CRANFIELD UNIVERSITY

Martina Sattlecker

Optimisation of Machine Learning Methods for Cancer Diagnostics using Vibrational Spectroscopy

Cranfield Health

PhD Thesis

CRANFIELD UNIVERSITY

Cranfield Health

PhD Thesis

Academic Year 2010-2011

Martina Sattlecker

Optimisation of Machine Learning Methods for Cancer Detection using Vibrational Spectroscopy

Supervisors:

Dr. Conrad Bessant

Prof. Nick Stone

January 2011

This thesis is submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy.

Abstract

Early cancer detection drastically improves the chances of cure and therefore methods are required, which allow early detection and screening in a fast, reliable and inexpensive manner. A prospective method, featuring all these characteristics, is vibrational spectroscopy. In order to take the next step towards the development of this technology into a clinical diagnostic tool, classification and imaging methods for an automated diagnosis based on spectral data are required.

For this study, Raman spectra, derived from axillary lymph node tissue from breast cancer patients, were used to develop a diagnostic model. For this purpose different classification methods were investigated. A support vector machine (SVM) proved to be the best choice of classification method since it classified 100% of the unseen test set correctly. The resulting diagnostic models were thoroughly tested for their robustness to the spectral corruptions that would be expected to occur during routine clinical analysis. It showed that sufficient robustness is provided for a future diagnostic routine application.

SVMs demonstrated to be a powerful classifier for Raman data and due to that they were also investigated for infrared spectroscopic data. Since it was found that a single SVM was not capable of reliably predicting breast cancer pathology based on tissue calcifications measured by infrared micro-spectroscopy a SVM ensemble system was implemented. The resulting multi-class SVM ensemble predicted the pathology of the unseen test set with an accuracy of 88.9%, in comparison a single SVM assessed with the same unseen test set achieved 66.7% accuracy. In addition, the ensemble system was extended for analysing complete infrared maps obtained from breast tissue specimens. The resulting imaging method successfully detected and staged calcification in infrared maps. Furthermore, this imaging approach revealed new insights into the calcification process in malignant development, which was not previously well understood.

Acknowledgments

This study was carried out at the Bioinformatics group at Cranfield University. I would like to thank my supervisor Dr. Conrad Bessant for his guidance throughout the project. Furthermore, I would like to thank my colleagues for their technical advisory and assistance.

In the same way, would like to thank the Biophotonics research group at the Gloucestershire Hospital for providing data for this study. Special thanks are expressed to my supervisor Prof. Nick Stone for taking time to introduce me well into the area of vibrational spectroscopy and his guidance throughout the project. I also would like to thank the other members of the biophotonics group who helped me in getting familiar with Raman and infrared imaging techniques.

I also would like to acknowledge the financial support from both Cranfield Health and Gloucester Hospital.

Ganz besonderer Dank gilt meiner Familie, im Besonderen meinen Eltern für ihre ununterbrochene und liebevolle Unterstützung über all die Jahre hinweg. Gleichermaßen möchte ich mich auch bei meinen Großeltern bedanken, die mir immer motivierend zur Seite standen.

Table of contents

List of Figures

List of Tables

Abbreviations

ANN: Artificial neural network

CHAP: Calcium hydroxyapetite

COD: Calcium oxalate dihydrate

CT: Computer tomography

EES: Elastic scattering spectroscopy

FCM: Fuzzy C-means clustering

FIR: Far-infrared

FT-IR: Fourier transform infrared

GI: Gastrointestinal

HCA: Hierarchical cluster analysis

IR: Infrared

KM: K-means

LDA: Linear discriminate analysis

LOOCV: Leave-one-out cross-validation

LV: Latent variable

MIR: Mid-infrared

MRI: Magnetic resonance imaging

NIR: Near-infrared

PCA: Principal component analysis

PC: Principal component

PLS: Partial least squares

PLS-DA: Partial least square discriminant analysis

PSA: Prostate-specific antigen

RBF: Radial basis function

SIMCA: Soft independent modelling by class analogy

SVM: Support vector machine

1 Introduction

1.1 Background

Cancer is one of the leading causes of death in western countries and incidences increase constantly all over the world. For example, in the UK around 298,000 people are newly diagnosed with cancer every year. In 2008 alone, over 156,000 deaths were caused due to the course of the disease (Cancer Research UK, 2010a).

The survival rate is strongly influenced by stage of the malignancy at the point of detection. Thus an early detection permits an earlier intervention of therapeutic treatment and helps to reduce the mortality and morbidity rate. In that manner, methods that allow early diagnosis or even population screenings are desirable. An example for such a test is mammography. This method allows detecting calcifications in breast tissue, which are often indicators for a malignant lesion. When a calcification is detected, biopsy is required for distinction between malignant or benign (Stone *et al.*, 2007).

Typically, biopsied tissue is examined by applying histological (tissue based) and cytological (cell based) techniques. These techniques show several limitations. For instance they are subjective because the diagnosis depends on the opinion of a pathologist. Thus, the result can vary when the same sample is examined at different times by the same pathologist or among different pathologists. A further disadvantage is that biopsy is an invasive procedure, which brings risk and discomfort for patients. Finally, the total examination procedure is time consuming (Crow *et al.*, 2003).

Due to the disadvantages of recent diagnostic and screening methods other options should be considered. The requirement for such methods would be that they are non-invasive, fast, objective, low in costs, high-throughput and only need a minimal amount of training. Spectroscopic approaches could meet these demands, such as, for example vibrational spectroscopy, including techniques like Raman and infrared spectroscopy, demonstrated to be very promising techniques for the purpose of disease diagnosis and further medical application (Ellis *et al.*, 2006).

As recently reviewed by Kendall, Isabelle *et al.* (2009), vibrational spectroscopy is capable of diagnosing different diseases in various types of human tissues. Visual inspection of resulting vibrational spectra normally does not allow a distinctive differentiation between healthy and diseased tissue. Thus, accurate computational pattern classification strategies are required to permit future diagnostic applications.

1.2 Biology

Cancer is a potentially lethal disease in humans and because of its raising appearance a growing problem in today's society. Cancer is caused when cell division gets out of control and leads to unregulated growth. A schematic loss of normal growth control is illustrated in Figure 1.1. Consequently, the growth of tumour cells leads to the formation of a tissue lump. The emerging tumour can become malignant by spreading over to other tissues and organs. There are two possibilities of spreading either by invading nearby tissues and organs or by formation of metastases. Generally, metastases are secondary tumours which are formed by tumour cells, which were previously transported over blood and lymph vessels to other parts of the body (Isabelle *et al.*, 2008). The growth of the tumour and additional spread in other organs

can result in organ failure, obstruction of the gastrointestinal tract, ducts and hollow organs and finally death (Prierce *et al.*, 2006).

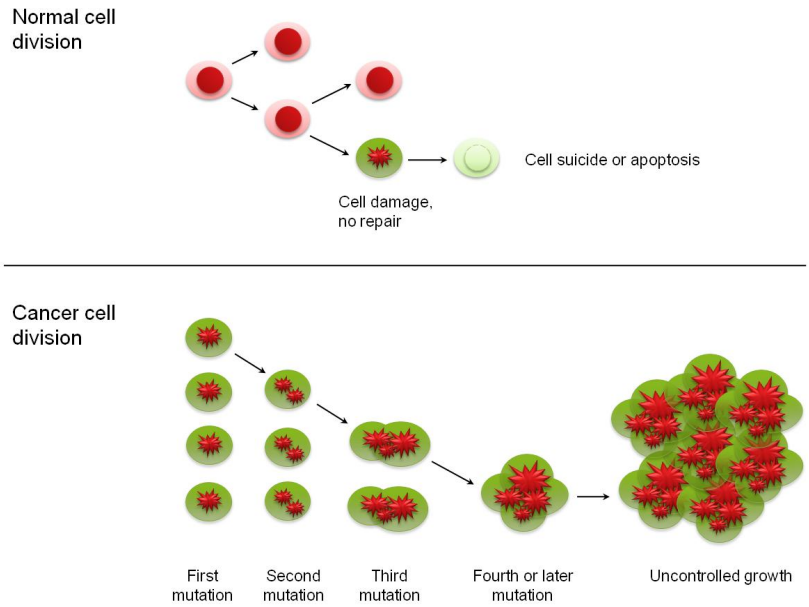


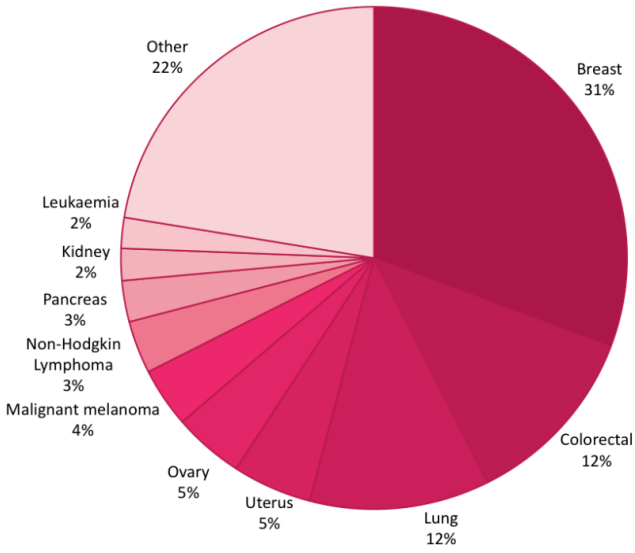
Figure 1.1 Cancer development: Typically, in the case that an error occurs during the normal cycle of cell division this damage is repaired. If this is not possible apoptosis is initiated, which is equivalent to a cell suicide in order to protect the whole organism. Under the circumstances that this protection mechanism fails uncontrolled cell growth is caused. Usually several mutations take place before uncontrolled cell growth begins. Adapted from: http:// www.cancer.gov/cancertopics/ understandingcancer/cancer/Slide4

On the molecular level cancer development changes the cell content. Thus, the nucleic acid, protein, lipid and carbohydrate content of cancerous cells differ from normal cells. This includes an increased chromatin to cytoplasm ratio, disordered chromatin and also changed levels of proteins and lipids (Mahadevan-Jansen *et al.*, 1996). Due to the fact that many of the biological molecules are vibrationally active they can be studied by Raman as well as by infrared spectroscopy. It has been shown already that slight changes in the molecular content of cells are reflected in the appearance of vibrational spectra. For this reason vibrational spectroscopic methods are capable of detecting already minor changes in tissue, which makes them an ideal tool for cancer detection as well as cancer grading (Mahadevan-Jansen *et al.*, 1997).

1.3 Incidence and mortality

Currently cancer can be considered as epidemic since the number of incidences is growing rapidly. In 2002 about 10.9 million new cases were diagnosed and 6.7 million deaths were counted worldwide (Ferlay J *et al.*, 2004). Thus, in Europe over 3.1 million new cases occurred in the year 2006. In the same year 1.7 million deaths caused by cancer were counted in this region (Ferlay *et al.*, 2007). Related to the UK about 298,000 new cancer cases were counted in the year 2008. According to that, one in three persons is likely to develop cancer during their life time (Cancer Research UK, 2010a).

The predominant type of cancer in females is breast cancer with over 45,000 new incidences in the year 2007. Due to that, breast cancer is responsible for almost one third of all diagnosed cancer incidences in females. In the UK breast cancer incidence increased over 50% within the last 25 years. (Cancer Research UK, 2010b). By now about one in ten women is likely to develop breast cancer in western countries (Stratton *et al.*, 2008). An overview of all counted cancer incidences in females and males in the year 2007 is illustrated in Figure 1.2.



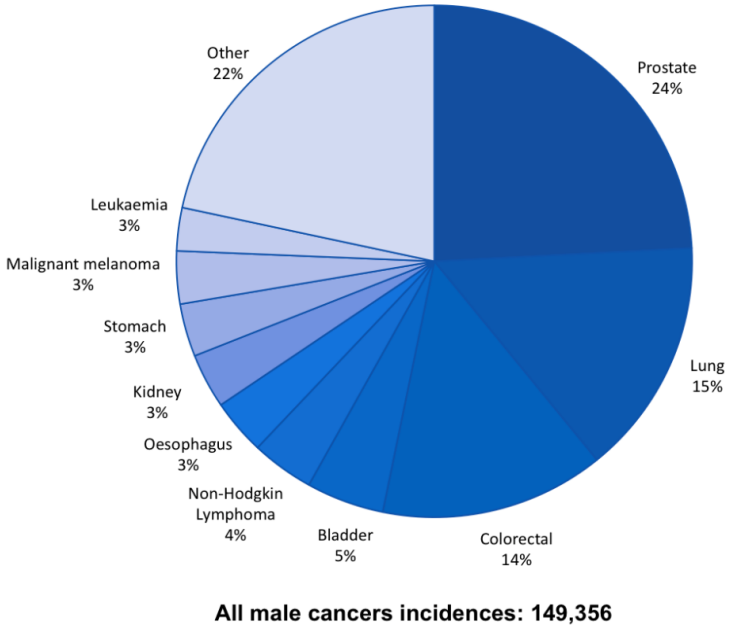


Figure 1.2 Cancer incidences in the UK in 2007: In these graphs the 10 most common cancers in females as well as in males are illustrated. The most common types of cancer are breast, prostate colorectal and lung cancer. These types of cancer were responsible for more than half of all cancer incidences in the UK in the year 2007. Adapted from: Cancer Research UK (2010a)

In 2008 cancer was responsible for 27% of all deaths in the UK and thus one in four deaths was caused by cancer. The majority of deaths is caused by lung cancer, colorectal cancer, breast cancer and prostate cancer (Cancer Research UK, 2010c). An overview of the mortality rates for the most common types of cancer in the UK is provided in Figure 1.3.

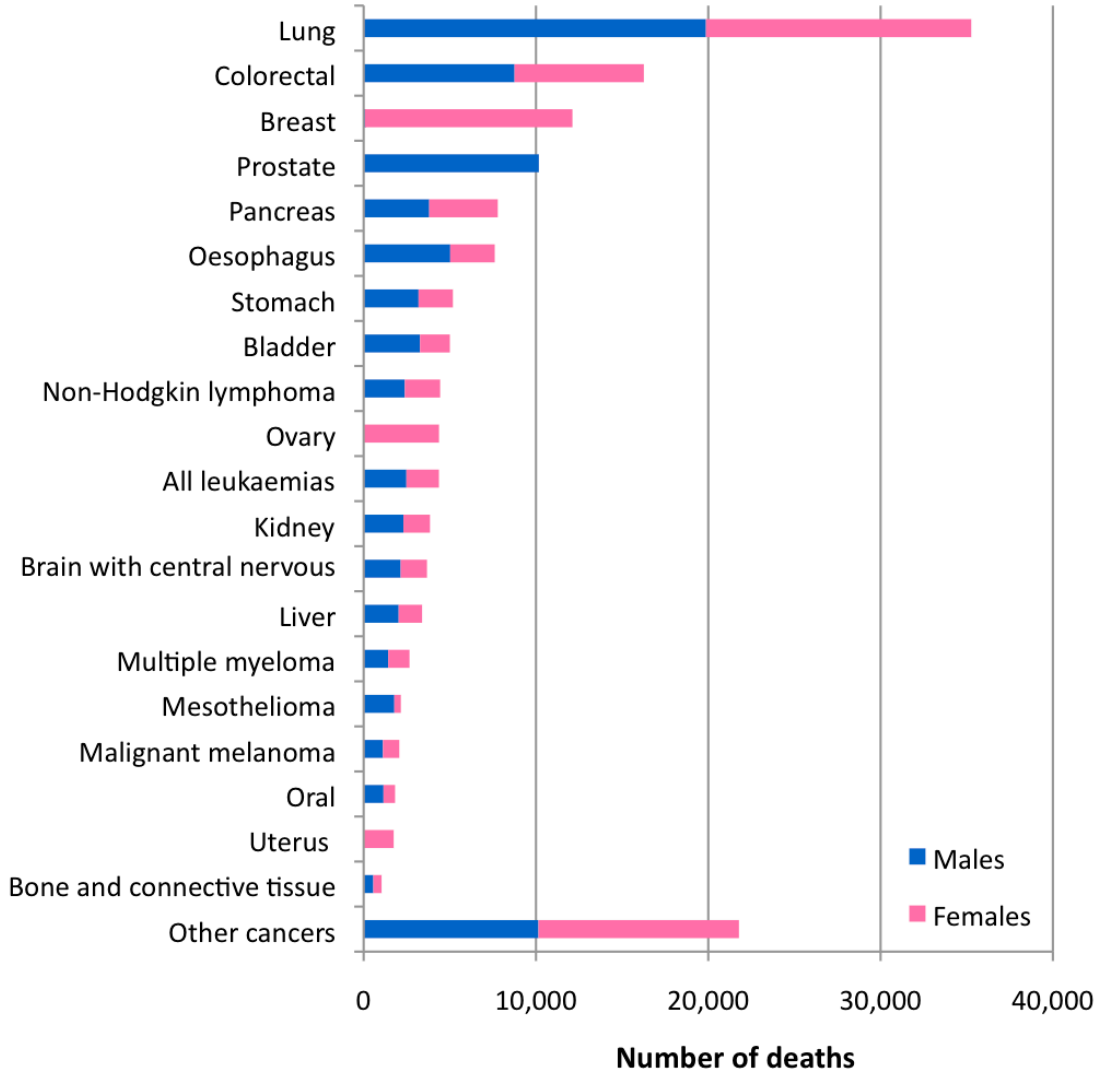


Figure 1.3 Cancer mortality in the UK in 2008: The cancer with the highest mortality rate is lung cancer. Although breast cancer has the highest incidence rate in females and prostate cancer has the highest incidence rate in males, lung cancer shows the highest mortality rate. Adapted from: (Cancer Research UK (2010c)

1.4 Thesis objectives

The aim of this project was to develop and optimise new data analysis strategies for vibrational spectroscopy data and images. For this reason this work was carried out in cooperation with the Biophotonics group at Gloucester Royal Hospital, which has generated

and analysed several high quality data sets derived from different types of cancerous and noncancerous tissue.

