the flow diagram shown in Figure 13.

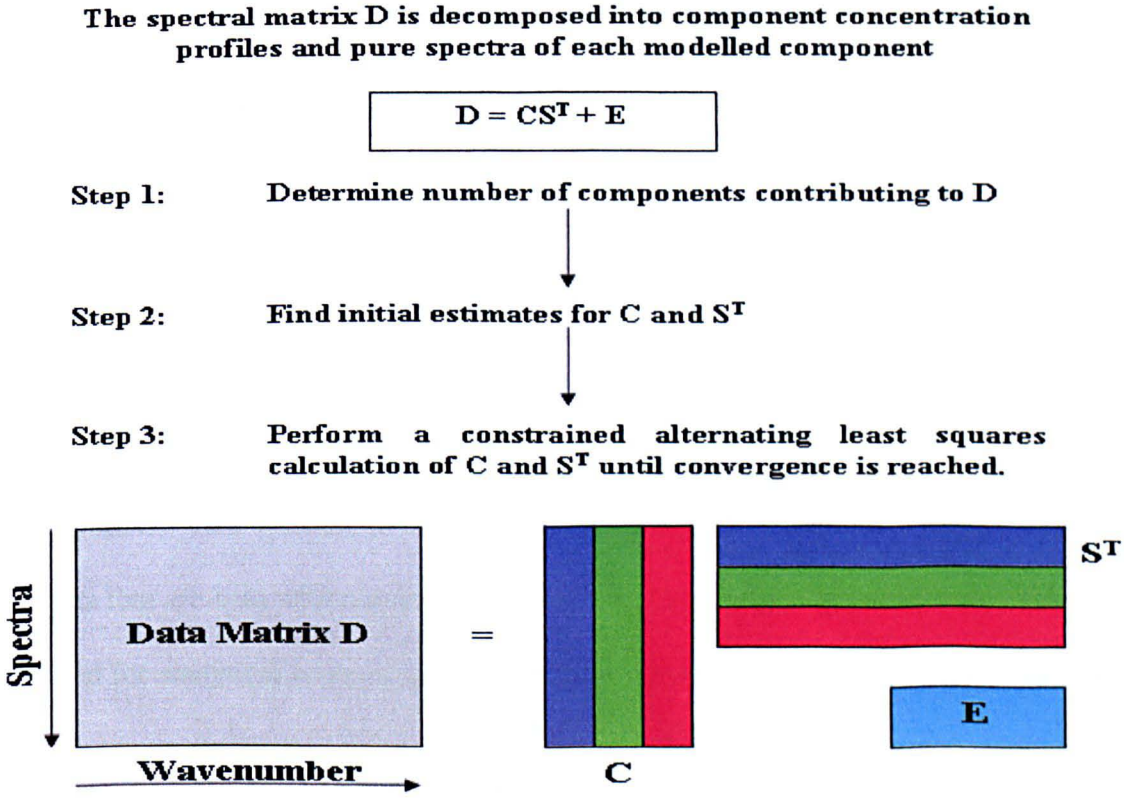

Figure 13: *The algorithm utilised for MCR analysis described pictorially*

4.5.4 Unsupervised Clustering Techniques

The aim of clustering techniques is to group a given set of unlabelled data into a number of predefined clusters so that data held within the same group are as similar to each other as possible, and data held within different groups are as dissimilar as possible. Algorithms used to achieve this initially convert the original or suitably processed experimental data into a matrix of dissimilarity or similarity measures [53]. These measures now describe the difference or similarity that is found between each sample held within a dataset. The algorithm then proceeds to cluster the data into groups so that a minimal separation is found between data held within a cluster, whilst also ensuring a maximum separation between clusters is achieved. It must be noted that the output from clustering processes, such as cluster membership or number of clusters produced, is dependent upon the similarity measure used and how they are applied. A full understanding of these processes is therefore essential to allow meaningful interpretation of the results.

4.5.4.1 Hierarchical Cluster Analysis (HCA)

Hierarchical Cluster Analysis (HCA) is a common technique employed for pattern recognition. This method utilises a similarity or distance matrix to cluster similar objects that are held within a dataset. A common similarity or association coefficient utilised for analytical analysis is the correlation coefficient. Alternative measures for similarity are rarely used being poorly defined and difficult to apply mathematically. Such measures will therefore not be discussed being inappropriate for spectroscopic

data analysis. Only similarity measures that best suit spectroscopic data will be discussed in this section. To help fully understand and visualise the basic steps of hierarchical clustering, a simple example using correlation coefficients as a similarity measure will be used to demonstrate the procedures involved. To appreciate how a correlation coefficient matrix is first calculated and subsequently utilised for cluster analysis, we must first understand the basic principles of covariance. The data shown in Table 3, displaying trace metal concentrations of soil sampled from different locations, will act as a simplified example to demonstrate how these processes work in practice.

	A	B	C	D	E	F
Cadmium	10.00	9.80	9.70	3.00	2.70	2.60
Copper	3.00	2.80	2.60	7.50	7.40	7.20
Lead	4.00	3.90	3.80	8.00	7.80	7.80
Nickel	4.10	4.00	3.70	9.50	9.20	9.30
Magnesium	3.00	2.80	2.60	9.80	9.70	9.70

Table 3: *Trace metal concentration of soil sampled from 6 alternate locations (A – F), expressed in mg kg^{-1}. The measurements in this matrix were manually generated to aid visualisation of a 2 cluster pattern within the data.*

For a single variable, the distribution around the mean value is classically described by its variance (Equation 2). By expanding the calculation to assess the shared variability between variables, using a common mean, the spread of multivariate data can also be determined. This measure of interaction between variables is more commonly termed covariance and is defined in Equation 3. To help understand the steps involved in this calculation, the covariance between site A and B is described in Table 4. The full variance–covariance matrix for our example dataset is shown in Table 5 and can be said to have diagonal symmetry, whereby the covariance between

site A and B is identical to that between site B and A. Variance for each variable, or sample site in this case, lie along the diagonal of the matrix.

$$S^2 = \sum_{i=1}^{n} (X_i - X_{im})^2 \Big/ (n-1) \qquad \textbf{(Equation 2)}$$

$$Cov_{jk} = \sum_{i=1}^{n} (X_{ij} - X_{jm})(X_{ik} - X_{km}) \Big/ n-1 \qquad \textbf{(Equation 3)}$$

						Σ	X_m	S
A (X_i)	10.00	3.00	4.00	4.10	3.00	24.10	4.82	2.9431
B (X_j)	9.80	2.80	3.90	4.00	2.80	23.30	4.66	2.9312
$X_i - X_{im}$	5.18	-1.82	-0.82	-0.72	-1.82			
$X_j - X_{jm}$	5.14	-1.86	-0.76	-0.66	-1.86			
$(X_i - X_{im})(X_j - X_{jm})$	26.6252	3.3852	0.6232	0.4752	3.3852	34.4940		

Covariance_{AB} = 34.4940 / 4 = 8.6235

Table 4: *The calculation of covariance between Site A and B*

	A	B	C	D	E	F
A	8.66	8.62	8.75	-7.34	-7.54	-7.60
B	8.62	8.59	8.71	-7.28	-7.48	-7.53
C	8.75	8.71	8.85	-7.43	-7.63	-7.69
D	-7.34	-7.28	-7.43	7.44	7.56	7.73
E	-7.54	-7.48	-7.63	7.56	7.69	7.85
F	-7.60	-7.53	-7.69	7.73	7.85	8.03

Table 5: *Variance-covariance matrix for the sample sites shown in Table 2.*

The next step in the process is to convert the covariance values into correlation coefficients. This is an important and necessary step as it allows the interrelation between variables to be calculated, which are independent of the measurement units used to describe them. This linear measure of interdependence between two variables is defined by:

$$Corr_{jk} = Cov_{jk} \Big/ (S_j \times S_k) \qquad \textbf{(Equation 4)}$$

Correlation coefficient values always lie between -1 to +1 as the covariance can never exceed the product of the standard deviations. A positive value close to 1 indicates that the two variables have a strong interdependence and increase together at a similar rate. However, the opposite is true for a negative value close to -1. In this case, as one variable is increasing the other is moving in the opposite direction and decreasing. Values close to zero on the other hand indicate that the variables are linearly independent from each other. The full correlation matrix for the sample dataset is shown in Table 6.