First of all, data analysis techniques, which have been applied by the Biophotonics group were evaluated and optimised. In progression of this work, further machine learning methods were investigated for their capability of discriminating between malignant and begin tissue samples based on Raman and infrared spectra. These techniques should be able to take into account the variability between patients and demonstrate robustness towards system-to-system variations. Thus, varying classification techniques, such as Linear Discriminant Analysis (LDA), Partial Least Square Discriminant Analysis (PLS-DA) and support vector machines (SVMs) were explored for their capability to classify tissue according to cancer state.

In order to work towards a fully automated classification of vibrational spectroscopic data, strategies for image analysis were developed. These techniques were able to detect features of interest and predict the pathology of tissue samples.

2 Cancer diagnostics

2.1 Current techniques

Currently there is no single test for accurately diagnosing and staging cancer. Typically, patients who are suspected of developing malignant disease are examined using different imaging techniques, which include X-ray imaging, magnetic resonance imaging (MRI) and computer tomography (CT) (Weissleder *et al.*, 2008). For instance a frequently applied breast cancer screening tool is mammography, which belongs to the group of X-ray imaging techniques (Blamey *et al.*, 2000). Another possibility of detecting malignant development is the application of measurement of specific components in body fluids. An example for such a test is the estimation of prostate-specific antigen (PSA) levels in blood serum, which can indicate the presence of prostate cancer growth (Oesterling, 1991).

Under the circumstances that suspicious lesions, as for instance in a mammogram, or the presence of biomarkers, such as increased levels of PSA in blood, are detected tissue samples are removed from the concerned area. Obtained biopsy samples are analysed by pathologists applying different tissue fixation, sectioning and staining techniques (Kendall *et al.*, 2009). Accordingly, histopathology is the gold standard to finally confirm the presence or absence of cancer and furthermore to stage the present cancer. Examples for H&E (haematoxylin and eosin) stainings of a breast tissue sample are illustrated in Figure 2.1.

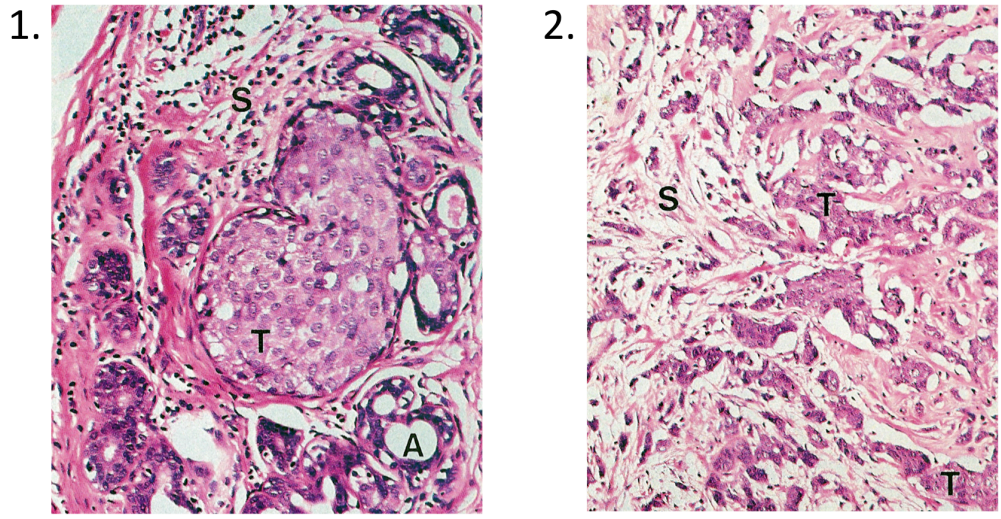


Figure 2.1 H&E stained sections of breast tissue. 1) Shows ductal carcinoma *in situ* (A: normal acini, S: stroma, T:tumour cells). 2) This staining shows invasive breast cancer (S: stroma, T: tumour) (Burkitt *et al.*, 1993).

Histopathology features several disadvantages, which includes that it results in a delayed diagnostic result and it also relies upon a subjective method, which can result in inter-observer disagreement (Kendall *et al.*, 2003, Montgomery *et al.*, 2001). Furthermore, excisional biopsy of vulnerable organs, including the central nervous system and vascular system, can be of increased hazard (Kendall *et al.*, 2009). In light of these limitations, an ideal diagnostic test would be rapid, non-invasive, high-throughput and would not require any tissue processing before analysis. Methods, including vibrational spectroscopic methods, such as Raman and infrared spectroscopy, have shown to be promising techniques for aiding histopathologists in the procedure of cancer detection and staging. These methods are discussed in detail below.

2.2 Emerging techniques

2.2.1 Biomedical photonics

Photonics deals with electromagnetic radiation, which can be defined as energy propagation by waves that feature electric properties as well as magnetic properties. The electromagnetic spectrum is generated by the extent of the energy, which in turn is proportional to the wavelength. The shorter the wavelength, the higher is the energy of an electromagnetic wave. An overview of the electromagnetic radiation is illustrated in Figure 2.2.

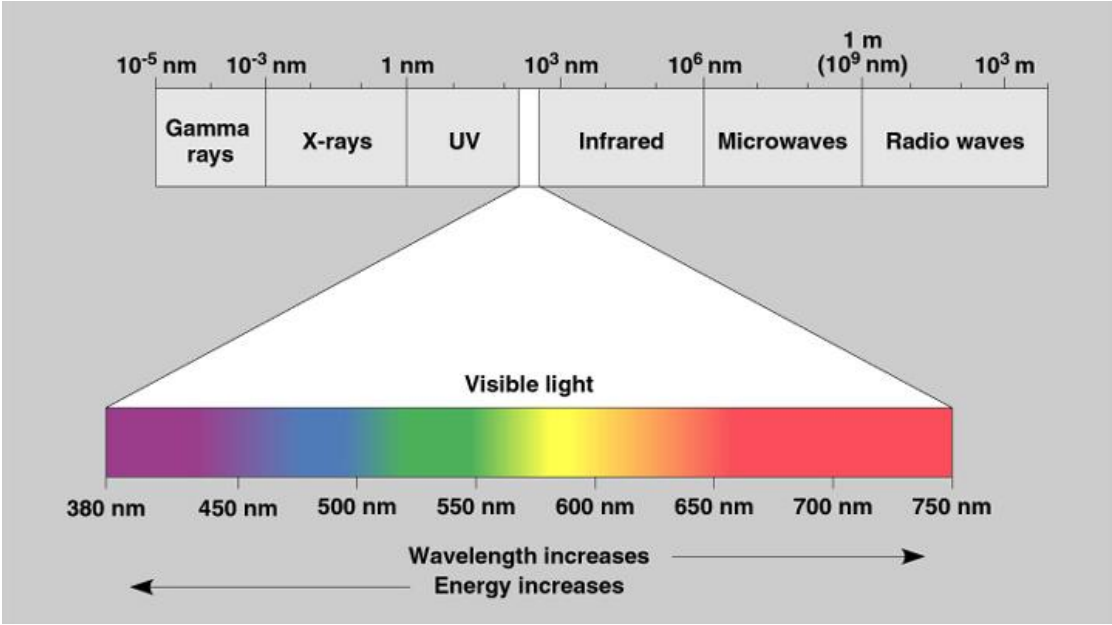


Figure 2.2 Electromagnetic spectrum.
(Adapted from: http://fig.cox.miami.edu/~cmallery/150/phts/spectra.htm)

The field of photonics can be split into optical and non-optical technologies. Whereas optical methods work with the visible light and non-optical methods use the much broader field of non-visible electromagnetic radiation. In that sense, biomedical photonics can be summarised as the research area and technology that uses the whole range of electromagnetic radiation for medical applications. For this reason electromagnetic radiation is applied in different ways, for instance absorption, emission, transmission, scattering, amplification and detection.

Photonic methods and technologies for medical application are lasers and other light sources, electro-optical instrumentation, fibre optics, microelectromechanical systems and also nanosystems. In general, photonic devices are applied for medical diagnostics, therapy and as well for the prevention of diseases (Vo-Dinh, 2003).

The application of spectroscopic techniques usually generates a large amount of data. Due to the size and complexity of spectral data obtained from tissue studies, computational methods are required for further downstream analysis. Therefore, the application of multivariate data analysis strategies is essential. Especially classification methods are of high interest, when applying spectroscopic techniques for the purpose of medical diagnoses. In order to make biophotonic methods applicable in clinical routine analysis, classifiers are needed, which can reliably differentiate cancerous from noncancerous tissue (Ellis *et al.*, 2006).

2.2.2 Vibrational spectroscopy

Atoms in a molecule are hold together by electron bonds. The relative positions of electrons and atom nuclei can change within the bonding orbitals. Commonly such a changed position is called vibrational mode and can only take arrangements as described by the quantum mechanic laws. A simplified illustration of vibrational modes for triatomic molecules, including symmetrical stretch, asymmetric stretch and bending deformation are shown in Figure 2.3. Molecules consisting of more than three atoms may have multiple complex vibrational modes (Hollas, 2002).

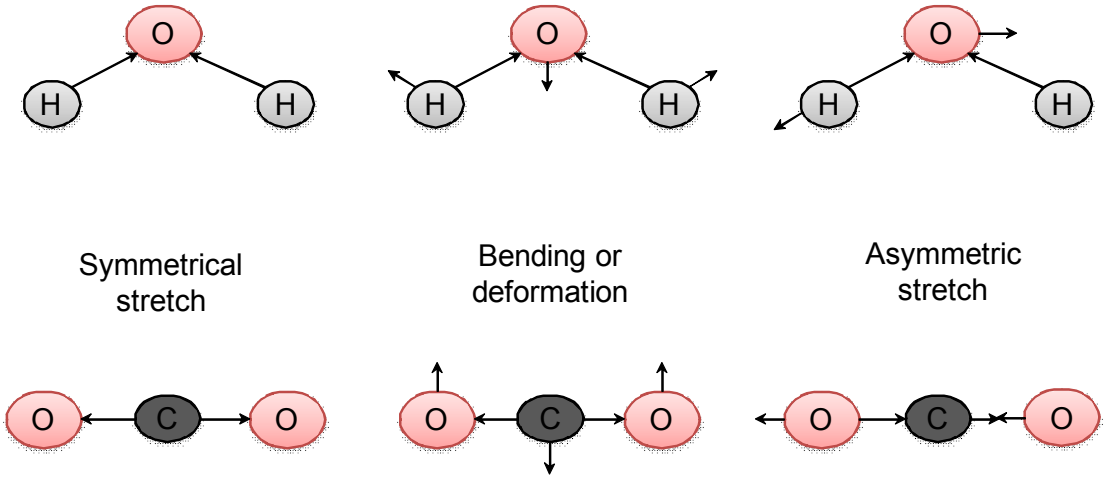


Figure 2.3 Vibrational modes of water and carbon dioxide: This spring and ball model illustrates the three possible vibrations of triatomic molecules. [Adapted from: Smith and Dent (2005)]

Vibration causes the nucleus to alter its relative position to the electronic cloud and consequently may result in a change of the dipole moment. Depending if a vibrational mode induces such a change of the dipole moment it can be differed between Raman-active and IR-active vibrational mode. Thus, Raman-scattering activity can be observed when the vibrational deformation does not result in a dipole alteration. On the other hand, a vibration that induces a dipole change is infrared absorption active (Pistorius, 1995). The difference between Raman active vibrations and IR-active vibrations is illustrated in Figure 2.4.

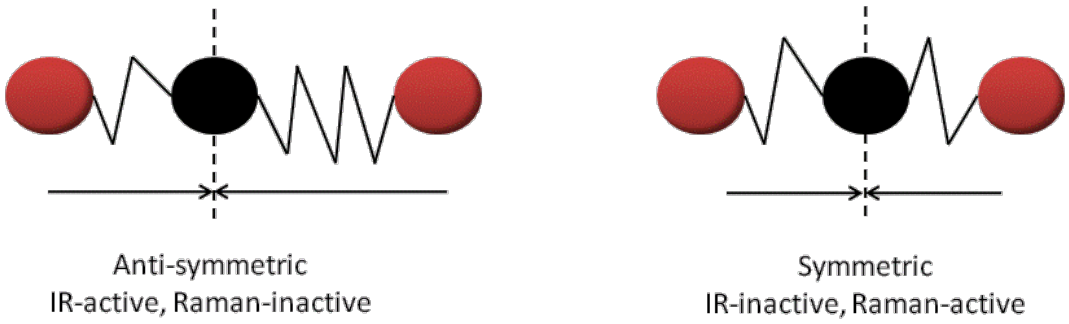


Figure 2.4 Stretching vibrations of CO_2: An anti-symmetric stretching leads to a change of the dipole moment which in turn results in an IR-active vibration. A symmetric stretching does not interfere with the dipole moment of the molecule and thus it causes a Raman-active vibration. [Adapted from: Pistorius (1995)]

Since both techniques, Raman scattering as well as and IR spectroscopy, show high potential for future diagnostic application they are described in more detail in the following.

2.2.3 Raman spectroscopy

2.2.3.1 History of Raman spectroscopy

The beginning of Raman spectroscopy can be dated back to the first quarter of the 20th century. Smekal, an Austrian quantum physicist, was the first to predict the inelastic scattering of monochromatic light in 1923 (Smekal, 1923). However, five years later Raman and his co-worker Krishnan were the first to actually observe this phenomenon (Raman *et al.*, 1928). Almost at the same time Landsberg and Mandelstam made the same observation independently in Moscow. In 1930 Raman received the Nobel Prize in Physics for the discovery of the scattering of monochromatic radiation. Since then this effect bears his name (Laserna, 2001).

After discovery further developments evolved slowly, this was influenced by several reasons. One of them was that the early experimental work was limited by the radiation source. Raman and Krishnan used filtered sunlight for their experiments and later mercury lamps became the standard radiation source. The invention of the laser in 1960 brought a significant upturn in the development of Raman spectroscopy. Another limitation in the early days was the absence of suitable electronic measuring devices. Once they were available many aspects, such as detection, data analysis and instrument miniaturisation, were improved (Laserna, 2001).

Nevertheless, for a long time IR spectroscopy was much more popular than Raman spectroscopy. The main reason therefore was that IR instruments have been commercially

available since the mid-1950s. A significant change occurred within the 1990s when Raman systems became simpler and smaller in size (Adar *et al.*, 2003). Since then Raman technology has been applied increasingly in many different areas, for instance for studies of polymers, inorganics and minerals as well as for biological, pharmaceutical and forensic applications (Smith *et al.*, 2005).

2.2.3.2 Raman theory

Several different interactions between light and matter, such as tissue, are possible and include absorption, transmission and scattering. The Raman effect is a light scattering phenomenon, which is a result of the change of the vibrational state of a scattering molecule of (Mahadevan-Jansen, 2003). Not every type of vibrational mode is Raman active as it has been described in more detail in section 2.2.2.

When light is scattered by a molecule each scattered photon predominantly possesses the same wavelength as the incident photon. This phenomenon is called elastic or Rayleigh scattering. The second mode is inelastic or Raman scattering. Under these circumstances the incident and the scattered photon features a different wavelength than the incident photon. However, only a very small number of photons, approximately 1 in 100 million photons, are inelasticly scattered (Zeng *et al.*, 2004). Such an inelasticly scattered photon can either gain or lose in wavelength. Accordingly, it is called Stokes-Raman scattering, when the wavelength of the scattered light is shorter and therefore the energy is higher than the incident photon. The opposite phenomenon is called anti-Stoke-Raman scattering (Petry *et al.*, 2003). The different scattering modes are illustrated in Figure 2.5.

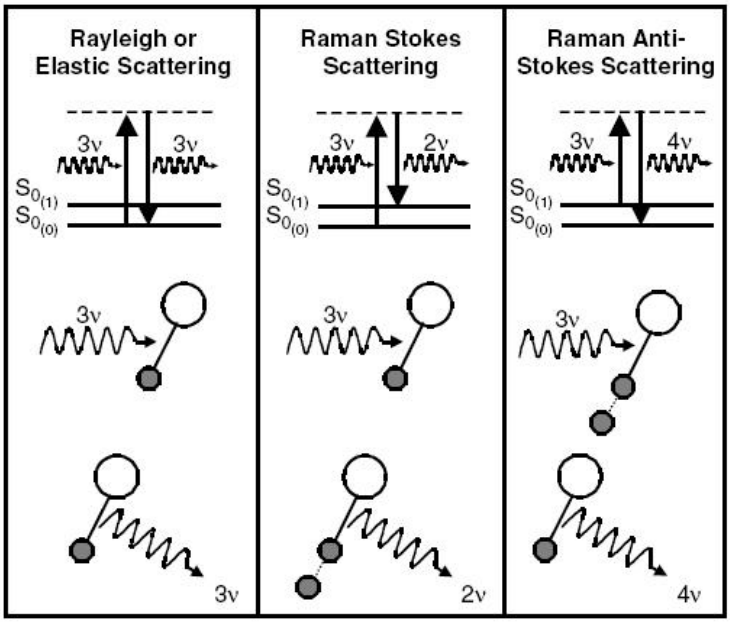


Figure 2.5 Illustration of the different scattering modes. $S_{0(0)}$ and $S_{0(1)}$ represent two vibrational states ($S_{0(1)}$ is higher in energy than $S_{0(0)}$) and ν the frequency of the incident photon. Thus, Elastic Scattering occures when the molecule returns to the same vibrational mode ($S_{0(0)}$) and the frequency of the incident photon does not shift. In comparison Raman Stokes Scattering occurs when the molecule returns to a higher vibrational state ($S_{0(1)}$) and the energy of the incident photon increases (2ν). Raman Anti Stokes Scattering occurs when the molecule returns to a vibrational lower state ($S_{0(0)}$) and the incident photon looses energy (4ν)[Adapted from: Mahadevan-Jansen (2003)].

As mentioned earlier, spontaneous Raman scattering is a very weak effect since only a very low number of photons are converted into Raman photons, since the efficiency is proportional to the fourth power of the frequency of the incident photon. A higher-frequency excitation source increases the number of scattered photons and therefore enhances the Raman scattering. Many organic substances and biological systems are fluorescent. Thus, a higher frequency also might stimulate molecules to fluoresce, which can mask the weak Raman scattering (Wartewig *et al.*, 2005). For tissue studies a high-frequency might not be favourable and due to the fact that higher-energetic photons may damage the sample through burning (Smith *et al.*, 2005).

2.2.3.3 Raman instrumentation

The basic Raman instrumentation does not differ drastically from any other spectroscopic system. The basic setup of Raman instrumentation can be dived into four main building blocks:

- A light source for excitation (traditionally a laser)
- A light delivery and collection system
- A wavelength selector (monochromator)
- A detection and processing unit (Ferraro *et al.*, 2003).

According to these building units the basic setup is illustrated in Figure 2.6.

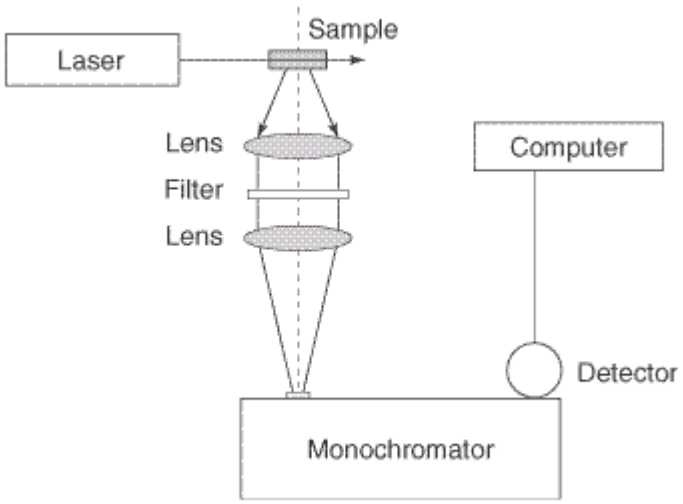


Figure 2.6 Basic Raman Instrumentation. [Adapted from: Popp and Kiefer (2003)]

The two most frequently applied Raman systems in biospectroscopy are Fourier transform (FT) Raman spectrometers (Hirschfeld *et al.*, 1986) and charge-coupled device (CCD)-based dispersive Raman spectrometers (Wang *et al.*, 1989). The main differences between these two systems are the applied laser and the methods by which the Raman scattering is detected and analysed (McCreery, 2000). Dispersive Raman spectrometers apply lasers that operate in

the ultraviolet, visible or near infrared region (Wartewig *et al.*, 2005). The most commonly applied detection system for dispersive Raman spectrometers is a charge coupled device. This multichannel device allows the simultaneous detection of a large spectral range (Popp *et al.*, 2003). On the other hand, FT-Raman spectrometers apply only near-infrared lasers, typically a neodymium:yttrium aluminium garnet (Nd:YAG) laser radiation at 1064 nm. These systems apply interferometric optics that also allows multiplex signal detection. A fast Fourier transform algorithm converts the resulting interferogram into a power density spectrum (Petry *et al.*, 2003).

Another well established technique is Raman microspectroscopy, which combines the properties of a Raman spectroscope with a microscope. Thus, a laser beam is focused by a microscope objective and allows pinpoint analysis as well as the generation of images. In a resulting image one point typically represents one Raman spectrum. A definite advantage of this method is that all kinds of objects can be analysed since it is possible to put almost any type of object under a microscope (Dhamelincourt, 2002).

2.2.4 Infrared spectroscopy

2.2.4.1 Theory of infrared spectroscopy

Infrared spectroscopy is a powerful method, which allows the qualitative and quantitative detection of many different types of materials and is applicable for solids and liquids as well as for gases (McKevly, 2000). Currently, infrared spectroscopy is one of the most important analytical methods (Günzler *et al.*, 2002). Accordingly, IR spectroscopy is applied in many areas, for example in polymer science, analysis of inorganic materials such as zeolites and metal oxides and analysis of semi-conductor structures. Furthermore, IR spectroscopy has

been highly investigated for life science research, such as pharmaceuticals (quality control and product monitoring), studies of blood and tissue, studies of the biological cell and cancer research (Meier, 2005).

IR spectroscopy belongs to the group of vibrational spectroscopic techniques. The energy of the IR incident radiation must correlate with the vibrational frequencies of the functional groups within the sample. If this pre-condition is met the molecular vibrations are stimulated. In addition, the molecular vibration must cause a change in the dipole moment in order to enable IR absorption. The different vibrational modes were outlined in more detail in section 2.2.2. The vibrational modes and the resulting IR absorption are very specific for a molecule and can be directly related to a (bio) chemical species. Thus, an infrared spectrometer allows the generation of a molecular fingerprint (Ellis *et al.*, 2006). Beside the facility of sample characterisation quantification is possible. This is possible due to the fact that the absorption is directly related to the concentration of a molecule within a sample (Jackson *et al.*, 2000).

The infrared spectral region is located between the visible light and the microwave. In addition, the infrared region can be divided into the near-infrared (NIR), mid-infrared (MIR) and far-infrared (FIR). The mid-infrared region ranges from 400-4000cm^{-1}. This region is of great importance for biomedical studies since the majority of molecules feature characteristic vibrations within this area. As a matter of fact this area is frequently referred to as the "fingerprint region". The near infrared region ranges from 4000-14,000cm^{-1} and typically leads to a broad and overlapping absorption. Due to that, previously only little attention was paid to this spectral region. However, the technology matured and this technique gained in importance in many research areas, including clinical and diagnostic analysis (Shaw *et al.*, 2000).

2.2.4.2 Infrared instrumentation

In general, the basic instrumental setup for infrared absorption spectroscopy is similar to other spectroscopic techniques. Accordingly a typical instrument setup, consisting of a light source providing the incident light, a spectral apparatus for spectral splitting, a detector and a computer measure the electromagnetic radiation transmittance of a sample of interest (Günzler *et al.*, 2002). A schematic design of an IR spectrometer is provided in Figure 2.7.

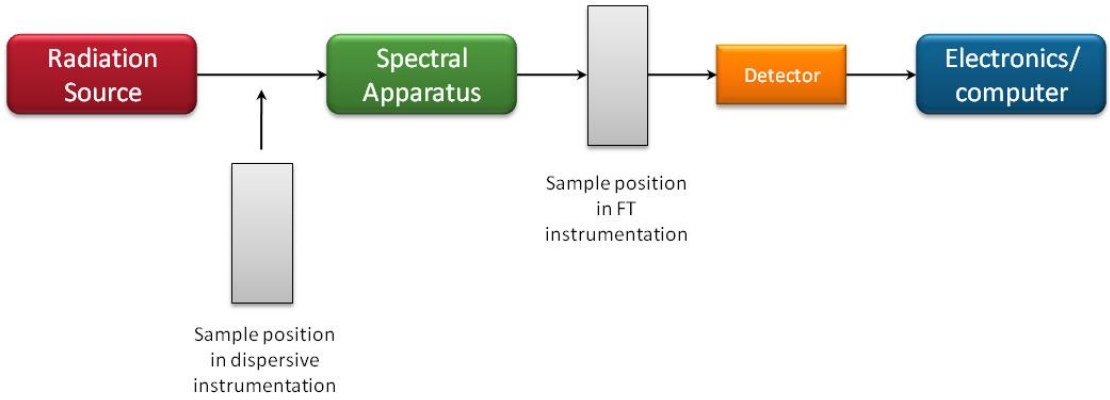


Figure 2.7 Schematic design of an IR spectrometer. [Adapted from: Günzler and Gremlich (2002)]

Infrared detectors are not capable of simultaneous differentiation between various wavelengths of light. Thus, a spectral apparatus is required for the separation of light into individual wavelengths. Depending on the spectroscopic apparatus, two main types of infrared spectrometers can be differed, dispersive instruments and FT instruments. Dispersive instruments use monochromators, for instance gratings or optical filters in order to select the wavelength which should reach the detector (Tomellini *et al.*, 2000). In contrast a FT spectrometer uses an interferometer for the splitting of the light. Commonly a Michelson two-beam interferometer is applied, which consists of two mirrors – one fixed and one moveable, and a beam splitter. The two split beams are reflected by the mirrors and recombined again, causing them to interfere (Günzler *et al.*, 2002). In that manner the detector detects an

interferogram, which is then mathematically transformed into a spectrum by applying the Fourier transformation algorithm (Wartewig *et al.*, 2005).

By now nearly all modern infrared spectrometers are Fourier transformation instruments since they show many advantages over dispersive instruments. FT instruments feature a higher robustness and are easier to handle. An FT infrared spectrometer can be combined with an microscope, which allows to generate infrared images of a sample with an spatial resolution in the range of ~10 µm (Meier, 2005). An FTIR microscope is illustrated in Figure 2.8.

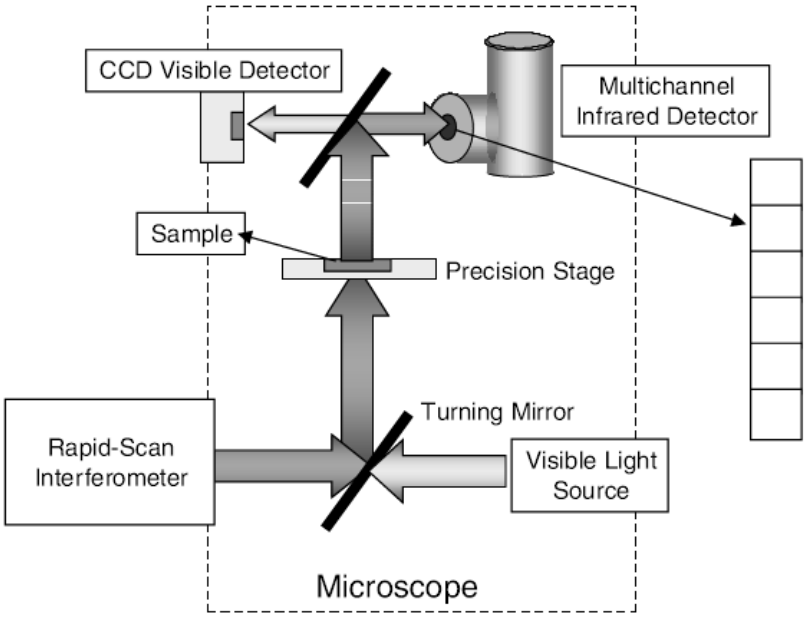


Figure 2.8 Schematic illustration of an FTIR micro spectroscope. [Adopted from: Bhargava and Levin (2003)]

2.2.5 Fluorescence spectroscopy

Fluorescence belongs to the field of luminescence spectroscopy. Generally, luminescence is caused when an electron goes from an electronic excited state into an electronic lower state (Skoog *et al.*, 1998). In order to elevate an electron into an excited state a molecule or an atom absorbs energy which is provided by photons at specific wavelength. For this purpose, near-ultraviolet or visible light is commonly used. Not every molecule that relaxes back to the ground state transmits fluorescence, instead they generate thermal energy. This phenomenon is called nonradiative transition. In comparison, molecules which are capable of radiative transition are called fluorophores (Norgaard *et al.*, 2007). Typically, the emitted fluorescence is longer in wavelength due to the fact that small amount of the energy is transformed into thermal energy. The emitted light is detected and undergoes further analysis (Ramanujam, 2000). The principles of fluorescence spectroscopy are illustrated in Figure 2.9.

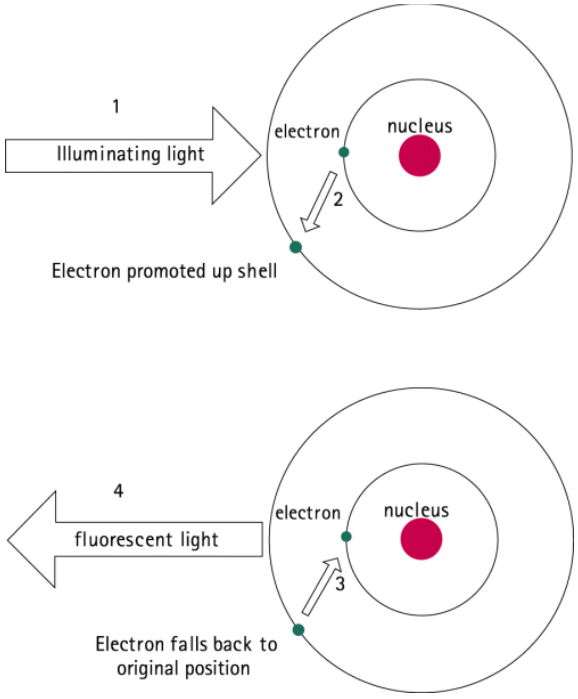


Figure 2.9 The principle of fluorescence spectroscopy: Light illuminates the molecule (1) and as a result the electron is elevated into an exited state (2). When the electron returns back to the original state (3) fluorescence is emitted (4). [Adopted from: Crow, Stone *et al.* (2003)]

Fluorescence techniques can use the features of endogenous fluorophores or exogenous fluorophores. The first category uses the ability of autofluorescence, which can be observed in many organic molecules. Components that are not autofluorescent can be aggregated with an exogenous fluorophore in order to become fluorescent and detectable (Bigio *et al.*, 1997).

Biological tissue contains several autofluorescent components, for instance amino acids, structural proteins, enzymes and co-enzymes, vitamins lipids and porphyrins (Ramanujam, 2000). The biochemical composition of a cell varies during the different states of disease and as a result the concentration of fluorophores can change. This can be explained by the fact that the disease development interferes with the cellular metabolism, which can even prevent production of the individual fluorescent components. In other cases the distribution of fluorophores may vary between diseased and normal tissue (Vo-Dinh *et al.*, 2003). Due to that fluorescence spectroscopy is highly capable of malignancy detection and was investigated for the common types of cancer, among them are cervical cancer (Mahadevan *et al.*, 1993), head and neck cancer (Schantz *et al.*, 1998), colorectal cancer (Mayinger *et al.*, 2003) and oral cancer (De Veld *et al.*, 2005).

2.2.6 Elastic scattering spectroscopy

As outlined earlier there are different ways in which a photon can interact with a molecule and thus another possible type of interaction between an incident photon and a molecule is elastic scattering. In this light interaction the incident photon is reflected by a molecule and does not experience any change of energy. Accordingly, elastic scattering spectroscopy (EES) detects photons that were scattered by the sample molecules. The elastic scattering capability of a sample increases with its refractive index or its density (Crow *et al.*, 2003).

In biological samples, such as tissue, the nucleus contributes strongly to cellular light scattering features. Carcinogenesis is accompanied with the change of the nucleus, which includes an increased nucleus and a change in the structure of the nucleus. Thus, EES may be used to detect and diagnosed cancer and was explored for this purpose by several groups (Mourant *et al.*, 2003). Lovat, Johnson *et al.* (2006) showed that the EES is capable of detecting high grade dysplasia and cancer within Barrett's oesophagus *in vivo*. EES was also successfully investigated for the detection of skin cancer by differing between primary melanomas and benign nevi *in vivo* (Marchesini *et al.*, 1992), detection of cervical cancer *in vivo* (Mourant *et al.*, 2007), for the assessment of bony resection margins in oral cancer (Jerjes *et al.*, 2005) and the detection of bladder cancer *in vivo* (Mourant *et al.*, 1995).

3 Machine learning in vibrational spectroscopy

3.1 Classification theory

Machine learning derives from the idea of learning by experience. Thus, an algorithm is trained to distinguish groups of a predefined data set where the class of each sample is known. This data set is commonly called the training set. This data set is used to establish a mathematical model, which in turn should be capable of predicting the class membership of unseen samples (Izenman, 2008). In order to assess the predictive power of the resulting model it has to be tested, ideally with unseen data. A common approach is to split the available data into training and testing sets before building the model. Thus, a part of the data is kept untouched during the procedure of the model optimisation (Mosteller *et al.*, 1977).

For the model optimisation (also termed training), two methods are frequently applied, cross-validation (Stone, 1974) and the bootstrap (Efron, 1979). For K-fold cross-validation the training set is randomly separated into *K* groups. It is important that these groups do not overlap. The sub-training set is formed by *K*-1 groups and the left out group is used for validation. This procedure can be repeated *K* times and the average of the prediction error can then be used to assess the resulting model (Izenman, 2008). In contrast, the bootstrap randomly selects samples from the parental data set in order to generate a validation set. This procedure is repeated several times, for example 200 times, and each time the prediction error is calculated. The average of the resulting prediction error values is used to assess the model quality (Brereton, 2007). Typically, cross-validation and bootstrapping are used for model optimisation but under the circumstances that the available data set is too small to be split into train and test set these methods can also be used for performance testing.

Once the model is optimised it can be assessed with the left out testing data and a confusion matrix is established as shown in table 3.1.