	A	B	C	D	E	F
A	1.0000	**0.9998**	0.9996	-0.9137	-0.9235	-0.9109
B	**0.9998**	1.0000	**0.9997**	-0.9102	-0.9205	-0.9074
C	0.9996	0.9997	1.0000	-0.9154	-0.9253	-0.9125
D	-0.9137	-0.9102	-0.9154	1.0000	**0.9995**	**0.9999**
E	-0.9235	-0.9205	-0.9253	0.9995	1.0000	0.9994
F	-0.9109	-0.9074	-0.9125	**0.9999**	0.9994	1.0000

Table 6: *Correlation matrix calculated from the example dataset*

Now that the correlation matrix has been correctly calculated, the algorithm can now start the clustering process. Initially the variables with the highest correlation are sort as these will form the centres of the clusters and have been highlighted in bold for each column in the matrix (Table 6). Studying these values it can be seen that sites D and F form the highest correlated pair with a shared correlation of 0.9999. The second highest mutual correlation is found between sites A and B with a value of 0.9998. These two pairs of sites will therefore form the centres of our clusters and can be graphically displayed as a dendrogram shown in Figure 14a. At this point, sites D and F, and A and B, are now thought upon as being one object having associate properties. Thus further similarities between these clusters and other objects are calculated by averaging their combined values. The newly calculated correlation matrix utilised for the next stage of the clustering process is shown in Table 7. Calculation of the correlation coefficient between new objects DF and AB was therefore achieved by the summation and subsequent averaging of the individual correlations between D to A, D to B, F to A and F to B. This process whereby new objects are assigned to a cluster giving a new reduced correlation matrix is repeated until all data has been grouped forming a tree diagram. In our example, two more iterations are required to cluster all of our data, producing the final correlation matrix shown in Table 8.

	AB	DF	C	E
AB	1.0000	-3.6420	-0.9997	-0.9220
DF	-3.6420	1.0000	-0.9140	0.9995
C	0.9997	-0.9140	1.0000	-0.9253
E	-0.9220	-0.9995	-0.9253	1.0000

Table 7: *Recalculated correlation matrix after first iteration of clustering*

	ABC	DFE
ABC	1	-0.915
DFE	-0.915	1

Table 8: *Final correlation matrix produced by clustering process*

As can be seen in the dendrogram shown in Figure 14b, during the second iteration, site E has joined cluster DF and site C joins cluster AB. The final dendrogram shown in Figure 14c depicts the final step whereupon all objects are linked together completing the tree diagram. This clearly enables visualisation of the similarity between objects and gives detail to the underlying structure of the dataset. The agglomerative clustering process is completely unsupervised and requires only the information contained within the dataset and subsequent similarity matrices to be completed. However, the division of objects into separate classes is finally defined by "cutting" the completed dendrogram. This is a subjective process and requires informed user input to help identify the optimal amount of clusters that best characterise the patterns held within the data. In our worked example, the optimal amount of clusters that best describes our dataset is most likely to be two, with the dendrogram being cut at the point where DFE connects to ABC having a negative correlation between them (Figure 14c). The differences between data held within these two clusters could additionally be visualised by plotting both the original dataset and the average values of the individual clusters.

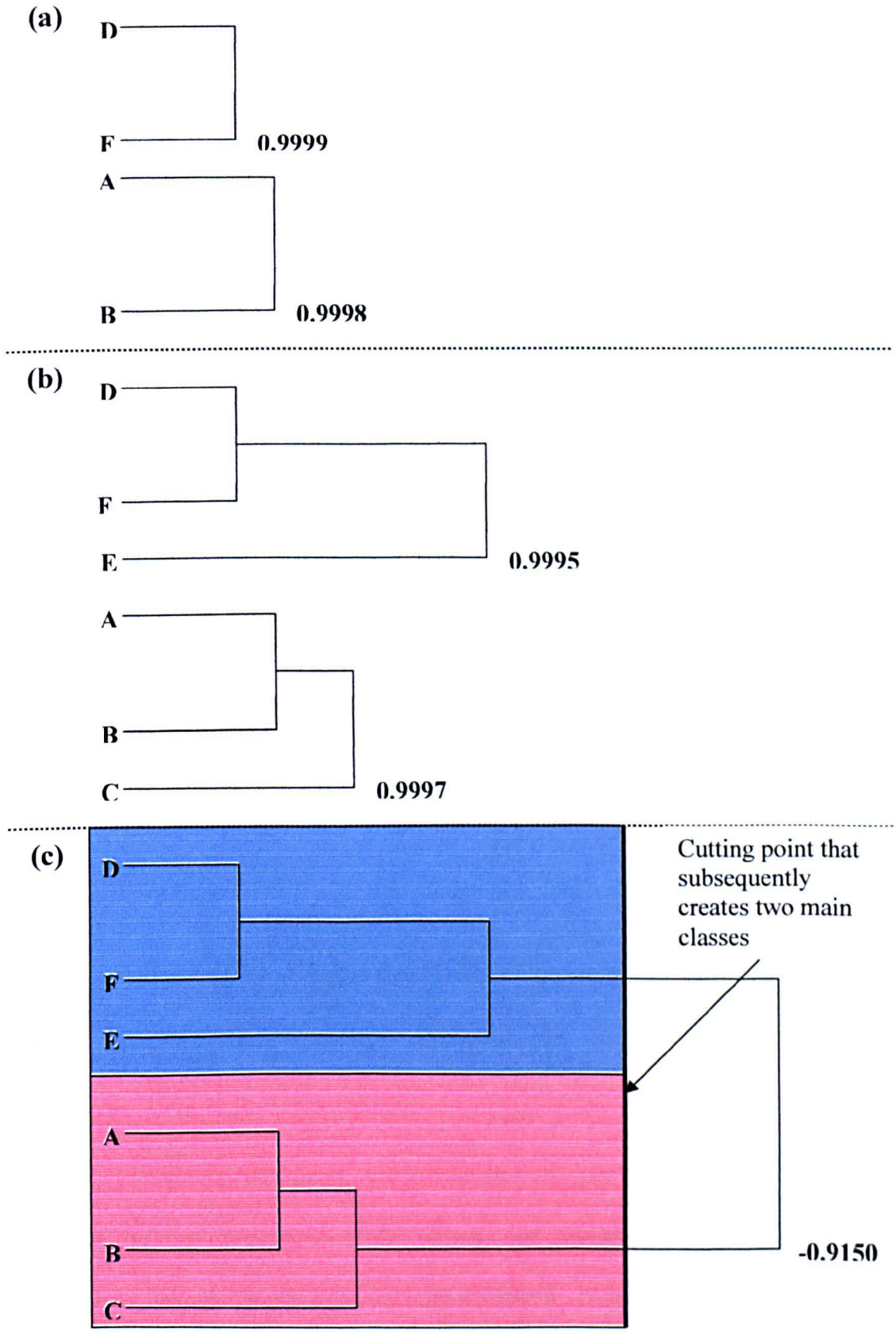

Figure 14: *Dendrograms describing the three stages of hierarchical clustering for our example dataset. Iterations (a) – (c) describe the correlation and subsequent links calculated in tables 4-6 respectively.*