Table 3.1 Confusion matrix scheme [Adopted from: Fielding (2007)]

		Actual class	
		1	2
Predicted class	1	Correct True positive a	Incorrect False positive b
	2	Incorrect False negative c	Correct True negative d

According to the confusion matrix the correct classification rate is calculated as follows (Fielding, 2007):

$$\text{Correct classification rate} = (\text{True positive} + \text{True negative})/\text{Number of samples} = (a+d)/N$$

For diagnostic approaches it is also common to express sensitivity and specificity. In a binary classification, such as benign and malignant, two types of errors can be found, false positive and false negative. Sensitivity is a measurement for the true positive cases and the specificity is a measurement for the true negative cases (Altman *et al.*, 1994). These two values are calculated as follows (Fielding, 2007):

$$\text{Sensitivity} = \text{True positive}/(\text{True positive} + \text{False negative}) = a/(a+c)$$

$$\text{Specificity} = \text{True negative}/(\text{True negative} + \text{False positive}) = d/(d+b)$$

3.2 Pattern recognition methods

One of the main interests of data analysis for medical applications is to identify patterns or groupings within a data set. Pattern recognition methods can be divided into different groups, the major ones are exploratory data analysis (EDA), unsupervised pattern recognition and supervised pattern recognition (Brereton, 2007).

3.2.1 Exploratory data analysis

Principal component analysis (PCA) is probably the best known exploratory data analysis technique. This method reduces the dimension of a given data set and creates new variables, called principal components. In many cases only two or three principal components are sufficient for capturing the major variance within data. Plotting of the principal components reveals similarities and differences within the data (Lavine, 2000). For this reason, PCA can be used to examine if Raman or infrared spectra, derived from tissue samples, can be grouped into cancerous and non-cancerous states as illustrated in Figure 3.1. Further information on PCA and how principal components are calculated can be found in Brereton (2007).

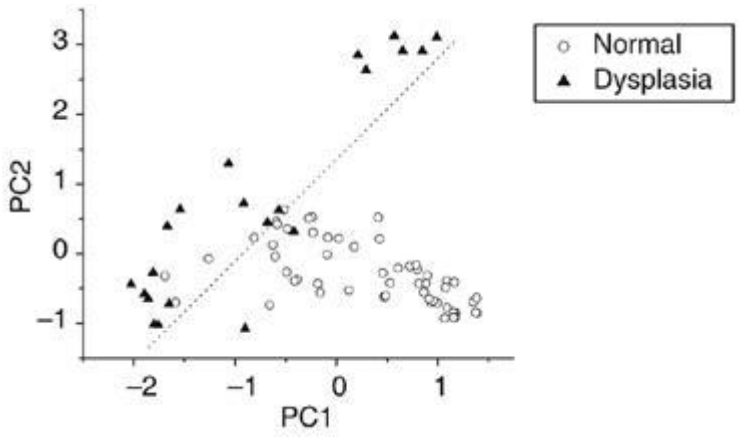


Figure 3.1. Scatter plot of the first and the second PC of Raman spectra. The circles represent normal gastric tissue and the triangles dysplastic gastric tissue. The grouping of the two classes is highlighted by the line separating them. Adapted from: Teh, Zheng et al. (2008b)

3.2.2 Unsupervised methods

Unsupervised pattern recognition is frequently referred to as cluster analysis. In general, cluster analysis aims to discover groupings within a data set by drawing a picture in order to uncover similarities (Brereton, 2007). Cluster analysis methods can be divided into hierarchical and non-hierarchical (Izenman, 2008). Out of the available methods hierarchical clustering tends to be the most commonly applied. The first step in hierarchical clustering is to establish a similarity matrix by calculating the distance between the samples (Lavine, 2000). Different algorithms are available for calculating the distance, for instance Euclidian, Manhatten and Minkowski. After generating the similarity matrix the samples need to be joined together into clusters. For this purpose different linkage methods are available, for example single-linkage, complete linkage or average-linkage (Izenman, 2008). The resulting clusters are then illustrated as a dendrogram, but also phylograms or cladograms can be generated (Brereton, 2007). A dendrogram hierarchal cluster of infrared spectra is shown in Figure 3.2. More details on the different clustering methods can be found in Izenman (2008).

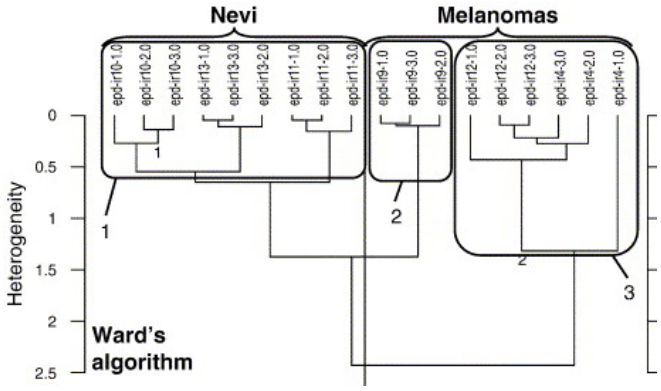


Figure 3.2 Hierarchical dendrogram cluster of infrared spectra: This dendrogram shows the grouping of IR spectra obtained from skin tissue samples. The first cluster contains all the benign nevus samples where cluster 2 and 3 group all the malignant samples. [Adapted from Tfayli, Piot et al.(2005)]

3.2.3 Supervised classification methods

3.2.3.1 Linear discriminant analysis

Linear discriminant analysis (LDA) is a frequently applied classification method due to its simplicity. This classifier produces a linear boundary between classes. For the calculation of the LDA distance to each class, Mahalanobis distance is commonly applied, which defines the distance of an sample x to class A, where $\boldsymbol{S}_A$ is the variance covariance matrix of all training samples belonging to class A:

$$d_A^2 = (x - \bar{x}_A)\mathbf{S}_A^{-1}(x - \bar{x}_A)' \quad (3.1)$$

Frequently, principal component analysis (PCA) is executed prior to building a LDA model. The resulting principal component (PC) scores are then used to generate the LDA model. Using PCs allows simplification of the data by maintaining the overall information content despite using fewer variables. Using a reduced data set is of special importance if the observed data has a higher number of variables than the number of samples due to the fact that Mahalanobis distance fails under these circumstances (Brereton, 2009). In this manner, the optimisation of the LDA model includes the estimation of the ideal number of PCs fed into the LDA. This is commonly done by leave one sample out cross validation (LOOCV), where one sample is left out and the remaining data is used to build a model, which is then used to predict the class membership of the left out sample.

3.2.3.2 Partial least square discriminant analysis

Partial least squares (PLS) has a long tradition in chemometrics. Similar to PCA, PLS is a data reduction method. The main difference between these two methods is that PLS tries to relate the two types of variables, in this case the spectral data and the pathology class. Thus, PLS attempts to maximise the covariance between these two building blocks. Typically, PLS is presented as follows:

$$\mathbf{X} = \mathbf{T}.\,\mathbf{P} + \mathbf{E} \tag{3.2}$$

$$\mathbf{c} = \mathbf{T}.\,\mathbf{q} + \mathbf{f} \tag{3.3}$$

X represents the measurements (spectra) and ***c*** the classes. The score matrix ***T*** models the measurements as well as the classes ***c*** and is common in both equations. The PLS loadings are represented by ***P*** in Equation 3.2 and ***q*** in Equation 3.3. Finally, ***E*** is an error matrix and ***f*** and error vector respectively. Commonly a PLS-DA model is optimised by estimating the number of PLS components (latent variables), which can be done for instance by cross-validation or bootstrapping (Brereton, 2007).

3.2.3.3 Support vector machines

Support vector machines (SVMs), which were first introduced by Vapnik (1995), are a relatively new member in the community of machine learning methods. SVM theory can be traced back to structural risk minimisation (SRM), which aims to estimate a classification decision function by minimising the empirical risk *R*. For a two class problem data $\boldsymbol{X} = \{x_1,\ldots x_i\}$ and $y_i \in \{1,-1\}$ this can be expressed as:

$$\mathrm{R} = \frac{1}{\mathrm{L}} \sum_{i=1}^{L} \left| f(\mathbf{x}_\mathrm{i}) - \mathrm{y}_\mathrm{i} \right| \tag{3.4}$$

Where *L* represents the number of the samples and *f* the decision function. In the simplest case, a linear separable problem, a linear decision function can be determined to separate the two classes:

$$f(\mathbf{x}) = sign(\mathbf{w} \cdot \mathbf{x} + \mathrm{b}) \tag{3.5}$$

The weight ***w*** and the bias *b* must be determined from the training set.

SVMs not only aim to separate data by a hyperplane that gives a low generalisation error, they also aim to maximise the margin between the different classes. In order to achieve this, a separation hyperplane must be optimised by satisfying the following conditions:

$$\min \frac{1}{2}\|\mathbf{w}\|^2 \tag{3.6}$$

subject to $y_i(\mathbf{w}\cdot \mathbf{x_i} + b) \geq 1$

Since not all classification problems are linearly separable, the minimisation problem needs to be modified. This extension introduces a soft margin that allows data points to be misclassified, but penalises resulting errors. This is achieved by a penalty parameter *C*, which is the trade-off between error ξ and the margin. Thus the modified minimisation problem can now be generalised as:

$$\min \frac{1}{2}\|\mathbf{w}\|^2 + \mathrm{C}\sum_{i=1}^{L}\xi_i \tag{3.7}$$

subject to $\xi_i > 0,\ \mathrm{y_i}(\mathbf{w}\cdot \mathbf{x_i} + \mathrm{b}) \geq 1 - \xi_i$

A further extension for non-linear separable problems is the application of kernel functions. A kernel function is a non-linear function that maps all data points into a higher dimensional feature space. This allows us to overcome the restriction that data points might not be separable in the original input space. The most frequently applied kernel function is the radial basis function:

$$K(\mathbf{x_i}, \mathbf{x_j}) = \exp\frac{-\left\|\mathbf{x_i} - \mathbf{x_j}\right\|^2}{2\sigma^2} \tag{3.8}$$

Originally, SVMs were designed as binary classification method, however in order to extend them to multiclass problems several methods are applied. Nonetheless, the most popular methods are the 'one-against-all' (OAA) and the 'one-against-one' (OAO) approach. Both methods split the multiclass problem into a series of binary problems. Therefore, the OAA method generates for a *N* class problem, *N* binary classifiers, one for each class. Every single SVM is then trained to separate samples of one class from the remaining samples. The finally assigned class corresponds to the SVM with the highest decision value (Vapnik, 1998). In

comparison, the OAO method generates $N(N\text{-}1)/2$ SVMs, which is equivalent to one SVM for each pair of classes. In order to get the final prediction from all the individual classifiers a voting strategy, commonly maximum voting, where each SVM votes for one class, is applied (Milgram *et al.*, 2006). Multiclass SVMs are therefore a particular implementation of an ensemble SVM system in which a multiclass problem is split into several binary problems.

3.2.3.4 Artificial neural networks

Artificial neural networks (ANNs) are an attempt to simulate biological neural networks. Real neural networks are composed by high number of neurons, which are connected with each other but independent. Thus an ANN consists of several simple processing elements termed nodes or neurons. The function of each node is to convert the input values to a bounded output value. The function a node uses for this is called a transfer function, which can be for instance a sigmoid function. Several types of ANN architectures are available, where an ordinary feed forward network is made up of three layers as shown in Figure 3.3 (Fielding, 2007). ANNs are considered a good technique when the underlying structure of the input data is not well known (Hand *et al.*, 1997). On the other hand ANNs lack transparency since they do not allow insight in how classification results are generated and due to that they are often considered to be 'black box' classifiers, though information about the most significant input variables can be determined using clamping (Green *et al.*, 2009).

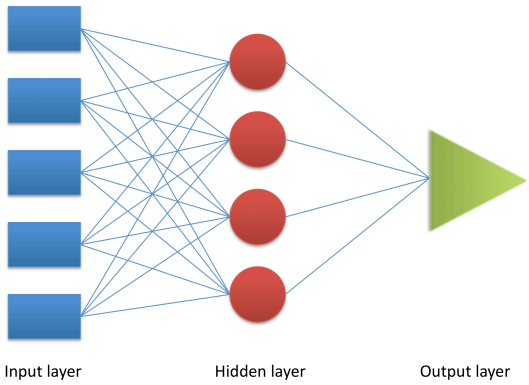


Figure 3.3 General illustration of a three layer neural network [Adapted from Fielding (2007)].

3.2.3.5 Ensemble methods

A frequently encountered problem in machine learning is that classifiers that achieve a good training performance frequently exhibit a low generalisation performance on unseen data. Indeed, classifiers achieving similar training performance can result in varying prediction of unseen data. A possibility to overcome these limitations is to create an ensemble consisting of several classifiers and average the output of all independent classifiers (Polikar, 2006). An ensemble can be built for any kind of classifier, including LDA, PLS-DA, SVM and ANN.

3.2.3.6 Random forest

Random forest, which was first proposed by Breimann (2001), belongs to the group of ensemble classifiers in which the output of several classification trees is combined. In random forests different classification trees are grown without pruning. In order to grow different trees the several train subsets a generated *via* bootstrapping. Consequently, each bootstrap set is used to grow a tree by randomly selecting a predefined subset of variables at each node. The output of all individual classification trees is typically combined by a majority vote (Breiman, 2001).

3.3 Classification approaches for cancer diagnostics

Malignant development can occur in various cell types of the human body and accord to the cellular origin cancers are separated into groups. Carcinomas, representing the most common type of cancer, start developing in epithelial cells, which build covers and cavities, including glands throughout the human body. Cancer can also develop in the connective tissue and bones, blood, lymph and nervous system. In this section machine learning approaches for cancer diagnostics using vibrational spectroscopy are presented for the most common types of cancer.

3.3.1 Epithelial cancers

3.3.1.1 Lung cancer

Lung cancer is one of the most common types of cancer and causes the highest number of deaths among all cancer types (Cancer Research UK, 2010c). Near infrared Raman spectroscopy was investigated by Huang, McWilliams *et al.* (2003) for lung cancer diagnostics. They analysed a total of 28 bronchial tissue specimens (12 normal, ten squamous cell carcinoma and six adenocarcinoma) derived from ten patients. They report a diagnostic sensitivity of 96% and a specificity of 92% estimated by comparing by comparing the ratio of two specific spectral features (1445:1655 cm^{-1}). In a similar manner FTIR microspectroscopy was applied for lung cancer diagnostics based on peak comparison (Yano *et al.*, 2000).

In more recent work Raman microspectroscopy has investigated for the diagnosis and prognosis prediction of nonsmall cell lung cancer. A total of 62 lung tissue samples (28 normal, 34 cancerous) derived from 43 patients were analysed. A random forest classification model was developed and assessed with an independent test set. This model yielded a

diagnostic sensitivity of 90% and a specificity of 75%. For predict postoperative cancer recurrence PCA was investigated, which achieved sensitivity of 73% and a specificity of 74% respectively (Magee *et al.*, 2010). PCA was also applied for *ex vivo* diagnosis of lung cancer by using a Raman miniprobe, which is suitable for possible future *in vivo* application. In this study lung tissue specimens derived from seven patients were investigated (Magee *et al.*, 2009).

Raman images have been explored for the characterisation of normal bronchial tissue section in order to gain further understanding of biochemical changes accompanying cancer development. For this imaging study 12 Raman maps were analysed by K-mean clustering. Spectra featuring similar characteristics where grouped into clusters and the generated maps were compared with histopathology sections (Koljenovic *et al.*, 2004).

3.3.1.2 Gastrointestinal cancer

Vibrational spectroscopy has been increasingly investigated for epithelial cancers of the gastrointestinal (GI) tract, including oesophageal, stomach and colon cancer. The great attention for these types of disease can be explained by the fact that the GI tract is easily accessible for future *in vivo* applications.

Oesophageal cancer:

In several works the capability of Raman spectroscopy as a prospective tool for cancer diagnostics was demonstrated. In the majority of works LDA was used for building diagnostic classification models. For example, Kendall, Stone *et al.* (2003) built a PC-fed LDA model for the prediction of oesophageal pathology based on spectra obtained from 87 histopathologically homogeneous samples (44 patients). The eight-class LDA model

achieved sensitivities of 73-100% and specificities of 90–100% when tested with leave one out cross-validation (LOOCV). In a similar approach LDA achieved sensitivities of 84–97% and specificities of 93–99% in a three-group classification approach when assessed by LOOCV (Stone *et al.*, 2002). In a more recent approach a novel Raman probe design has been investigated for the potential of *in vivo* diagnosis. For this work 114 oesophageal biopsy samples were collected from 45 patients and measured using varying acquisition times (2 and 10 seconds). The gathered spectra were used to develop a PC-fed LDA model, which predicted the state of samples from an independent test set with a sensitivity of 66–84% and specificity of 81–96% (Kendall *et al.*, 2010).

FTIR spectroscopy in attenuated total reflectance has been investigated for premalignant mucosa. An LDA model achieved a sensitivity of 92% and a specificity of 80% when tested by LOOCV (Wang *et al.*, 2007a). Quaroni and Casson (2009) generated infrared maps from different oesophageal pathology types and consequently applied hierarchical clustering (HC) for studying different areas in biopsy samples.

Stomach cancer

For the implementation of classification models for stomach cancer prediction various types of classifiers have been investigated. Thus, for instance a PC-fed LDA model built for classifying dysplasia from normal gastric tissue based on Raman spectra gather (44 patients, 76 specimens) achieved a sensitivity of 95.2% and a specificity of 90.9% when assessed by LOOCV (Teh *et al.*, 2008b). Similar, Kawabata, Mizuno *et al.* (2008) applied a PCA based discriminant analysis for differentiation between cancerous and non cancerous Raman spectra. In this work only the measured intensity of the Raman shift at 1644 cm^{-1} was considered, which achieved a sensitivity and specificity of 70%. A different classification

approach was taken by Teh, Zheng *et al.* (2008a), who investigated classification and regression trees (CART) for differentiating between normal and cancerous gastric tissue specimens. For the model development the gathered Raman data, 73 tissue samples from 53 patients, were split into a train and a validation set. A sensitivity of 88.9% and a specificity of 92.9% have been estimated when tested with the independent validation set. A three-class model for diagnosing and typing ardenocarcinoma in the stomach was built by multinomial logistic regression (MNLR). This model predicted the pathology of 125 tissue specimens (72 patients) with sensitivities between 75-91% and specificities between 80-96% when assessed by LOOCV (Teh *et al.*, 2010).

Soft independent modelling of class analogies (SIMCA) was employed for prediction of three different stomach pathologies (normal, adenoma and cancer) based on IR spectroscopic measurements. Although the data set was small, consisting of only 11 patient samples, an independent test set was used to evaluate the classification model. The SIMCA model achieved predictive accuracies of 77% for normal samples, 30% for adenoma samples and 87% for cancer samples (Park *et al.*, 2007). An LDA model was developed by Li, Sun *et al.* (2005) for predicting four different stomach tissue pathologies (healthy, superficial gastritis, atrophic gastritis, and gastric cancer). The developed model achieved accuracies of 90% for healthy samples, 90% for superficial gastritis samples, 66% for atrophic gastritis samples and 74% for cancerous samples when assessed by LOOCV.

Colorectal cancer

The application of vibrational spectroscopy has been increasingly investigated for colorectal cancer, due to the fact that colorectal cancer is one of the most common types of cancer. Although, many Raman studies were dedicated to this type of cancer only a small number of

classification were reported. For instance Widjaja, Zheng *et al.* (2008) developed multi-class SVM models for predicting colon pathology (normal, hyperplastic polyps and ardenocarcinoma). In this study 105 tissue specimens from 59 patients were analysed by Raman spectroscopy and the gathered spectra were consequently used to establish an SVM model. A RBF SVM model achieved an overall accuracy of 99.3% (normal), 99.4% (hyperplastic polyps) and 99.9% (ardenocarcinoma) when tested by LOOCV. K-mean clustering has been investigated for image analysis of Raman microspectroscopic maps obtained from normal colonic tissue. In this study characteristics and variances between different tissue sections were investigated and the findings compared with maps obtained by coherent anti-Stokes Raman scattering (CARS) microspectroscopy (Krafft *et al.*, 2009).

Currently, the application of classification methods for studying FTIR maps obtained from colonic tissue focused on image analysis. Cluster Analysis (CA), PCA and artificial neural networks (ANNs) have been investigated for differentiation of types of tissue structure (Lasch *et al.*, 1998). In a more recent approach agglomerative hierarchical (AH) clustering (Ward's technique), fuzzy C-means (FCM) clustering, and k-means (KM) clustering have been applied to analyse infrared maps of ardenocarcinoma tissue sections. The correlation of infrared maps processed by hierarchical clustering and histopathology was demonstrated (Lasch *et al.*, 2004).

Krafft, Codrich *et al.* (2008) measured colon tissue sections with Raman as well as FTIR spectroscopy. Maps obtained from both techniques were subjected to KM cluster analysis. This study demonstrated that maps from both techniques coincided remarkably well.

3.3.1.3 Urological cancer

Prostate cancer

In the UK prostate cancer is the most commonly diagnosed cancer in men (Cancer Research UK, 2010a). In order to develop new diagnostic approaches several groups have investigated vibrational spectroscopy for this purpose. For instance Stone, Kendall *et al.* (2004) investigated Raman spectroscopy for differing between benign and neoplastic tissue samples. The developed LDA model achieved a sensitivity of 96% and a specificity of 91% when test by LOOCV. Similarly, LDA was applied to distinguish benign from malignant prostate sample (37 patient samples) measured with a fiberoptic probe, as suitable for laparoscopic and endoscopic use. A sensitivity of 87% and a specificity of 84% were achieved when the model was assessed by LOOCV (Crow *et al.*, 2005).

Infrared spectroscopy was also applied for grading of prostate cancer tissue specimens. In a study of 39 patients, classification models using a principal component discriminant function analysis achieved an overall sensitivity of 92.3% and a specificity of 99.4% when assessed with an independent test set (Baker *et al.*, 2008). Furthermore, Gazi, Baker *et al*. (2006) investigated infrared spectroscopy and LDA modelling for grading of prostate cancer. The resulting classifier was tested with an independent test set and achieved sensitivities between 70-78% and specificities between 81-89%. In comparison, both grading approaches yielded a lower specificity than sensitivity. Infrared spectroscopic images and genetics-based machine learning was studied for computer aided histopathology of prostate tissue. The developed algorithm was able to classify pixels, which represent different tissue areas with an accuracy up to 90% when assessed by a 10-fold cross-validation based on 20 patient samples (Llora *et al.*, 2009).

Bladder cancer

Several studies have taken the approach of applying of Raman spectroscopy for bladder cancer diagnostics. The predominantly applied classification algorithm was LDA, which was for instance used to develop a diagnostic model to discriminate between nontumour and tumour bladder tissue by de Jong, Schut *et al.* (2006). The resulting model, which was build of Raman data obtained from 15 patient samples, yielded a sensitivity of 94% and a specificity of 92% when tested by LOOCV. In a similar approach 24 patient samples, representing normal urothelium, cystitis and transitional cell carcinoma tissue samples were used to develop a diagnostic LDA model. This classifier achieved a sensitivity of 89% and a specificity of 79% when tested by LOOCV (Crow *et al.*, 2005). In more recent work the combined application of fluorescence with Raman spectroscopy was investigated for diagnostic prediction of bladder biopsies. The employed LDA model, build on data derived from 38 patient samples, achieved a sensitivity of 42.6 % and a specificity of 71.1% when assessed by LOOCV (Grimbergen *et al.*, 2009).

In a small study bladder samples obtained from three patients were investigated for the differentiation between normal and cancerous tissue based on phosphate bands (Romano *et al.*, 1995). Beside this, till present no further classification approaches based on infrared spectroscopic data and images have been reported.

3.3.1.4 Breast cancer

Raman spectroscopy has been investigated for breast cancer tissue analysis for almost 20 years (Alfano *et al.*, 1991). In a Raman study investigating *ex vivo* samples from breast tissue (normal, fibrocystic change, fibradenoma and invasive cancer) a logistic regression was

employed to differ between malignant and benign spectra. The model yielded a sensitivity of 94% and a specificity of 96% when tested by LOOCV (Haka *et al.*, 2005). The same algorithm was further investigated for the capability of classifying fresh resected tissue samples mimicking and *in vivo* application. Thus, 129 tissue sites from 21 patients were measured and their pathology predicted by the logistic regression mode. A sensitivity of 83% and a specificity of 93% were reported (Haka *et al.*, 2009). A different classification approach was taken by Moreno, Raniero *et al.* (2010), who employed quadratic discriminant analysis (QDA) for distinguishing invasive ductal carcinoma (22 patients), fibrocystic breast conditions (six patients) and normal breast tissues (six patients). The QDA model separated normal from altered tissue with an accuracy of 98.5%.

In an early FTIR approach Dukor, Liebman *et al.* (1998) investigated a LDA model for discrimination between benign, hyperplasia and malignant breast tissue specimens derived from one patient. Multiple two-class models were generated (benign *vs.* malignant, malignant *vs.* hyperplasia and hyperplasia *vs.* benign) and tested by LOOCV. The classifiers achieved accuracies of 90-100%. In a more recent approach ANNs were used to distinguish between infrared spectra representing four different breast tissue types, fibroadenoma, ductal carcinoma in situ, connective and adipose tissue. The ANN was tested with an independent test set, consisting of seven patient samples, and achieved accuracies between 85-100% (Fabian *et al*., 2006). In an infrared micro-spectroscopic imaging study cluster analysis was used to examine benign breast tumor tissue specimens. The maps, generated by the cluster analysis, were compared with the corresponding histopathology staining slides and it showed that this methodology allows differentiation between benign and malignant tumors types (Fabian *et al*., 2003).

Micro-calcifications are commonly found in breast tissue and often and indicator for malignant disease development. In an effort to exploit this, Haka, Shafer-Peltier *et al*. (2002) investigated Raman spectroscopy and logistic regression for predicting malignancies in breast tissue based on micro-calcifications. Spectra, derived from 11 patient samples, were classified with a sensitivity of 88% and specificity of 93%. Infrared spectroscopy was also investigated for the potential to diagnose breast pathology based on micro-calcifications. Pathology specific patterns (carbonate content and protein matrix: mineral ratios) were used to generate a two-matrix linear discriminant model for differentiating between benign, ductal carcinoma in situ and invasive malignancies. In this study sensitivities of 79-90% and specificities of 82-98% were reported (Baker *et al.*, 2010b).

3.3.1.5 Cervical cancer

NIR Raman spectroscopy in combination with LDA modelling has been investigated for *in vivo* diagnostics of cervical cancer. The classifier built by using spectra derived from 46 patients yielded a diagnostic sensitivity of 93.5% and specificity of 97.8% when tested by LOOCV (Mo *et al.*, 2009). In a different approach Raman spectroscopy was applied for measuring normal, cervical intraepithelial neoplasia and invasive carcinoma tissue samples from 40 patients. A LOOCV achieved sensitivities ranging from 98.5 to 99.5% and specificities from 99.0 to 100% (Krishnaa *et al.*, 2006).

FTIR spectroscopy and SVM classification were investigated for the differentiation between normal and dysplasia of cervix biopsies and furthermore for grading of dysplasia samples. An overall accuracy of 72% was reported (Njoroge *et al.*, 2006). Infrared spectroscopy has also been applied for image analysis of cervix tissue samples. Steller, Einenkel *et al.* (2006) used fuzzy C-mean clustering and hierarchical cluster analysis for identifying morphological

characteristics in infrared maps. The resulting maps were compared to the correlating H&E slides and this showed that a differentiation between basal layer, dysplastic lesions and squamous cell carcinoma was possible. Furthermore, hierarchical cluster analysis was applied for distinguishing between normal and diseased tissue based on infrared maps. It showed that the generated cluster maps correlated to the H&E stainings (Wood *et al.*, 2004).

3.3.1.6 Skin tumours

Skin is easily accessible and for this reason most suitable for *in vivo* diagnostics using vibrational spectroscopy. Raman spectroscopy and sparse multinomial logistic regression were used to distinguish between normal, basal cell carcinoma, squamous cell carcinoma and melanoma. In this study, based on 39 patients, an overall sensitivity and specificity of 100% was reported (Lieber *et al.*, 2008a). Based on the previous study, a Raman handheld probe was developed and used to measure skin samples in 19 patients *in vivo*. Sparse multinomial logistic regression was employed to differ between normal and abnormal (basal cell carcinoma, squamous cell carcinoma and inflamed scar tissues) spectra. The assessment by cross-validation achieved a sensitivity of 100% and a specificity of 91% (Lieber *et al.*, 2008b). Artificial neural networks were applied for diagnostic prediction of five different skin lesion types, including normal skin, pigmented nevi, seborrhoeic keratosis, basal cell carcinoma and malignant melanoma. In this study a total of 222 tissue samples were measured by Raman spectroscopy. The resulting spectra were used to build and test an ANN by LOOCV, which achieved accuracies for the varying pathology groups of 80.5-99.1% (Sigurdsson *et al*., 2004). For a Raman imaging approach 15 basal cell carcinoma were measured. The resulting Raman maps were subjected to K-mean clustering and the generated maps compared with the correlating histopathology slides. It showed that differentiation between normal and abnormal tissue is feasible (Nijssen *et al*., 2002).

In a small study consisting of six patient samples, FTIR microspectroscopy was investigated for its potential for differing between nevi from melanoma. It was shown that hierarchical clustering could successfully separate between the two groups (Tfayli *et al.*, 2005). FTIR-microspectroscopy and cluster analysis was also investigated for image analysis. Thus, infrared maps obtained from ten patient samples were subjected to K-mean clustering and hierarchical clustering in order to generate colour-coded images. The results demonstrated that the generated images could reproduce tissue histology (Ly *et al.*, 2010). In an earlier work the same group showed that infrared mapping in combination with K-mean clustering is capable of highlighting histological structures in skin tissue as well as colon samples (Ly *et al.*, 2008).

3.3.1.7 Lymph node metastases

Lymph node assessment is an important step in staging cancers especially since it is known that the presence of metastasis carries a worse prognosis for the patient. In order to allow a better assessment of the lymph node status in breast cancer patients Raman spectroscopy has been investigated for a potential inter-operative application. In this approach 38 lymph nodes have been measured with a Raman hand-held probe. The achieved spectra were used to develop a PC-fed LDA model, which achieved a sensitivity of 92% and a specificity of 100% when tested by LOOCV (Horsnell *et al.*, 2010). In a different study 103 lymph nodes, representing different pathologies including primary lymph nodes from Hodgkin's and non-Hodgkin's lymphomas as well as lymph nodes containing metastases from squamous cell carcinomas and adenocarcinomas. A developed LDA model, built for differing between these four groups achieved sensitivity of 75-100% and specificities of 86-99% when tested by LOOCV (Orr *et al.*, 2010).

Infrared spectroscopy has been specifically investigated for image analysis with the aim to visualise the presence of metastases in lymph nodes. Thus, 30 lymph node samples obtained from breast cancer patients were used to develop an imaging method combining hierarchical clustering and artificial neural networks. HCA was used to group spectra together into replicate the morphological structure of lymph nodes. The grouped spectra were further processed by the established ANN with the aim to detect metastasis within the image. It showed that this approach could accurately reproduce tissue pathology and highlight metastases (Bird *et al.*, 2008). HCA clustering was also applied for visualisation of micrometastases in infrared maps derived from lymph node samples. Based on ten patient samples it was shown that the visualisation of micrometastases is feasible (Bird *et al.*, 2009). In order to estimate the most suitable number of clusters for distinguishing different tissue types found in lymph node samples Wang, Garibaldi *et al.* (2007b) developed a method based on a fuzzy c-means clustering.

In a study comparing Raman and infrared spectroscopy for the assessment of lymph nodes of oesophageal cancer patients it was demonstrated that data derived from both methods is capable of predicting pathology status. For each spectroscopic method a PC-fed LDA model was developed and consequently for both models a training performance greater than 94% was estimated (Isabelle *et al.*, 2008).

3.3.2 Brain tumours

Excisional biopsy can be a potential hazard for vulnerable organs such as the brain. Taking this into account vibratonal spectroscopy would be an ideal tool for future *in vivo* application in brain tumour diagnostics. In addition, inter-surgery application for estimation of tumour during resection would be highly desirable since a too excessive resection might result in brain damage. Conversely an incomplete resection can cause the reoccurrence of the tumour.

Biopsies from three normal adrenal glands, 16 neuroblastomas, five ganglioneuromas, six nerve sheath tumours, and one pheochromocytoma were collected for a Raman study. Principal component analysis and discriminant function analysis were used to build a classification model, which separated the different pathologies with sensitivities of 95%-100% and specificities of 92.3-100% (Rabah *et al.*, 2007). The same group investigated in a similar manner if principal component analysis and a discriminant model built of Raman spectra obtained from frozen samples is capable of predicting the pathology of fresh samples. This classification approach yielded sensitivities of 80.8%-100% and specificities of 64.3%-100% (Wills *et al.*, 2009). In a small study of 20 patients a LDA model was applied to discriminate between meningioma from normal dural. The resulting LDA model achieved an overall accuracy of 100% when assessed by LOOCV (Koljenovic *et al.*, 2005).

Infrared spectroscopy has been extensively investigated for diagnosis of brain tumours. For instance Steiner, Shaw *et al.* (2003) applied a classifier system, consisting of a genetic algorithm for feature selection and a LDA model for discriminating cancerous brain tissue (astrocytoma, glioblastoma) from normal brain tissue (25 patients). The developed classifier separated the infrared spectra into four distinctive groups with accuracies of 83-96%. The same classification system was applied for a larger data set consisting of infrared data obtained from 59 tissue specimens representing four pathologies. The resulting model

achieved accuracies of 17-95% for the different pathology groups respectively, when tested by a four-fold cross-validation (Beleites *et al.*, 2005). In further work it was demonstrated how the performance of this classification approach can be enhance by integrating several independent LDA models into an ensemble. It was reported that the overall performance increased from 67% to 82% due to the ensemble approach (Beleites *et al.*, 2008). In a study regarding the application of IR spectroscopy as an as an intra-operative tool in cerebral glioma surgery 54 tissue samples from six patients were used. An LDA model was employed to generate colour-coded maps representing six different pathology groups. The comparison of the results with histopathology slide showed that in 98% of all cases the correct decision, whether continuing with surgical tumour resection or not, could have been made based on the LDA colour maps (Sobottka *et al.*, 2009).

3.3.3 Leukaemia

Leukaemia is a cancer originating in blood cells and bone marrow. Currently, routine diagnostics includes the presentation of the clinical manifestation and morphology, which are commonly applied in combination with molecular methods, cytogenic studies and flow cytometric immunopheonotyping (Kendall *et al.*, 2009).

Although Raman spectroscopy has been widely investigated for epithelial cancer diagnostics only a small number of studies were dedicated to Leukaemia diagnostics. It was reported that Raman micro-sperctroscopy was investigated for discriminating between normal and transformed lymphocytes. Based on a principal component clusters a sensitivity of 98.3% and a specificity of 97.2% was estimated (Chan *et al.*, 2006).

Infrared spectroscopy has also been broadly investigated for leukaemia research. However, classification methods for diagnostic approaches were only reported in a small number of studies. For instance, Babrah, McCarthy et al. (2009) applied FTIR spectroscopy and LDA for discriminating between T-cell lymphoma, B-cell lymphoid and myeloid leukaemia cells. The generated model yielded sensitivities of 79.9-100% and specificities of 93.8-100% when tested by LOOCV. In a different approach, infrared spectra obtained from plasma sample representing healthy patients and patients suffering from chronic lymphocytic leukaemia were investigated by HCA. It showed that cluster analysis can be used to distinctively differ between healthy and leukaemic samples (Erahimovitch *et al.*, 2006).