Although the correlation coefficient can be an effective similarity measure for clustering techniques, it is only a measure of the co-linearity between variables of the data. Thus non-linear relationships that may exist between variables are not taken into account, which could ultimately enhance characterisation and further pattern recognition. A more practical method that can allow such relationships to be considered is to use the distance that exists between objects as a measure of their similarity. Each object or datum within a dataset is ultimately characterised by the value of its individual variables. Objects can therefore be alternatively represented as a single point within multidimensional space, each dimension or axes symptomatic to a variable of the data. Distance measures between these objects in multidimensional space can therefore be calculated and a distance matrix similar to that of a correlation matrix produced and used for clustering. A number of distance measures have been proposed for clustering processes, but the most commonly referenced and adopted in this study is the Euclidean distance. This can be defined by:

$$D_{AB} = \left[\sum_{j} (X_{1j} - X_{2j})^2 \right]^{1/2} \qquad \textbf{(Equation 5)}$$

The calculated distance matrix for a given dataset is initially clustered in a similar fashion as previously described, whereupon the two most similar objects are linked. However, in this circumstance, the objects that display the smallest separating distance are then paired and a cluster centre formed. At this point in the algorithm, a variety of different metrics can be utilised to recalculate the between cluster distances and subsequent reduced distance matrix [54-56]. In our studies, the Wards

algorithm was applied having gained a large amount of support in the literature [26,36,57], and has a tendency to produce compact clusters with large between cluster distances [58]. The Wards algorithm can be defined as [58]:

$$D_{C(AB)} = (\alpha_i \times D_{CA}) + (\alpha_j \times D_{CB}) + (\beta \times D_{AB}) \qquad \textbf{(Equation 6)}$$

where: D_{AB} is the distance found between objects A and B

$D_{C(AB)}$ is the distance between object C and new object AB

The individual coefficients α_i, α_j and β are defined as:

$$\alpha_i = \frac{N_C + N_A}{N_C + N_B + N_A} \qquad \textbf{(Equation 7)}$$

$$\alpha_j = \frac{N_C + N_B}{N_C + N_A + N_B} \qquad \textbf{(Equation 8)}$$

$$\alpha_j = \frac{-N_C}{N_C + N_A + N_B} \qquad \textbf{(Equation 9)}$$

Whereby the number of objects contained within cluster x equals N_x.

The algorithm can then henceforth repeat these two steps until all objects contained in the dataset have been clustered.

4.5.4.2 Fuzzy C-Means Clustering (FCM)

An alternative method that can be utilised for clustering data is to employ a numerical routine that aims to optimise the intra-cluster distance. Such techniques utilise an iterative algorithm to continuously update the position of randomly selected initial cluster centres, until a minimum improvement in the clusters compactness is observed or no objects can be further reassigned. These types of iterative algorithm are best described by again imagining our data as single points that exist in multidimensional space, whereupon each dimension defines a variable of the data. To help fully understand the steps involved in this process, a random 2D data matrix, as shown in Figure 15a, will be clustered into 4 groups as an example. Initially, all objects in the 2D matrix are randomly given a cluster membership to 1 of the 4 predefined clusters (Figure 15b). Centre points for each cluster within this space are then computed (Figure 15c). The next step in the algorithm is to calculate the distance that exists between all objects in the dataset and the centre points. Using these measures, each object in the dataset is reassigned to the cluster with the minimum separation. Subsequently, new centre points for the clusters can be computed allowing the distance between all objects and these centres to be recalculated (Figure 15d). If an objects closest centre point in space no longer belongs to the cluster it is presently a member of, the object will change its membership to the cluster that has the closet centre point. This iterative process is repeated until no objects remain that have not been reassigned or a minimum improvement in cluster density has been reached (15e).

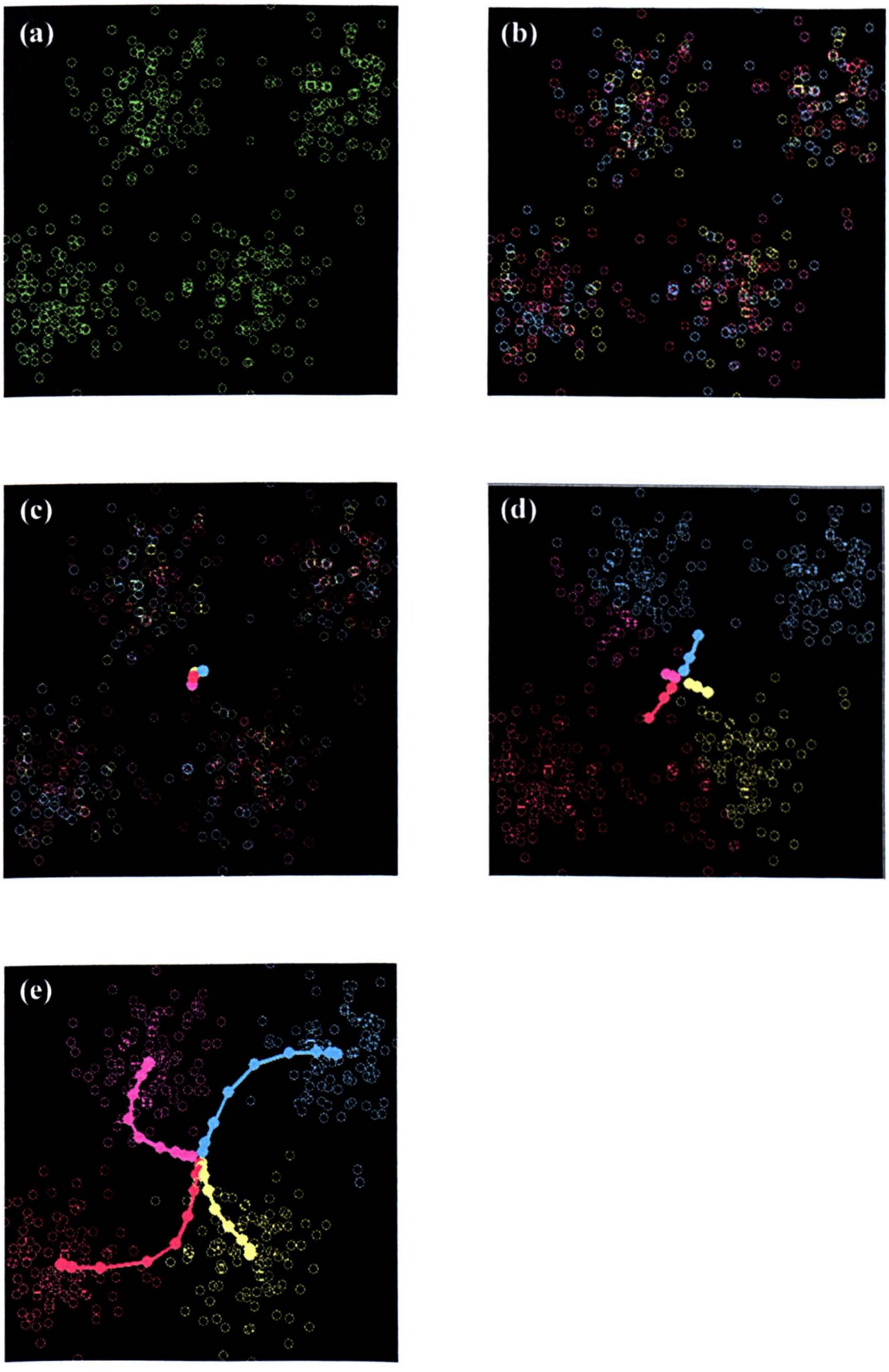

Figure 15: *Fuzzy clustering of random 2D matrix (a). (b) – (e) describe random initialisation, find initial centres, recalculation of fuzzy centres and termination steps respectively.*

Fuzzy clustering algorithms such as FCM have shown distinct advantages over more traditional techniques that use crisp and probabilistic methods to define cluster membership [59-62]. Rather than using a two-class system to define cluster membership (0 or 1) as described above, a membership function is employed that defines the degree of membership of each object to each cluster. FCM then clusters the data by minimising the objective function:

$$J(U,V)=\sum_{i=1}^{n}\sum_{j=1}^{c}(\mu_{ij})^{m}\,\|x_i-v_j\|^2 \qquad \textbf{(Equation 10)}$$

Where $X=\{x_1,x_2,...x_n\}$ is the set of data, μ_{ij} represents the membership degree of an object x_i to the cluster centre v_j. μ_{ij} must also satisfy the following conditions:

$$\mu_{ij}\in[0,1],\quad \forall i=1,..n,\forall j=1,..c \qquad \textbf{(Equation 11)}$$

$$\sum_{j=1}^{c}\mu_{ij}=1,\quad \forall i=1,...n \qquad \textbf{(Equation 12)}$$

The closer the object x_i is to the cluster centre v_j, the higher the value μ_{ij} will be, and vice versa. $\|x_i-v_j\|$ represents the Euclidean distance between x_i and v_j. The parameter m is used to control the fuzziness of the membership for each object, $m >$ 1. There is no theoretical basis for the optimal selection of m, but a value of $m = 2.0$ is conventionally chosen. $U=(\mu_{ij})_{n*c}$ is a fuzzy partition matrix and $V=\{v_1,v_2,...v_c\}$ is a set of cluster centres. FCM can be described by the following steps [59]:

1) Initialize membership matrix μ_{ij} with random value, satisfying conditions **(11)** and **(12)**.

2) Compute the fuzzy centres v_j for the defined amount of clusters using

$$v_j = \frac{\sum_{i=1}^{n} (\mu_{ij})^m x_i}{\sum_{i=1}^{n} (\mu_{ij})^m}, \forall j = 1,..., c \qquad \textbf{(Equation 13)}$$

3) Calculate the new distance d_{ij} between each object and the fuzzy centres

$$d_{ij} = \| x_i - v_j \|, \forall i = 1,...n, \forall j = 1,..c \qquad \textbf{(Equation 14)}$$

4) Update the fuzzy membership μ_{ij} for each object to each cluster

If $d_{ij} \neq 0$ $\quad \mu_{ij} = \frac{1}{\sum_{k=1}^{c} (\frac{d_{ij}}{d_{ik}})^{\frac{2}{m-1}}}$

Else $\mu_{ij} = 1$ **(Equation 15)**

5) Repeat step 2) to 4) until a predefined minimum J value is achieved. When the analysis has come to completion, each object is assigned to a specific cluster for which the degree of the membership is maximal.

In this study, we used set parameters for the FCM analysis. The maximal number of allowed iterations was set to 100, the minimum objective function value was $1.0x10^{-7}$, and the number of clusters was subjectively increased from 2-8.

4.5.4.3 Combination of PCA and FCM clustering

As mentioned in section 4.3.2, the second type of application PCA can be used for is to reduce the dimensionality of a dataset. By representing the data on new orthogonal axes that are uncorrelated to each other, and account for a maximal amount of variance, the data can now be described by a reduced number of variables or PC's, without a significant loss of information. We have applied this compression technique in a newly developed algorithm that combines both PCA and FCM Clustering. Traditionally FCM Clustering has been directly applied to large vibrational datasets. In our experiments, we have used PCA to reduce the dimensionality of our datasets so that spectra are now described by only the first 10 PC's, normally accounting for 95% and above of the total variance. FCM clustering is then directly applied to these datasets. By reducing the amount of dimensions, the computation time, especially for very large datasets, is dramatically reduced.

4.5.4.4 Novel Automated FCM Merge Method Algorithm

Although clustering techniques have shown the ability to group tissue spectra according to their clinical diagnosis, the number of clusters (i.e. types or subtypes of tissue) that best describe a sample is still an unknown feature and usually requires

human input. A clustering method that could automatically cluster different tissue types would therefore be of huge benefit, providing a more convenient and efficient approach in practice.

In our previous work, a fuzzy clustering algorithm that featured simulated annealing (SAFC) was used to automatically detect the 'optimal' number of clusters that would describe tissue spectra collected from several different tissue types [46,63]. The algorithm utilised initially choose a random number of clusters, and then traversed the search space using three different neighbourhood operations: i) *perturb centre*, ii) *delete centre* and iii) *split centre*. The final clustering result was therefore dictated by the data structure in multidimensional space that gave the smallest cluster validity index value. The experiments showed that the algorithm obtained the identical number of clusters as defined by clinical analysis in four out of the seven datasets analysed. Although these results were promising, the algorithm did over estimate the number of clusters in almost half of the experiments and could become time consuming with larger datasets. With the aim of overcoming this problem, in our latest work [64,65], a refined FCM based clustering algorithm was used to find the 'optimal' number of clusters. But while the algorithm was better suited for larger datasets, the analysis again identified an excessive number of clusters in a number of cases. This was partly due to the FCM algorithm and cluster validity index, where all distances between data points and cluster centres are calculated using their Euclidean distance. This means that when the shape of the clusters were significantly different from spherical, the clustering and validity measures were not effective. However, the complexity and range of the different cell types (e.g. pre-cancerous and mature cancer) may also lead to an excessive number of clusters being

identified. At this stage of our study, we only want to cluster spectra into groups that can be identified by pathology and therefore assess natural variation within these defined tissue types. In order to achieve this goal, we need to combine the clusters that have the most similar characteristics, e.g. the suspected pre-cancerous and mature cancerous cell types, as they may exhibit similar properties to one another even though they are at different stages of malignancy. This information may be contained in the existing infrared spectra or data analysis. A new method is therefore proposed to automatically merge greatly similar clusters and detect a more suitable clustering structure using the characteristics of the tissue spectra.

To help describe this newly proposed technique, a problematic dataset previously analysed [64,65], will be used as an example. In this particular dataset, 159 cancerous (from different areas of cancerous tissue), 72 normal and 45 reticular tissue spectra were collected and analysed via FCM clustering. When the number of clusters was set at three to match the clinical analysis, the clustering results did not match the clinical diagnosis (possibly due to the Euclidean distance measurement in FCM). Some cancerous spectra were incorrectly clustered into the same group as normal tissue spectra. When we subsequently applied the automatic FCM based clustering algorithm, four clusters were obtained. However, two of the four clusters corresponded to one type of tissue (cancerous). In the remaining two clusters, the overwhelming majority of the data was correctly classified into their corresponding clinical group (only two spectra were misclassified). These results are shown in Figure 16. The excessive number of clusters may be caused by the fact that the cancerous spectra, which were taken from diverse areas of tissue, may contain cells at different stages of the cancer (e.g. pre-cancerous and mature cancer). As

mentioned before, at this stage, we only wish to cluster cell spectra that have the same clinical diagnosis.