3.3.4 Summary

As the literature review showed the most commonly employed method for developing diagnostic models is LDA. More complex classifiers such as ANN and SVMs are less frequently applied. It was observed that the model testing is most frequently done by LOOCV, although it is known that assessment with an unseen data set is more accurate. The predominant application of LOOCV might be explained by the fact that data sample availability is restricted. However, in many cases a model achieving a good LOOCV performance yields a significantly lower result when tested with an independent test set. Therefore, in order to make this classifiers applicable as routine diagnostic tools much more thorough assessment of models is needed. Furthermore if diagnostic models are applied in routine clinical analysis they might be confronted with imperfect data, which can be caused by system to system variation, working condition and by the operator. So far no approach has been taken to investigate the robustness of these classification model towards the mentioned error sources.

For the image analysis the most frequently applied data analysis technique was found to be cluster analysis. Although, clustering methods of images demonstrated to be useful for highlighting tissue features in images they do not allow an automated diagnosis based on images. Thus, more sophisticated methods such as ANN or SVM must be investigated for their capability for automated image analysis and diagnosis.

This work aimed to address the earlier mentioned limitations in order to advance the development of classification models for future diagnostic applications of vibrational spectroscopy. The approaches taken for achieving this are presented in the following chapters.

4 Machine learning and Raman spectroscopy for lymph node diagnostics

4.1 Introduction

Breast cancer is the most common cancer in women worldwide and, due to the increasing number of newly diagnosed cases, is a growing healthcare problem (Cancer Research UK, 2010b).

Most frequently breast cancer originates in the glandular elements of the breast, the lobules and the ducts. Malignant transformation includes such changes as nuclear enlargement, changes in the number of chromosomes and variations in shape and size (Kumar *et al.*, 2005). These changes affect the chemical composition but do not cause a large-scale production of new chemicals. One of the most significant changes in malignant disease development is the change of the nuclear-to-cytoplasm ratio. This causes malignant tissue to differ from benign tissue in terms of the concentration of the main building blocks, nucleic acid, proteins, lipids and carbohydrates (Shafer-Peltier *et al.*, 2002).

Progressing breast carcinoma metastasizes to the regional lymph nodes over the efferent lymphatic vessels and enters the subcapsular sinus. For this reason early lymph node involvement is often found in the subcapsular sinus. An invaded lymph node may respond by displaying secondary follicles with reactive germinal centres, sinus histiocytosis and granulation. A further very specific change is desmoplasia, the change in the formation of collagenous fibrous stroma around the metastatic cells. With growing involvement genuine lymph node architecture gets increasingly replaced by metastases, in the majority of cases reflecting the features of the primary tumour (Ioachim *et al.*, 2009).

Lymph node involvement is an important prognostic factor for breast cancer patients. Thus, breast cancer staging includes the assessment of the lymph nodes in the ipsilateral axilla. For this reason lymph node biopsy is carried out in order to determine the presence of metastasis (Arnaud *et al.*, 2004). A frequently applied method is sentinel lymph node biopsy, where the first node or nodes with direct lymphatic drainage from the tumour are identified. These lymph nodes are considered to be the first ones to be involved when a tumour metastasizes (Morton *et al.*, 1992). Lymph node involvement has a major impact on further treatment of the patient, including extensive dissection of axillary lymph nodes, chemotherapy and occasionally radiotherapy.

Current routine histopathology methods for lymph node assessment encounter several limitations. Traditional histological staining techniques are subjective, resulting in missed lesions and significant disagreement of inter- and intra observers (Cserni *et al.*, 2005). Frequently histopathology laboratories do not have the human resources to analyse every section of removed lymph nodes and due to that micrometastases might be missed. It was shown that an extended histological assessment of lymph nodes reduces the number of false negative samples. Accordingly, 7% to 30% of negative nodes were reclassified as positive nodes as a result of an more exhaustive assessment (Chatterjee *et al.*, 2002).

Ideally, lymph nodes are assessed intra-operatively, which facilitates a lymph node clearance within the same surgery as the tumour resection. Thus, patients undergo only one surgical procedure and benefit from reduced stress levels and no delay in adjuvant treatment. Alternatively, methods have been developed to allow faster intra-operative assessment of lymph nodes. These methods, including touch imprint cytology (Salem *et al.*, 2003, Salem *et al.*, 2006) and frozen section analysis (Grabau *et al.*, 2005). Touch imprint cytology includes

bisecting of lymph nodes and pressing the surfaces onto a slide, which are typically reviewed by a histopathologist. A meta-analysis, reviewing 31 studies, reported an average sensitivity of 63% and an average specificity of 99% for imprint cytology (Tew *et al.*, 2005). In comparison, sensitivities of 57-87% and specificities greater than 99% were reported for frozen section analysis (Creager *et al.*, 2002). Nonetheless, frozen section analysis features a high processing time and also needs the immediate assessment by an experienced histopathologist.

According to the previously presented facts, new methods are needed which allow a more sensitive and objective lymph nodes assessment. Furthermore, theses methods should be applicable during surgery. These requirements can be met by spectroscopic methods such as Raman spectroscopy. For instance Smith (2005) developed a PC-fed LDA model for lymph node classification based on Raman spectroscopic data. This model achieved a sensitivity of 88% and a specificity of 80% when tested by LOOCV. In more recent work, a PC-fed LDA model was employed for intra-operative lymph node diagnostics, which achieved a sensitivity of 92% and a specificity of 100% when tested by LOOCV (Horsnell *et al.*, 2010). Beside these studies, no further evidence of the application of Raman spectroscopy for lymph node assessments in breast cancer patients was found. Both works report high sensitivities and specificities, nonetheless they were only tested by LOOCV, which is not considered to be a very thorough assessment method and subsequently such models are likely to fail when tested with unseen data. Thus, diagnostic models are required that result in even greater accuracies as the reported ones as well as supersede the accuracies of other diagnostic techniques such as imprint cytology or frozen section analysis. Furthermore, the accuracy of these diagnostic models must be maintained throughout rigid testing procedures in order to demonstrate reliability for future diagnostic applications.

Concluding, a further step towards the clinical application of Raman spectroscopy in lymph node assessment of breast cancer patients requires the development of diagnostic models, which allow reliable classification of tissue samples, without the need for human interpretation. The development and assessment of such classification methods is reported in this chapter.

4.2 Materials and Methods

4.2.1 Samples

A total of 43 axillary lymph nodes were collected after surgical resection of breast cancer patients. All samples were obtained with the full consent of patients and approved by the Gloucestershire Research Ethics Committee. Each lymph node was cut into halves. One half was placed onto acetate paper and snap frozen in liquid nitrogen in order to maintain the freshness of the tissue. From the frozen sample a 7μm section was cut and placed on a calcium fluoride slide and stored in a -80°C freezer for Raman spectroscopy. The other half of the node was sent for the routine histopathology, which found that out of the 43 samples 13 were positive and 30 were negative for metastases.

4.2.2 Raman microspectroscopy

A Renishaw System 1000® Raman microspectometer coupled to a diode laser, a Leica® microscope, a Prior® electronic stage, a video viewer and a desktop computer with customized Grams® software was used for all measurements. The diode laser had an output of 350 mW and was set to a wavelength of 830 nm with the aim to reduce autofluorescence from tissue. Raman mapping was executed in steps of 100 μm in *x* and *y* directions across the

sample surface. At each point the spectra were integrated for a total of 30 seconds. More details on sample preparation and the carried out Raman mapping can be found in Smith (2005).

4.2.3 Data analysis workflow

The first step was image processing and thus all generated Raman maps were loaded into Matlab (Mathworks, USA) and converted into 3D hyperspectral matrices. For each individual map principal component analysis (PCA) was executed. The resulting first three principal components were used to transform the Raman map into a composite image. According to the composite images heterogeneous regions were selected manually by avoiding obvious fat or areas likely to be contaminated with blood. Thus, spectra were collected which represent homogenous regions of positive or negative nodal parenchyma. An example of this process is demonstrated in Figure 4.1.

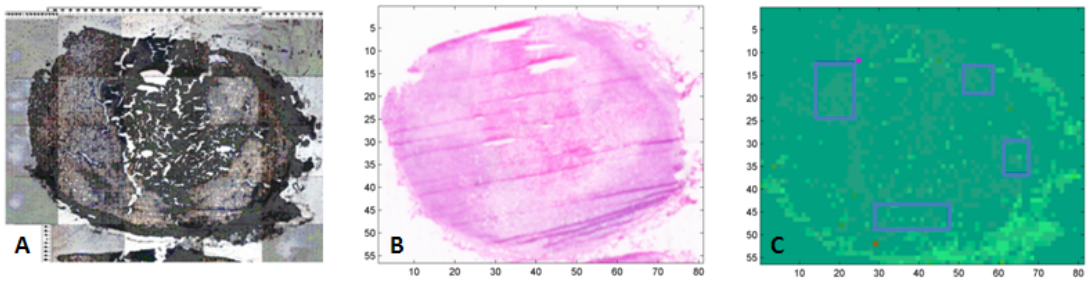


Figure 4.1 A. White light image of a lymph node effaced with metastatic tumour. B. The H&E staining shows pale stained areas, which are a result of the marked desmoplastic reaction to the metastatic tumour. C. Composite image of the lymph node sample. Each pixel represents one spectrum. Thus, the grid units represent spectra for both axes (81 spectra in x direction and 56 in y direction). Areas boxed in blue represent the selected spectra, which were then used for further investigations.

Saturated spectra were removed from the extracted data and spectra containing evidence of cosmic rays were corrected. The resulting data set consisted of 10,477 spectra, where 3,385

were from positive samples and 7,092 were from negative samples. The number of spectra available for each sample varied from four to 1,014 spectra per sample. The high variation of available spectra for each sample was caused by the fact that some nodes contained a large amount of fat, resulting in an increased number of saturated spectra, which had to be removed before further analysis. The data at this stage was the starting point for the model development.

As described in section 3.1, for classification model development it is good practice to separate all samples into two data sets before the model development. The larger data set forms the training set and the smaller one the test set. How to split the data depends on the amount of available samples for each pathology class. For this approach the data set was split randomly into test set and training set, at which the test set represents approximately a quarter of each pathology group. Accordingly the test set was independent and contained 12 samples, eight negative and four positive, while the training set contained 31 samples, nine positive and 22 negative. A summary of the generated training and test set is shown in Table 4.1.

Table 4.1 Summary of the obtained training and test set, which were used for the diagnostic model development.

Data set	Total number of samples	Positive samples	Negative samples
Training set	31	9	22
Test set	12	4	8

The resulting training set was used to optimise the parameters of each classification model. The resulting parameters were then used to build the final model. In order to estimate the predictive power each model was tested with an independent test set. An overview of the described workflow is illustrated in Figure 4.2.

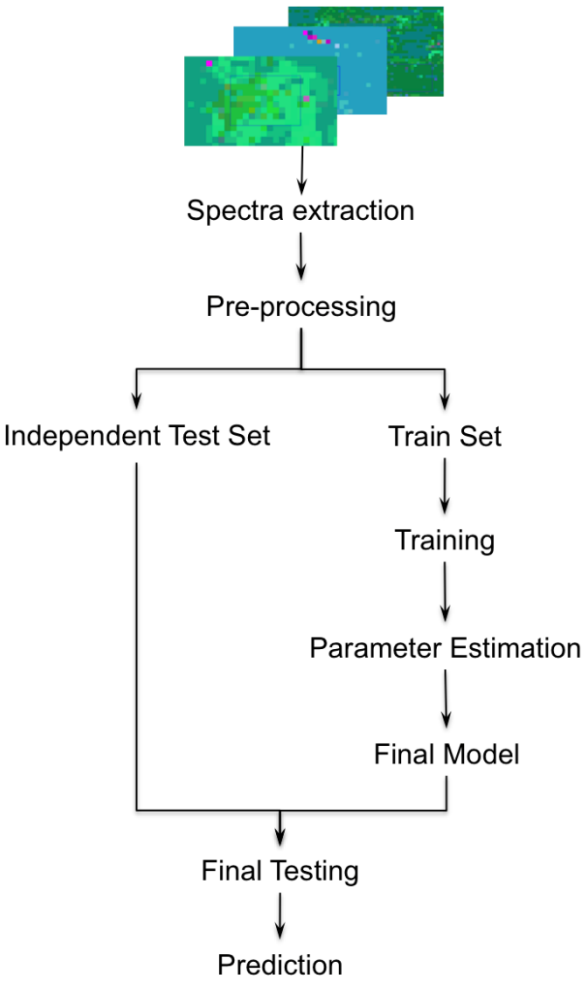


Figure 4.2 Classification workflow for Raman lymph node maps.

4.2.4 Data subsets

The number of spectra consisting for each sample varied between four and 1014. Due to the fact that some samples are overrepresented where others are underrepresented. This can reflect on the classification result, for example by over fitting a classifier to overrepresented samples. Further, a smaller data set reduces the computing time of the training procedure. For the described reason balanced data sets were generated.

For the first approach, the data set was balanced by maintaining 70 spectra for each sample. It was decided to select 70 spectra for each sample since it was found that this is the highest number of spectra the majority of all samples had in common. For this purpose spectra were selected systematically in order to represent the whole sample. This was executed by dividing the number of spectra by 70. The result was rounded down to the next integer x. Estimated x was then used to extract every x^{th} spectra. Consequently the training set consisted of 2520 spectra and the test set of 374 spectra.

In a different approach the original data set was reduced by a specifically developed spectra selection method, which aims to decrease the variance within a sample. For each sample 50 spectra, which was identified as the highest number of spectra the majority of all samples had in common, were selected randomly from the pre-processed data set. In this manner the resulting training set consisted of 1550 spectra and the testing set of 355 spectra.

4.2.5 Targeted spectra selection

4.2.5.1 Spectra variance

The number of available spectra for each sample varied from four to 1014 spectra per sample. This was caused by the fact that some samples contained more fat than others, which caused an increased number of saturated spectra and consequently these spectra had to be removed during the data cleaning step. The investigation of the remaining spectra showed that there was a high variance within each individual sample. The observed variance is illustrated in Figure 4.3.

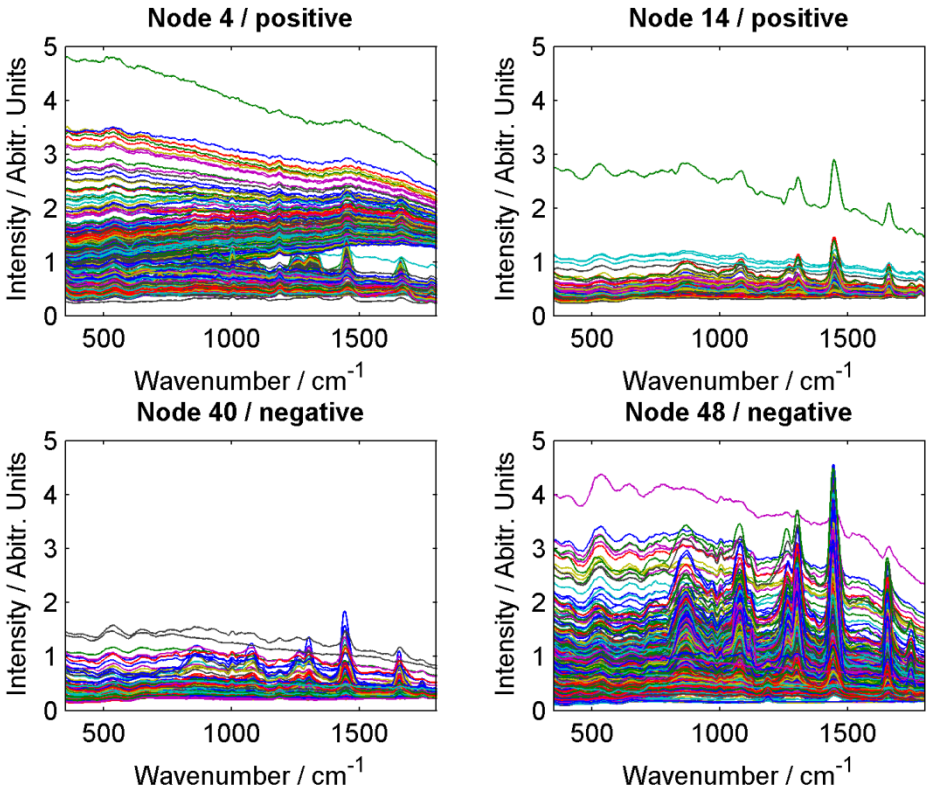


Figure 4.3 Spectra variation for four different lymph node samples. As these graphs show the spectra within one sample can very significantly and even contain spectra not representing biological features as shown in the graph in the top left corner.

It was assumed that a reduction of variance within a sample, as well as the removal of spectra not representing biological signals leads to an increased classification performance. For this purpose, a method was developed that allows targeted spectra selection for each sample. The developed method is presented in the following section in greater detail.

4.2.5.2 Method

First of all the mean of intensity per arbitrary units was calculated for every individual spectrum. According to the obtained mean values a histogram was generated separately for each lymph node sample. The highest bar, representing the most common number of mean value of spectra (median), was estimated. The mean value represented by this bar was

determined and used to build the lower cut off limit. Spectra under this value were neglected because they are assumed to be low signal spectra. Additionally, the standard deviation of the mean values was calculated. The lower cut off limit plus the standard deviation were used to establish the upper cut off limit. Spectra above this limit, which are considered to be of higher intensity than the average of the residual spectra, were rejected. Finally, spectra falling between these two limits were extracted. In the histograms representation, as illustrated in Figure 4.4, selected spectra are illustrated in red whereas rejected spectra are illustrated in blue.