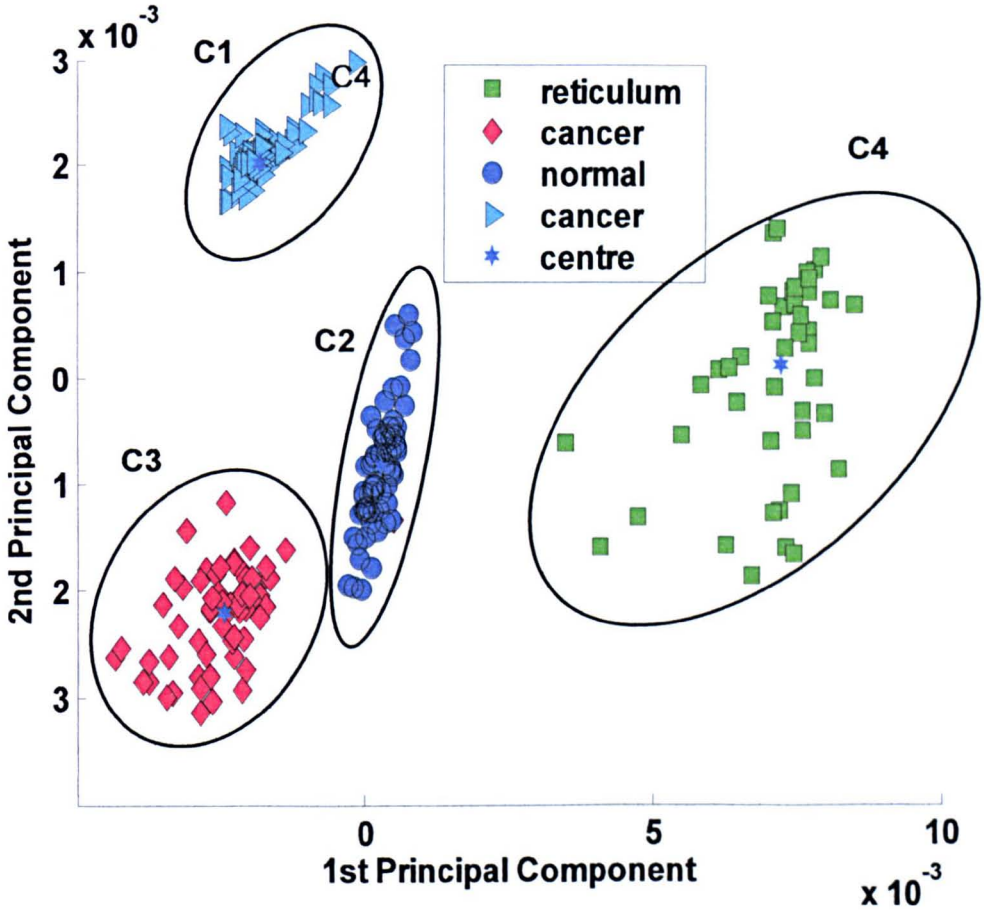

Figure 16: *The example spectral dataset loaded onto the first two principal components after applying the automatic FCM model selection algorithm. C1-C4 describe the cluster numbers respectively.*

Previously, many algorithms have been proposed to merge clusters [66,67]. The different approaches can generally be divided into two groups. The first group are those that select the clusters which are 'closest' to each other [66] and, the second, those that choose the 'worst' two clusters (judging by some cluster validity function) [67]. When applying these principles to the dataset shown in Figure 1, the two closest clusters within the existing four clusters are C1 and C2 (see the distance between each cluster centre). Generally, a good cluster is defined by the property

that data points within the cluster are tightly condensed around the centre (compactness). In this dataset, C1 and C2 are more compact than the other two, so the worst two clusters are C3 and C4. However, the two clusters that should be merged together are C1 and C3 (both are cancer). Hence, neither of these approaches for merging clusters is suitable for solving the problem here. Therefore, we looked for a new solution based on examining the original infrared spectra rather than searching for a relationship using the data structures in the PCA plot. Plotting the mean spectra from the separate clusters allows the major differences between them to be more clearly visualized. The similarity between clusters is more obvious at the wavenumber where the biggest difference between any two mean spectra is located. Our proposed automated merge clustering method is based on this observation and can be divided into two main stages. The first stage is to discover a reference wavenumber which the cluster merging process will use. The second step is to repeatedly determine the most similar clusters and merge them, until certain a termination criteria has been reached. In the following section the two steps are described in detail.

Step1: Determine a reference frequency

The reference frequency is defined as the wavenumber at which the biggest difference between any two mean spectra is found. The full procedure of determining this frequency is:

1) Obtain the clustering results from the automatic FCM based clustering algorithm.

2) Calculate the mean spectra $\overline{A_i}$ for each cluster,

$$\overline{A_i} = \frac{1}{N_i}\sum_{j=1}^{N_i} A_{ij} \qquad (i=1...c) \qquad \textbf{(Equation 16)}$$

where N_i is the number of data points in the cluster i; A_{ij} is the absorbance of the spectrum for each data point j in cluster i; c is the number of clusters. The size of $\overline{A_i}$ is p, the number of data points in each spectrum (each mean spectrum is a vector of p elements).

3) Compute the vector of pair-wise absolute differences D_{ij} between all mean spectra,

$$D_{ij} = \left| \overline{A_i} - \overline{A_j} \right| \quad (i=1...c,\ j=1...c) \qquad \textbf{(Equation 17)}$$

4) Find the largest single element, d_{max}, within the set of vectors D.

5) Determine the frequency corresponding to the maximal element d_{max}.

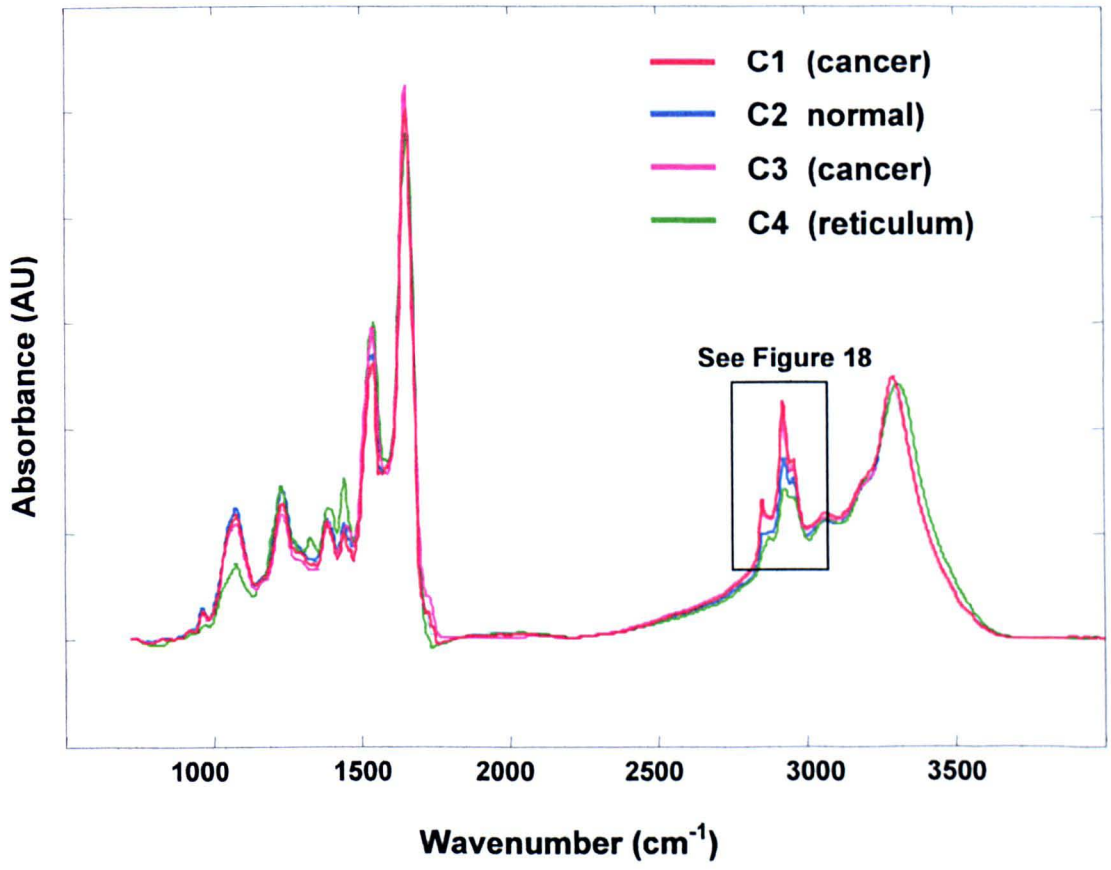

Figure 17: *Mean infrared spectra calculated for each cluster partitioned by the automated FCM clustering technique. Note the 'optimal' validity measure was reached when adopting a 4 cluster stricture.*

Mean average IR spectra for the four resultant clusters are displayed in Figures 17 and 18. The set of differences, D, between each pair of mean spectra was calculated using equation (17). The largest difference d_{max} exists between C_1 and C_4, as shown in Figure 18. The wavenumber that corresponds to d is 2924 cm^{-1}.