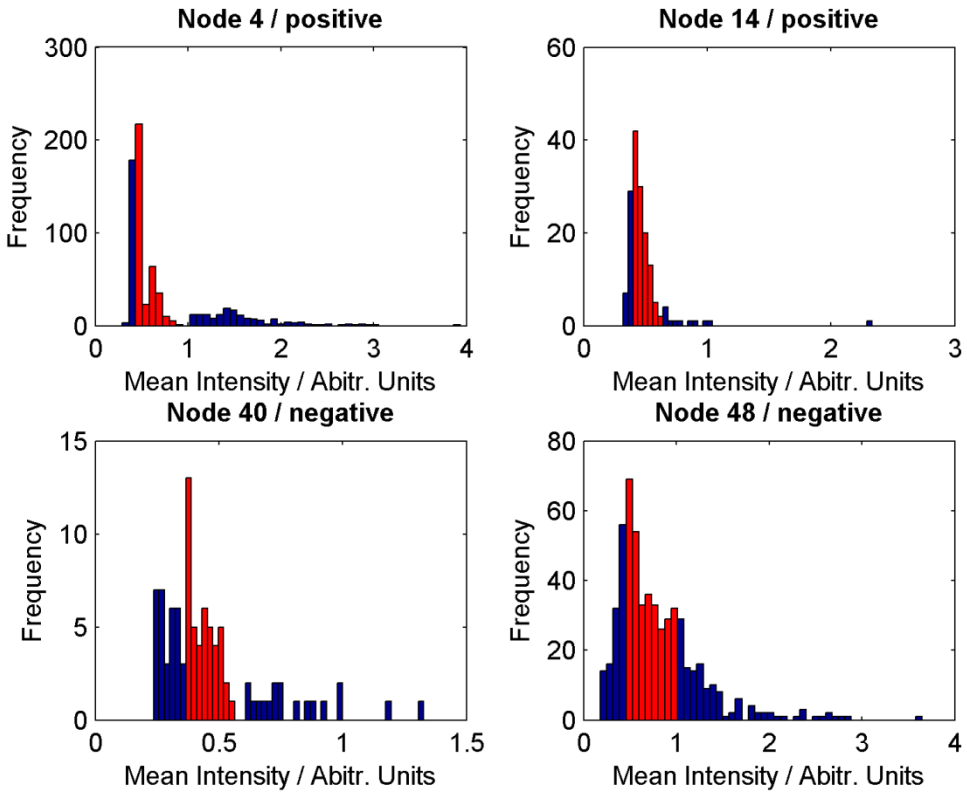


Figure 4.4 Histogram illustrating the mean intensity values of spectra for individual nodes. Spectra with a mean intensity value between the cut off limits are shown in red where the excluded spectra are shown in blue.

For all lymph node samples spectra were extracted according to this method a result the variance within each sample was reduced. In addition to that spectra suspected not to

represent significant biological signals, for example inside ducts, were removed. The normalising effect of this method is illustrated in Figure 4.5.

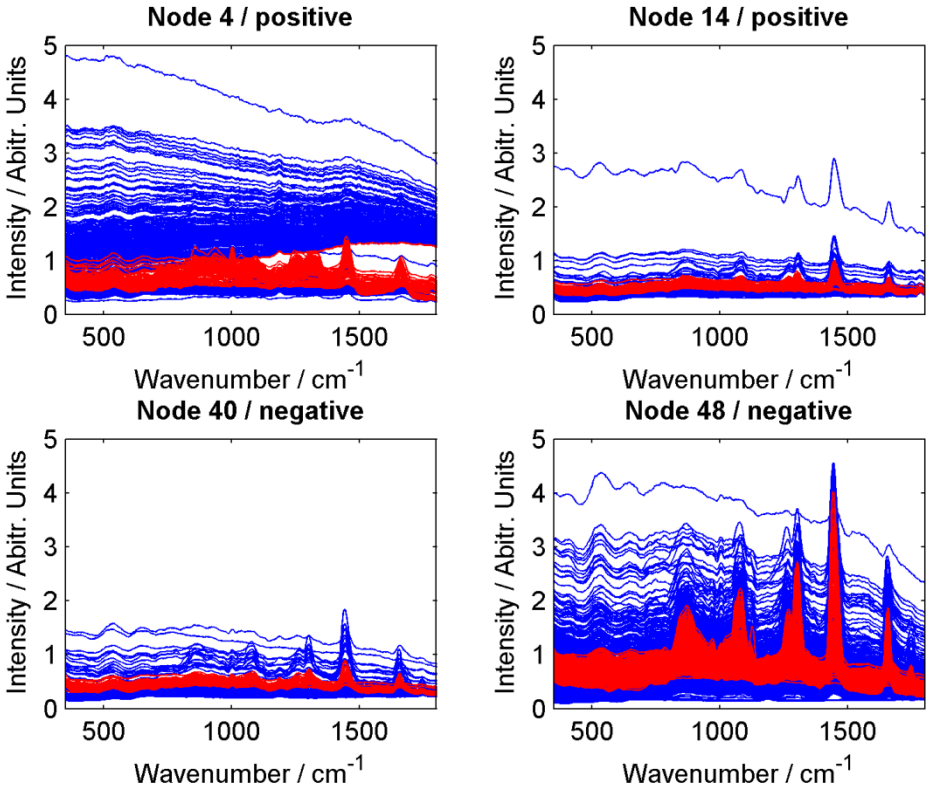


Figure 4.5 Illustration of the selected spectra as suggested by the selection method: the selected spectra are shown in red where the discarded spectra are shown in blue.

4.2.6 Classification methods

4.2.6.1 Linear discriminant analysis

Linear discriminant analysis (LDA) has been investigated for classification of the Raman lymph node data set by Smith (2005). Data pre-processing included mean centring and normalisation. After this step principal component analysis (PCA) was executed and from the resulting first 25 principal components (PCs) only the most significant PC scores were retained. The significance testing for this step was carried out by ANOVA. The remaining

scores were used to build the LDA model, which was finally validated by leave one out cross-validation (LOOCV).

In previous work, as described earlier, no approach was taken for optimising the number of PCs. Therefore, the first aim was to examine if the number of components can be optimised and consequently leads to an improved classification result. This was done by leave one out cross-validation with constantly increasing number of PCs, starting from one up to 25 PCs. The number of components was determined by the overall prediction accuracy resulting from the LOOCV. The final model was build with the optimum number of PCs and finally tested with the independent test set. This procedure was done for both data subsets (data set A and data set B).

4.2.6.2 Partial least square discriminant analysis

Matlab and the PLS Toolbox (Eigenvector Research Inc.) were used for developing partial least squares discriminant analysis (PLS-DA) models.

For this approach, the data were normalised and different scaling methods were investigated, including mean centring and auto scaling. This allowed the assessment of the influence of different scaling methods on the PLS-DA model performance. For the optimisation of the number of latent variables (LVs) LOOCV was executed, starting with one LV up to 25. According to the prediction result of the LOOCV the optimum number of LVs was estimated and used to build the final model. The resulting model was then assessed with the independent test data.

4.2.6.3 Support vector machines

All SVM models were built by using the toolbox libsvm developed by Chang and Lin (2001). This freely available toolbox allows SVM classification and can be used in different environments including Matlab, R, Perl and Python.

For this study three different types of SVM kernels were investigated: linear, polynomial and radial basis function (RBF) kernel. It is required to optimise kernel specific parameters for each SVM, which is of significant importance since the kernel parameters strongly influence the classification performance. For this purpose, a loose parameter optimisation was executed, which was done by cross validating a subset of the training data. Using only a subset of the training data helped reduce the computing time. A loose search resulted in a temporary approximation of the kernel parameters. These parameters were fine tuned by a second more rigid parameter optimisation. This time the whole training set was used. For this final optimisation step, the range of the kernel parameters was set close around the previously approximated parameters and cross-validation was executed (Hsu *et al.*, 2008). It is assumed that the best parameters resulted in the best cross-validation result. The estimated kernel parameters were then used for creating the final model. In order to evaluate the predictive power of the built model it was tested with the independent test set. The described SVM classification workflow is illustrated in Figure 4.6.

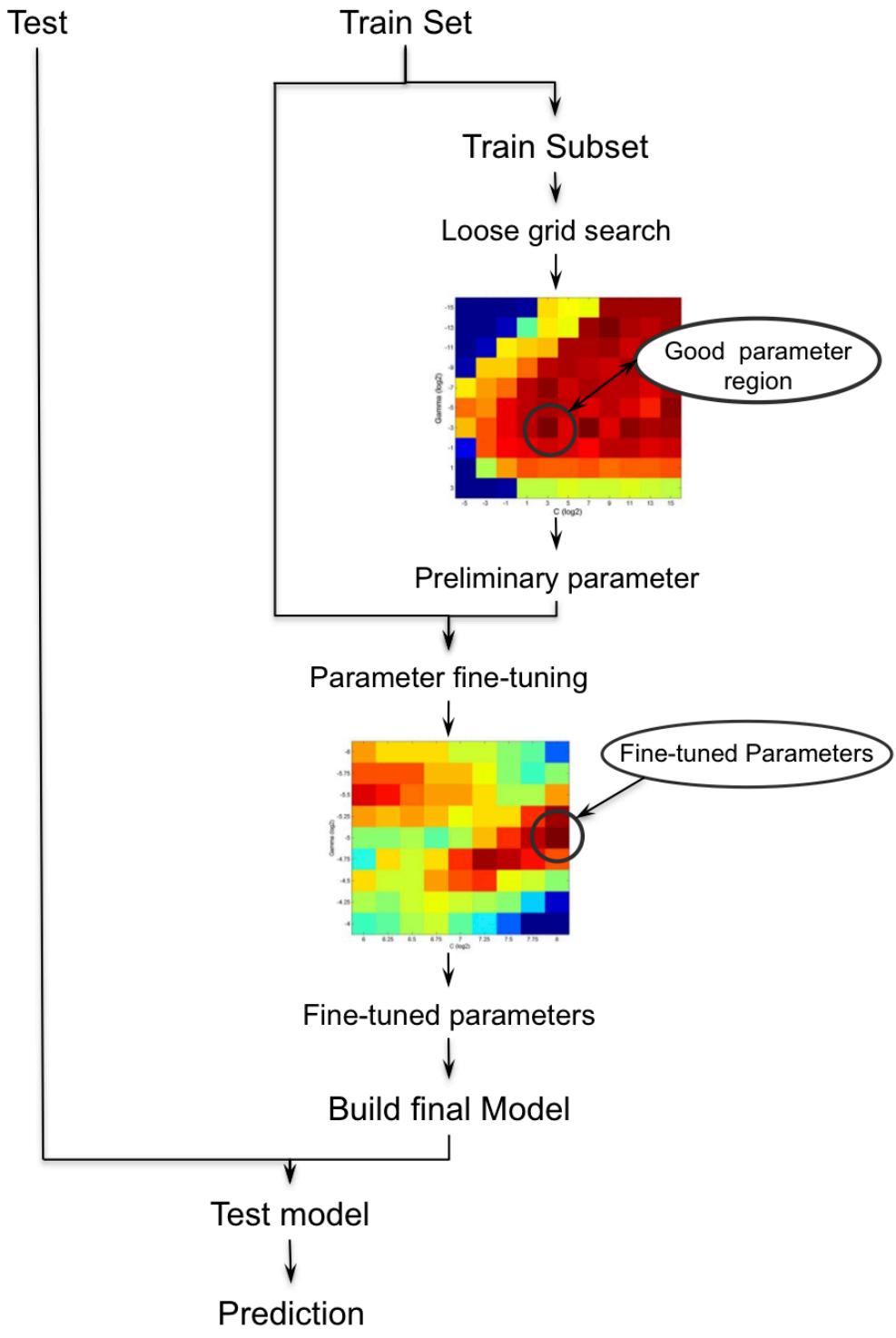


Figure 4.6 SVM classification workflow

4.2.7 Assessment of model significance

Monte Carlo methods were employed for the empirical assessment of the model significance.

All samples of the training set, which was used to build the original model, were assigned

randomly with a class membership, either positive or negative. These samples were then used to build a model using the same parameters as the originally estimated ones. The newly generated model was then tested with the test set, which also has been randomly assigned with sample classes. For both data sets the relative number of positive and negative samples was maintained. This procedure was executed multiple times, where every time the class membership of the data was randomly permuted. The achieved testing accuracies were then used to create a null distribution. The comparison of the null distribution with the observed testing performance allows an empirical assessment of the model significance. This approach is inspired by similar work in which Wongravee *et al.* (2009) utilised Monte Carlo methods for estimating the significance of variables.

4.2.8 Investigation of key features

For the model development the spectral region from 350-1850 cm^{-1} was used. In order to investigate spectral features that have the greatest impact on the model performance alternating intervals of 100 wavenumbers were eliminated from the data set systematically. The remaining data set, containing a total of 1401 wavenumbers, was used to build and test an optimised classifier. In this manner, for a spectral range starting at 350 to 1850 cm^{-1}, 15 models were built and tested. A decrease in testing performance is assumed to be caused by the fact that the left out spectral features have a high impact on the model performance

4.3 Results and Discussion

As described in section 4.2.4 two different data sets were created. In the following, to the classification approach using balanced data set consisting of 70 spectra per lymph node it is

referred to as data set A. Hence, data set B refers to the classification approach accomplished for the data set, which was generated by applying the targeted spectra selection method.

4.3.1 Data set A

4.3.1.1 Linear discriminant analysis

First of all the number of PCs was optimised, which was found to be 18. The training achieved an overall accuracy of 93.1%, a sensitivity of 100% and a specificity of 91.1%. This good result could not be achieved when testing the model with the independent data set. Although the testing set could be classified with sensitivity of 100%, only a specificity of 68.6% was obtained. The results of this LDA approach are summarised in Table 4.2.

Table 4.2 LDA results for data set A

Training			**Testing**		
Sens. / %	**Spec. / %**	**Acc. / %**	**Sens. / %**	**Spec. / %**	**Acc. / %**
100	91.1	93.1	100	68.6	80.2

This model managed to classify all positive samples correctly but it misclassified a high number of negative spectra as positive. Assuming that this model would be used for deciding if lymph node dissection is required or not several patients would undergo unnecessary surgery. However, all cancerous samples would be identified correctly.

These data were used in previous work to develop a LDA model, which was only tested by LOOCV. This model was built by using the 25 PCs without prior optimisation of this number. The generated model yielded a sensitivity of 80% and a specificity of 88%

respectively. In contrast, the new model development included the optimisation of the number of PCs fed into the LDA model. The optimisation procedure resulted in an increased cross-validation sensitivity of 84.9% and specificity of 88.9%. Interestingly, the required number of PCs is still very high, especially since 99.8% of variance are captured by the first three PCs. For this reason it must be assumed that subtle spectral differences are of high importance for the LDA model in order to allow a good separation between the classes.

Finally, this approach demonstrates that a model performing well when assessed by cross-validation does not perform equally when tested with an independent test set. Therefore, rigid test methods, such as the use of an independent test set, are needed in order to allow a more reliable assessment of the predictive performance.

4.3.1.2 Partial least squares discriminant analysis

Data set A was normalised and scaled using alternated different scaling methods, including mean-centering and auto-scaling. The differentially pre-processed data sets were then used to generate partial least squares discriminant analysis (PLS-DA) models.

The optimum number of latent variables (LVs) was estimated by LOOCV of the training set. For the approach without applying any scaling method, it was found that ten LVs are the optimum number of components to build the final model. For the two approaches applying scaling methods the optimised number of LVs was nine. Figure 4.7 illustrates the model optimisation and the determination of the optimum number of PLS components.

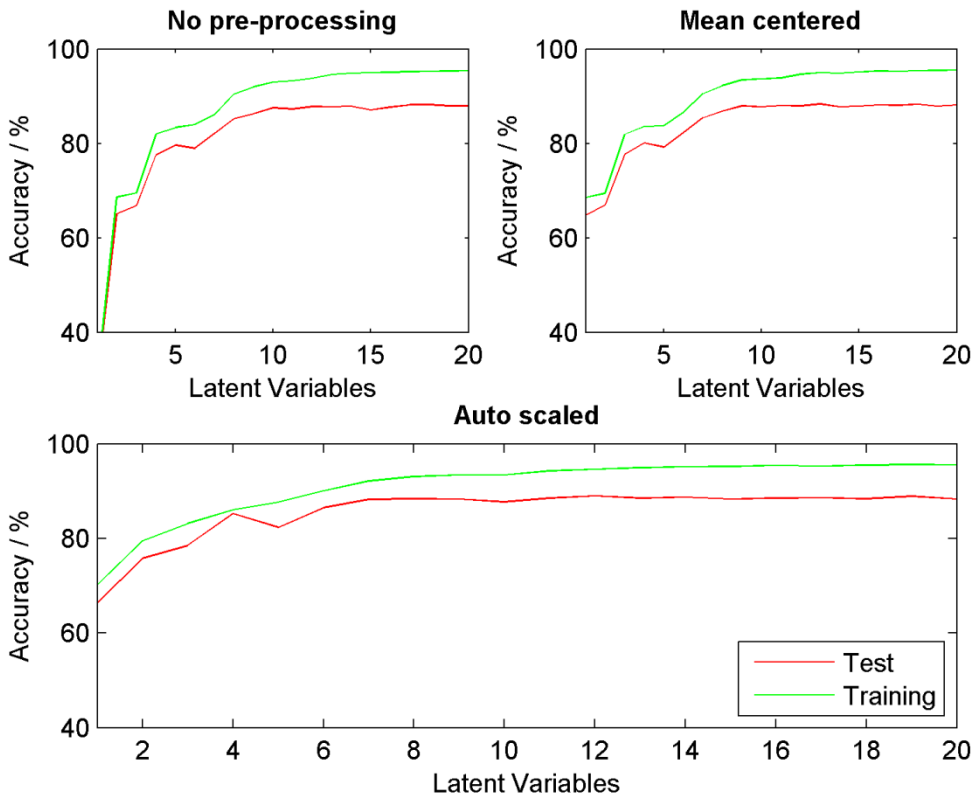


Figure 4.7 Optimisation of PLS-DA models for data set A. The green graph shows the average training performance and the red graph the average testing performance obtained by cross-validation.