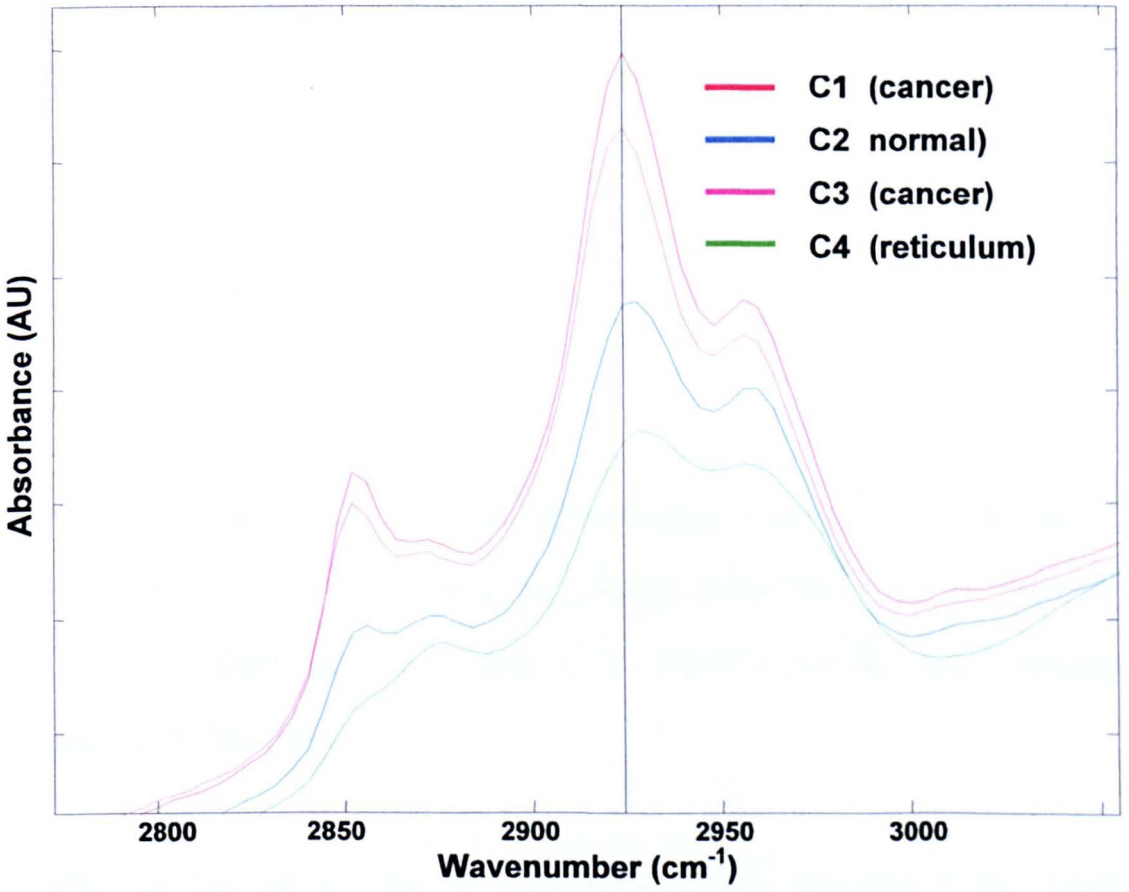

Figure 18: *Enlarged spectral region indicated in Figure 17.*

Step2: Automatically merging clusters

After finding the reference frequency, the next step is to choose the most similar clusters based on the absorbance value of each mean spectrum at this wavenumber, and then merge them together. As this is an iterative process, the merging procedure

will end when at least one of the termination criteria has been satisfied. Assume currently there are C mean spectra. The detailed information can be described as below:

1) Obtain C absorbance values of the mean spectra at the reference frequency from step 1, re-sort them in ascending order.

2) Calculate the distance, *dist,* between these sorted absorbance values (note that the size of *dist* is now *C-1*)

3) Identify the smallest distance, $dist_{min}$, and find out the two most similar clusters which correspond to this distance.

4) Merge these two clusters if they satisfy the merging condition: $dist_{min} \leq$ *average of rest of dist (without* $dist_{min}$*)*. The average of these two clusters mean absorbencies is then calculated and considered as a new object to join the rest of merging iteration. Go back to 1.

5) When there are only two *dist* left. The merging condition changes to: if the current $dist_{min} \leq 1/2$ *rest of dist* or $(dist_{min}$-*1/2 rest of dist)/* $dist_{min} \leq 0.1$, then the two clusters which corresponding the $dist_{min}$ are merged together. Again, the average of the two mean absorbencies is considered as a new object to replace them.

6) The merging process stops if there are only two clusters left or no merging conditions are satisfied.

The same example dataset will be used to help illustrate this process. In Figure 19, $\overline{A_i}$ *(i=1...4)* is the mean spectral absorbance value from each obtained cluster respectively. $\overline{A_1}$= 0.0045, $\overline{A_2}$=0.0034, $\overline{A_3}$=0.0041, $\overline{A_4}$=0.0028. The straight line corresponds to the reference frequency 2924 cm^{-1}. After sorting $\overline{A_i}$ in ascending order, their new arrangement is $\overline{A_4}, \overline{A_2}$, $\overline{A_3}$ and $\overline{A_1}$, obviously $\overline{A_1}$ is the maximum absorbance value, corresponding to cluster C_1 in Figure 18. *Dist* = $\{d_1, d_2, d_3\}$, it is then trivial to calculate d_1 = 0.0006, d_2 = 0.0007, d_3 = 0.0004, which is the $dist_{min}$. The *average of rest of dist* = (0.0006+0.0007)/2 = 0.00065 and greater than $dist_{min}$. This satisfies the merging condition in 4). Therefore, the two clusters which correspond to $dist_{min}$ (C_1 and C_3) are merged together. After this, the average of these two clusters mean spectral absorbance, $\overline{A}_{new}$=0.0043, replaces the previous two values. Re-sort the new array of the mean spectral absorbencies to, $\overline{A_4}$, $\overline{A_2}$ and $\overline{A}_{new}$, as displayed in Figure 20. The corresponding new distances are d_{new1} = 0.0006 and d_{new2} = 0.0009. Reference 5), $dist_{min}$ (0.0006) is not smaller than or equal to *1/2 rest of dist* (0.00045); additionally, it does not satisfy the second condition either. Hence, in this situation, no merging conditions are satisfied, and so continued iteration stops as defined in 6).

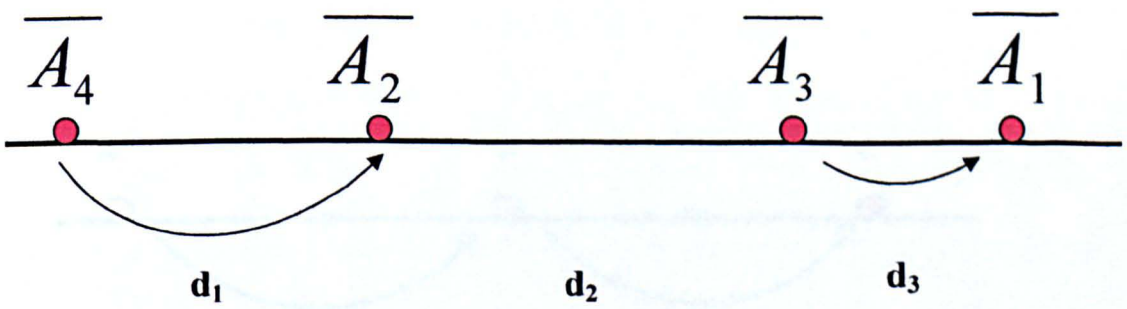

Figure 19: *Schematic representation of the absorbance intensity values found at the reference frequency for each of the four clusters partitioned by the automated FCM analysis.*

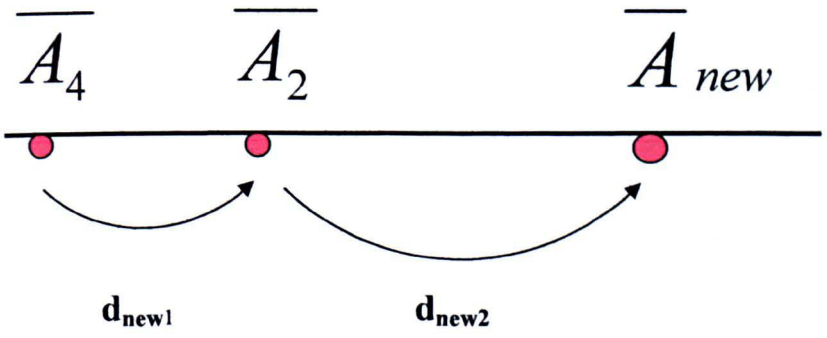

Figure 20: *Schematic representation displaying the new distribution of absorbance intensities found at the reference frequency after the two of the most similar clusters have been merged.*

As may be noticed, the merging condition in 5) is different from the one when there are more than two *dist* left, as depicted in 4). This is because when there are only two *dist* (3 clusters) left, if the same merging condition as in 4) is used, this may lead to two clusters being merged in which their corresponding mean absorbance distance is slightly less than and nearly equal to the other distance. For example, in Figure 21, if d_2 is a slightly less than d_1, then clusters *b* and *c* will be merged together. Visibly, it is not convincing. In order to alleviate this situation, the merging conditions described in 5) are pursued. For example, in Figure 22, if d_2 is smaller than half of d_1 then cluster *b* and *c* are merged together.

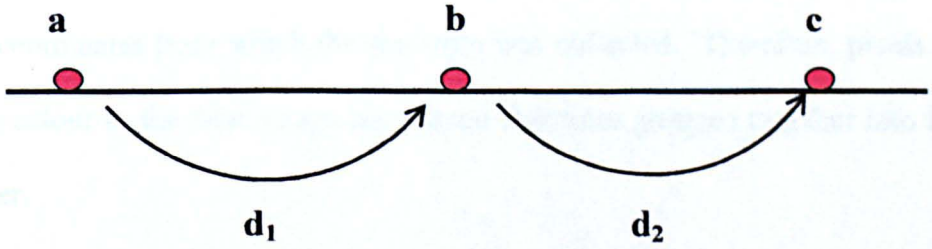

Figure 21: *Schematic representation showing the merging process when there are two dist left (type 1).*

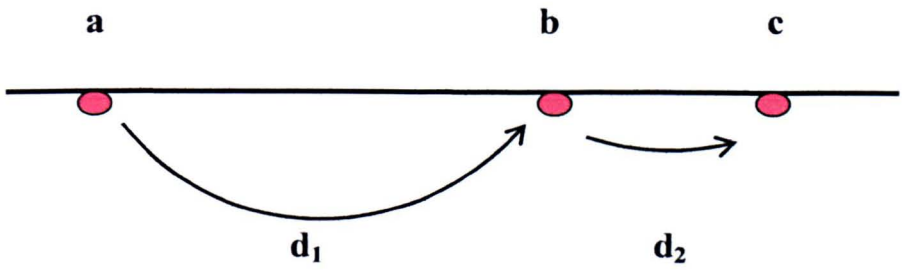

Figure 22: *Schematic representation showing the merging process when there are two dist left (type 1).*

4.5.4.5 Spectroscopic Cluster Imaging

As part of this study, IR spectroscopic maps have been collected from a variety of different tissue sections. In each instance, a parallel tissue section was cut and stained in the conventional way to allow diagnosis via histology. Photomicrographs could then be taken from the same region upon the stained section where a spectral map had been collected. The clusters created during the analysis should contain spectra from histological regions that display comparable spectral characteristics. In contrast, spectra contained in different clusters should exhibit spectral features characteristic of different tissue types. False colour "cluster images" can thus be assembled from the same region and compared directly against these stained images. By assigning each cluster a colour, these colours can then be plotted as pixels at the x, y coordinates from which the spectrum was collected. Therefore, pixels with the same colour in the false image are spectra that were grouped together into the same cluster.

4.6 References

[1] G. N. Papanicolaou and H. F. Traunt, *Am. J. Obstet. Gynecol.*, 1941, **42(2)**, 193.

[2] M. Quinn, P. Babb, J. Jones and E. Allen, *Brit. Med. J.*, 1999, **318**, 904.

[3] S. Gutman, *Acta Cytol.*, 2000, **44(6)**, 1120.

[4] R. M. Austin and I. Ramzy, *Acta Cytol.*, 1998, **42**, 178.

[5] J. Linder, *Diagn. Cytopathol.*, 1998, **18**, 24.

[6] A. B. Carpenter and D. D. Darey, *Cancer*, 1999, **87(3)**, 105.

[7] K. R. Lee, R. Ashfuq, G. G. Birdsong, M. E. Corkill, K. M. McIntosh and S. L. Inhorn, *Obstet. Gynecol.*, 1997, **90**, 278.

[8] J. Monsonego, A. Autillo-Touati, C. Bergeron, R. Dachez, J. Liaras, J. Saurel, L. Zerat, P. Chatelain and C. Mottot, *Br. J. Cancer*, 2001, **84(3)**, 360.

[9] J. L. Papillo, M. A. Zarka and T. L. St, John, *Acta Cytol.*, 1998, **42**, 203.

[10] Taken from Cytyc Website, manufacturers of the ThinPrep instrument, located at URL http://www.cytyc.com.

[11] Taken from SurePath Website, manufacturers of the SurePath instruments, located at URL http://www.surepath.com.

[12] B. R. Wood, M. A. Quinn, F. R. Burden and D. McNaughton, *Biospectrosc.*, 1996, **2**, 143.

[13] J. Kent, 4th Year MSci Project Report, Nottingham Univeristy, 2002.

[14] H. Susi and D. M. Byler, *Biochem. Biophys. Research Commun.*, 1985, **115**, 391.

[15] W. K. Surewicz and H. H. Mantsch, *Infrared Absortion Methods for Examining Protein Secondary Structure, in Determination of Protein*

Stricture in Solution by Spectroscopic Methods, 1994, VCH Publishers, New York.

[16] E. Taillandier, J. Liquier and J. A. Taboury, *Infrared Spectral Studies of DNA Conformations, in Advances in Infrared and Raman Spectroscopy,* 1985, Wiley, New York.

[17] M. Jackson and H. H. Mantsch, *Spectrochimica Acta Review,* 1993, **15**, 53.

[18] H. H. Mantsch and D. Chapman, *Infrared Spectroscopy of Biomolecules,* 1996, Wiley-Liss, New York.

[19] H. L. Casal and H. H. Mantsch, *Biochim. Biophys. Acta,* 1984, **779**, 381.

[20] J. M. Chalmers and P. R. Griffiths, *Handbook of Vibrational Spectroscopy,* 2002, Wiley, Chichester.

[21] R. A. Meyers, *Encyclopaedia of Analytical Chemistry,* 2000, **1**, Wiley, Chichester.

[22] B. Rigas, K. Laguardia, L. Qiao, P. S. Bhandare, T. Caputo and M. A. Cohenford, *J. Lab. Clin. Med.,* 2000, **135**, 26.

[23] G. P. Willimas, *Infrared Synchrotron Radiation, Review of Properties and Prospectives,* National Synchrotron Light Source, 1999, New York.

[24] C. Matthaus, S. Boydston-White, M. Miljkovic, M. Romeo and M. Diem, *Appl. Spectrosc.,* 2006, **60**, 1.

[25] D. C. Fernandez, R. Bhargava, S. M. Hewitt and I. W. Levin, *Nature Biotechnol.,* 2005, **23(4)**, 469.

[26] M. Romeo and M. Diem, *Vibr. Spectroscoc.,* 2005, **38**, 115.

[27] M. Romeo and M. Diem, *Vibr. Spectrosc.,* 2005, **38**, 129.

[28] B. Mohlenhoff, M. Romeo, M. Diem and B. R. Wood, *Biophys. J.*, 2005, **88**, 3635.

[29] A. Pacifico, L. Chiriboga, P. Lasch and M. Diem, *Vibr. Spectrosc.*, 2003, **32**, 107.

[30] S. Boydston-White, T. Gopen, S. Houser, J. Bargonetti and M. Diem, *Biospectrosc.*, 1999, **5**, 219.

[31] H. C. Van-Der-Mei, D. Naumann and H. J. Busscher, *Arch. Oral Biol.*, 1993, **38**, 1013.

[32] B. R. Wood, M. A. Quinn, F. R. Burden and D. McNaughton, *Biospectros.*, 1996, **2**, 143.

[33] M. Romeo, B. R. Wood and D. McNaughton, *Vibr. Spectrosc.*, 2002, **28**, 167.

[34] P. Lasch and D. Naumann, *Cell. Mol. Biol.*, 1998, **44(1)**, 189.

[35] N. Stone, C. Kendall, N. Shepherd, P. Crow and H. Barr, *J. Raman Spectrosc.*, 2002, **33**, 564.

[36] B. R. Wood, L. Chiriboga, H. Yee, M. A. Quinn, D. McNaughton and M. Diem, *Gynecol. Oncol.*, 2004, **93**, 59.

[37] H. C. Van-Der-Mei, D. Naumann and H. J. Busscher, *Arch. Oral Biol.*, 1993, **38**, 1013.

[38] C. P. Schultz and H. H. Mantsch, *Cell. Mol. Biol.*, 1998, **44(1)**, 201.

[39] M. Jackson, B. Ramjiawan, M. Hewko and H. H. Mantsch, *Cell Mol. Biol.*, 1998, **44(1)**, 89.

[40] M. Diem, L. Chiriboga and H. Yee, *Biopolymers*, 2000, **57(5)**, 282.

[41] C. P. Schultz, K.-Z. Liu, J. B. Johnson and H. H. Mantsch, *Leukemia Res.*, 1996, **20(8)**, 649.

[42] J. R. Mansfield, M. G. Sowa, G. B. Scarth, R. L. Somorjai and H. H. Mantsch, *Anal. Chem.*, 1997, **69**, 3370.

[43] X. Y. Wang and J. M. Garibaldi, *A comparison of Fuzzy and Non-Fuzzy Clustering Techniques in Cancer Diagnosis,* in the *Proceedings of 2nd International Conference in Computational Intelligence in Medicine and Healthcare,* 2005.

[44] L. M. McIntosh, J. R. Mansfield, N. A. Crowson and H. H. Mantsch, *Biospectrosc.,* 1999, **5**, 265.

[45] L. Zhang, G. W. Small, A. S. Haka, L. H. Kidder and E. N. Lewis, *Appl. Spectrosc.,* 2003, **57(1)**, 14.

[46] X. Y. Wang and J. M. Garibaldi, *Eur. J. Inform.,* 2005, **29**, 61.

[47] I. T. Joliffe, *Principal Component Analysis*, 1986, Springer-Verlag, New York.

[48] P. D. A. Pudney, T. M. Hancewicz, D. G. Cunningham and C. Gray, *Food Hydrocolloids*, 2003, **17**, 345.

[49] J-H. Wang, P. K. Hopke, T. M. Hancewicz and S. L. Zhang, *Anal. Chim. Acta,* 2003, **476**, 93.

[50] P. D. A. Pudney, T. M. Hancewicz, D. G. Cunnigham and M. C. Brown, *Vibr. Spectrosc.,* 2004, **34**, 123.

[51] E. Malinowski, *Anal. Chem.,* 1977, **49**, 612.

[52] E. R. Malinowski, *J. Chemometr.*, 1987, **1**, 33.

[53] B. Everitt, *Cluster Analysis*, 1980, Heinemann Educational, London.

[54] E. Garett-Mayer and G. Parmigiani, *John Hopkins University, Dept. of Biostatistics Working Press Papers,* John Hopkinds Univeraity, The Berkeley Electronic Press (bepress).

[55] A. K. Jain, M. N. Murty and P. J. Flynn, *ACM Comput. Surv.,* 1999 **31(3)**, 264.

[56] D. Jiang, C. Tang and A. Zhang, *IEEE T. Knowl. Data En.*, 2004, **16(11)**, 1370.

[57] P. Lasch, W. Haensch, D. Naumann and M. Diem, *Biochim. Biophys. Acta*, 2004, **1668**, 176.

[58] J. H. Ward, *J. Am. Stat. Assoc.*, 1963, **58**, 236.

[59] J. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, 1981, Plenum, New York.

[60] F Hopper, *Fuzzy Clustering Analysis Methods for Classification, Data Analysis and Image Recognition*, 1999, Wiley, New York.

[61] Y. Zhao and G. Karypis, *Clustering Criterion Functions for Partitional Document Clustering: A summary of results, in the Proceedings of the 13th ACM International Conference of Information and Knowledge Management*, 2004.

[62] G. Hamerly and C. Elkan, *Alternatives to the K-means Algorithm that find better clusterings, in the Proceedings of the 11th ACM International Conference of Information and Knowledge Management*, 2002.

[63] X. Y. Wang, G. Whitwell and J. Garibaldi, *The Application of a Simulated Annealing Fuzzy Clustering Algorithm for Cancer Diagnosis, in the Proceedings of the 4th International Conference for Intelligent System Design and Application*, 2004.

[64] X. Y. Wang, J. M. Garibaldi, B. Bird and M. W. George, *Fuzzy Clustering in the Biochemical Analysis of Cancer Cells, in the Proceedings of the Fourth*

Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT), 2005, Barcelona, Spain, 1118.

[65] X. Y. Wang, J. M. Garibaldi, B. Bird and M. W. Geogre, *Appl. Intell.*, 2006 (In press).

[66] J. H. Ward, *J. Am. Stat. Assoc.*, 1963, **58**, 236.

[67] Y. Xie, V. V. Raghavan and X. Zhao, *3M Algorithm: Finding an Optimal Fuzzy Cluster Scheme for Proximity Data, in the Proceedings of the 2002 FUZZ-IEEE Conference,* 2002.