The estimated numbers of LVs were used to build the final models. All three models were then tested with the independent test set. The respective classification accuracy, sensitivity and specificity for the models built on data set two are summarised in Table 4.3.

Table 4.3 PLS-DA results for data set A.

	Training			**Test**		
Scaling	**Sens. %**	**Spec. %**	**Acc. %**	**Sens. %**	**Spec. %**	**Acc. %**
No scaling	92.4	94.8	94.2	87.8	71.9	78.9
Mean-centred	92.0	95.3	94.5	86.6	74.8	80.0
Auto-scaled	92.4	95.0	94.4	85.4	74.3	79.1

It showed that the training results of all three models are almost similar since they do not vary more than 0.3%. A similar result was obtained for the testing result, where the difference between the highest and the lowest classification result was not more than 1.1%. This difference is considered to be an improvement brought by the scaling method. Although, the scaling brought a slight improvement it cannot be considered as significant.

Overall the PLS-DA model did not perform better than the LDA model since both diagnostic models yielded test accuracies around 80%. Nonetheless, the PLS-DA model achieved a significant higher specificity. Thus, it must be assumed that the PLS-DA model separates more evenly, whereas the LDA model separates in favour of positive samples, which explains the 100% sensitivity in comparison to the low specificity of 68.6%.

4.3.1.3 Support vector machines

Linear kernel

Similar to the other model developments data set A was normalised and mean-centred. For a linear kernel SVM it is required to optimise parameter *C* (error cost) (Phan *et al.*, 2005) and due to that the first step was to approximate parameter *C* in a loose search where only a subset (five positive and five negative samples) of the training data was used. Using only a data subset helped to reduce computing time and was found to be a sufficient approximation. For this search the range of parameter $C = [2^{-5}, 2^{-3},\ldots, 2^{13}]$. The approximated parameter *C* was finally optimised by a second more thorough search. The fine-tuned parameter *C* was found to be 2^9, as shown in Figure 4.8. This Figure also illustrates that a too high value of *C* results in over fitting, which can be seen by the fact that *C* values higher than 2^9 still result in an improvement of the training performance where on the other hand the testing performance

starts to decrease. The reason for this is that the higher C the more thorough the decision boundary of the support vector is drawn, which consequently results in over-fitting.

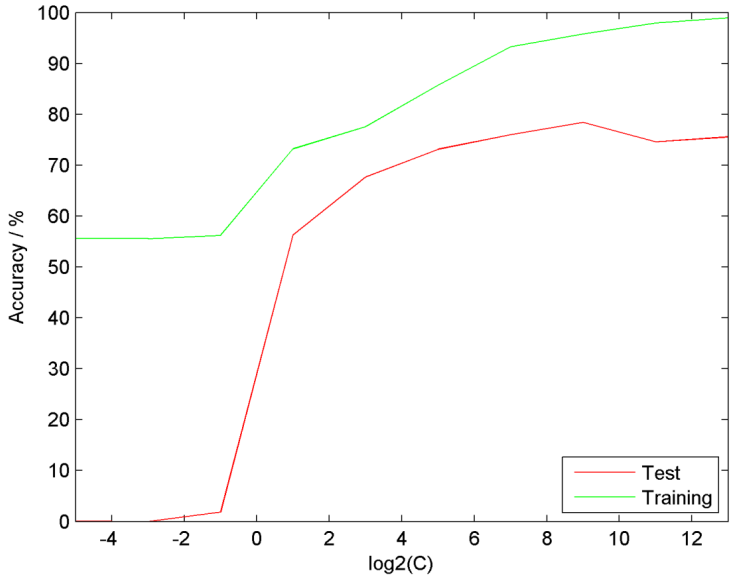


Figure 4.8 Loose parameter estimation for linear SVM (data set A).

A second optimisation step was used to fine tune the previously estimated parameter C. For this purpose, another search was implemented using a range of $C = [2^8, 2^{8.25},\ldots,2^{10}]$. The calculated cross-validation results of the fine-tuning step did not vary as significantly as in the loose parameter estimation. The difference between the highest and the lowest result did not exceed 0.4%. In this manner it was found that the optimum parameter C for creating a linear SVM model is 2^{10}. This parameter was used to build the model and test it with the independent data set, which resulted, as summarised in Table 4.3, in a testing accuracy of 80.0%, a sensitivity of 78.7% and a specificity of 81.0% respectively.

Table 4.4 Linear SVM results for data set A.

Param.	Training			Test		
log2(*C*)	Sens. %	Spec. %	Acc. %	Sens. %	Spec. %	Acc. %
10	89.4	98.83	96.5	78.7	81.0	80.0

Comparing the over all performance of this model with the LDA and the PLS-DA model it shows that the overall performance is similar. However, they differ significantly in sensitivity and specificity and thus the linear SVM yielded the highest specificities among these approaches.

Polynomial kernel

The application of a polynomial kernel requires the estimation of the optimum number of polynomial degrees and the optimisation of the error cost (*C*). An increase of *C* reduces the misclassification of the data points in the training set (Vapnik, 1995). As already shown in the previous section a too high value of *C* can result in over-fitting. Similar as for the linear SVM models, as previously all data were normalised and mean-centred before carrying out a loose parameter optimisation was. For this approximation the range for the polynomial degrees was set as d = [2,…,8] and for C = [2^{-5}, 2^{-3},…, 2^{15}]. For the loose parameter estimation only a subset of the training data was used, which consisted of five positive and five negative nodes. The results for the loose parameter search are shown in the Figure 4.9.

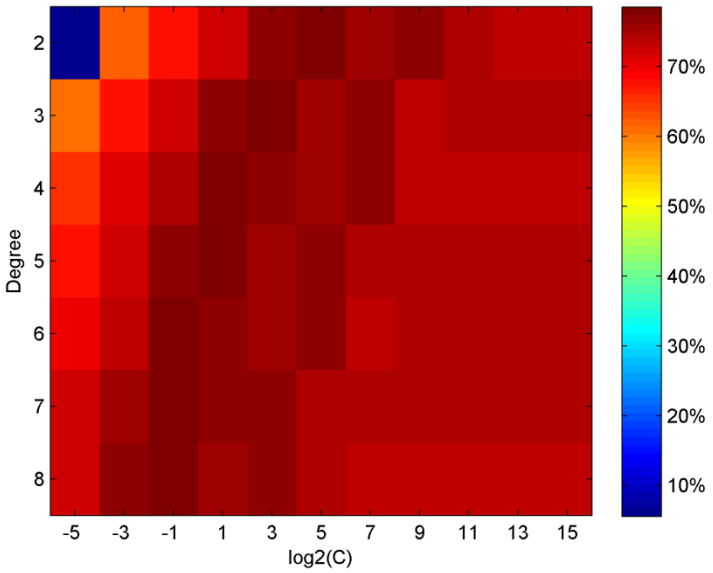


Figure 4.9 Result for loose parameter search for polynomial kernel (data set A).

According to Figure 4.9 the cross-validation testing results do not differ significantly between the different settings. For this reason it must be assumed that a polynomial kernel can not sufficiently model this data set. However, the selected parameters and the related classification results are summarised in the Table 4.5.

Table 4.5 Estimated parameters and grid search result for data set A.

Polynomial degree	**log2(*C*)**	**Accuracy %**
2	5	78.6
7	-1	78.6
8	-1	78.6

In the following the estimated parameters were fine-tuned and for this reason, the range of parameter *C* was extended around the values identified in the previous step. In contrast this time the whole training set was used for the LOOCV. The best three cross-validation results

and the related parameters are illustrated in Table 4.6. The classification accuracy is identical for all three approaches.

Table 4.6 Fine-tuned parameters for polynomial SVM (data set A).

Polynomial degree	log2(*C*)	Accuracy %
2	4	75.0
7	-1.75	75.0
8	-1.75	75.0

The settings as illustrated in Table 4.6 were used to build the final models, which were assessed with an independent test set. The obtained results are summarised in the Table 4.7. As the results show all model fails to classify positive spectra in the training and consequently in the testing. This can be seen in the fact that the sensitivity is 0% for all three models and the specificity is 100%. Therefore, a polynomial kernel is unsuitable for projecting this data into the higher dimensional feature space due to the fact that it does not result in an improved separability.

Table 4.7 Polynomial SVM results for data set A.

Parameters		Training			Test		
Degree	log2(C)	Sens. %	Spec. %	Acc. %	Sens. %	Spec. %	Acc. %
2	4.00	0	100	75	0	100	56.2
7	-1.75	0	100	75	0	100	56.2
8	-1.75	0	100	75	0	100	56.2

Radial basis function kernel

The application of a Gaussian radial bias function (RBF) kernel requires the optimisation of parameters C and γ. Parameter C regulates the error cost and γ controls the weight of the Gaussian kernel (Vapnik, 1995). First of all, data set A was normalised and mean-centred and subsequently a loose grid search using only a subset of the training data was carried out. This subset consisted of five positive and five negative samples. For the loose parameter search parameter C was set $C = [2^{-5}, 2^{-3},\ldots, 2^{15}]$ and $\gamma = [2^{-15}, 2^{-13},\ldots, 2^{3}]$ (Hsu *et al.*, 2008). In order to estimate the optimum parameters a colour map was generated as in Figure 4.10. The colour map shows that the cross-validation performance improves with increasing values of C and γ. According to the cross-validation results, shown in Figure 4.10, the best three parameter combinations were extracted and used for a further more precise search. These parameters and the respective cross-validation results are summarised in Table 4.8.

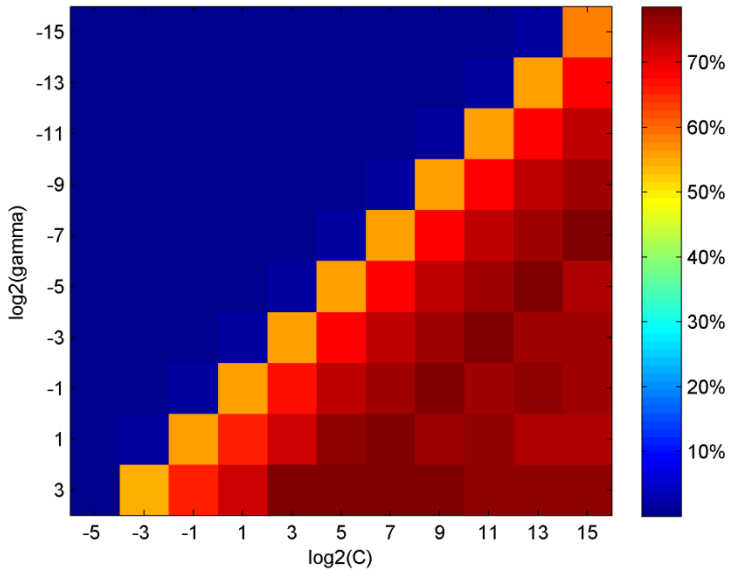


Figure 4.10 Loose grid search for RBF SVM (data set A).

Table 4.8 Estimated parameters for RBF SVM (data set A).

log2(C)	log2(γ)	Accuracy %
11	-3	78.6
15	-7	78.4
13	-5	78.4

The previously estimated parameters were fine tuned by a second grid search around their values. In contrast, for this optimisation step the complete training set was used. The range for each parameter values and the best classification accuracy are shown in Table 4.9.

Table 4.9 Fine-tuned parameters for RBF SVM (data set A).

Range log2(*C*)	Range log2(*γ*)	Best log2(*C*)	Best log2(*γ*)	Acc. %
[10, 10.25, …, 12]	[-4, -3.75, …, -2]	12	-2.75	70.5
[14, 14.25, …, 16]	[-8, -7.75, …, -6]	15.5	-6.25	72.5
[12, 12.25, …, 14]	[-6, 5.75, …, -4]	14	-4.75	72.5

The fine tuned parameters were then used to build the final model. All three settings performed almost identically in the training. Thus, all RBF SVM models obtained a training performance of over 95.0%. A similarly high result could not be obtained when testing the models. In this manner, the best overall classification accuracy was 74.1%. The final result was strongly affected by the lack in sensitivity, which was only 65.2%, where in comparison a specificity of 81.0% was achieved. All results are summarised in Table 4.10.

Table 4.10 RBF SVM results for data set A.

Parameters		Training			Test		
log2(*C*)	**log2(*γ*)**	**Sens. %**	**Spec. %**	**Acc. %**	**Sens. %**	**Spec. %**	**Acc. %**
12	-2.75	90.3	98.8	96.7	62.8	80.5	72.7
15.5	-6.25	90.2	98.9	96.7	65.2	81.0	74.1
14	-4.75	90.2	98.8	96.6	63.4	80.5	73.0

Although the RBF SVM achieved the highest training accuracy among all generated models it did not yield the highest test accuracy. For this reason it must be concluded that this model is over-fitted.

4.3.1.4 Summary

The result summary for classification models built for data set A in Table 4.11 shows that the highest testing accuracy, which is 80.2%, was obtained by the LDA model. The PLS-DA model and a linear SVM achieved a testing accuracy of 80.0%. Although, all three linear classifiers achieved a similar testing accuracy, they vary in sensitivity and specificity. Among all models the LDA model achieved the highest sensitivity of 100%, hence it also achieved the lowest specificity. Interpreting this result in a diagnostic way all positive samples would be identified. However, about 30% negative samples were misleadingly classified as positive, which would result in an unnecessary cancer treatment for a patient. For the sentinel lymph node data set used here, this would mean the patient undergoes unneeded full axillary lymph node dissection. Under this circumstance, a reduced specificity might be acceptable for achieving an optimum of sensitivity due to the fact that not identifying metastasis is more life threatening for a patient than unnecessary surgery. However, the classification performance is

still suboptimal for a future clinical application due to the lack of specificity. In comparison, the LDA, PLS-DA and linear SVM model achieved an equal or greater sensitivities as reported for touch imprint cytology and frozen section analysis. Nonetheless, these methods demonstrate specificities above 99%. Since high specificity is of great interest in order to avoid unnecessary lymph node clearance in breast cancer patients diagnostic models are needed that allow a specificity higher or at least equal than the diagnostic techniques mentioned earlier in order to apply Raman spectroscopy as a clinical routine tool.

Table 4.11 Summary of results for data set A.

	Training			**Testing**		
Method	**Sens. %**	**Spec. %**	**Acc. %**	**Sens. %**	**Spec. %**	**Acc. %**
LDA	100	91.1	93.1	100	68.6	80.2
PLS-DA	92.0	95.3	94.5	86.6	74.8	80.0
SVM:						
- linear	89.4	98.83	96.5	78.7	81.0	80.0
- polynomial	0	100	75	0	100	56.2
- RBF	90.2	98.9	96.7	65.2	81.0	74.1

For this data set linear methods, LDA, PLS-DA and the linear kernel SVM, performed better than non-linear methods. The polynomial SVM lacks completely in sensitivity and due to that it is unable to predict the class-membership of the independent test set. The RBF SVM resulted in a good training performance, however this performance could not be maintained when tested with the independent test set. For this reason it must be assumed that the RBF SVM is over-fitted. It is surprising that RBF SVM, which are usually a strong classification technique, were outperformed by LDA. In order to see if the performance of classifiers, especially SVMs can be improved by introducing a new data normalisation technique a

targeted spectra selection method was developed. The impact of this method on the performance of the different classification methods is outlined in the following section.

4.3.2 Data set B

4.3.2.1 Linear discriminant analysis

The LDA model was built by using spectra, which were selected according to the method described in 4.2.5. The training set was used to identify the optimum number of PCs by LOOCV. The results suggested that retaining 13 PCs leads to the best possible model. In comparison to the LDA model developed for data set A, which used a number of 18 PCs, the optimum number of PCs is lower. This can be explained by the impact of the spectra selection method, which reduced the variance within individual samples.

The optimised number of PCs was used to build the final model, which achieved a training accuracy of 89.5%. Training and testing results are presented in greater detail in Table 4.12. The testing result is notably higher than the training result and therefore it shows that this model is a good fit. The comparison of this result with the LDA results of data set A shows an increase of approximately 20%. In this manner, the spectra selection method increased the classification accuracy by increasing the specificity.

Table 4.12 LDA results for data set B.

Training			**Testing**		
Sens. %	**Spec. %**	**Acc. %**	**Sens. %**	**Spec. %**	**Acc. %**
90.2	89.3	90.3	100	91.9	93.8

4.3.2.2 Partial least square discriminant analysis

As a result the optimisation procedure suggested that eight LVs should be used to build a PLS-DA model. The resulting PLS-DA model achieved a training accuracy of 94.5% and testing accuracy of 95.2%. All results of this diagnostic model are summarised in Table 4.13. The testing performance of this model is similar to the training performance and thus, this model is considered to be a good fit since no over-fitting took place. Notably, the testing result is increased by almost 15% in comparison to the PLS-DA model for data set A. Hence the application of the spectra selection method also improved the testing result of the PLS-DA model.

Table 4.13 PLS-DA results for data set B.

Training			**Testing**		
Sens. %	**Spec. %**	**Acc. %**	**Sens. %**	**Spec. %**	**Acc. %**
89.1	90.0	94.5	100	93.7	95.2

4.3.2.3 Support vector machines

Linear kernel

First of all a loose parameter search was executed in order to optimise the error cost *C*. For this purpose, the range of C was set as $C = [2^{-5}, 2^{-3},\ldots, 2^{13}]$. As for the previous linear SVM for the temporary parameter estimation only a subset of the training data was used. According to the graph in Figure 4.11 a parameter *C* of the value 2^7 was identified as the most suitable one. This *C* value led to a cross-validation testing result of 75.4% and a cross-validation training performance of 99.8%.

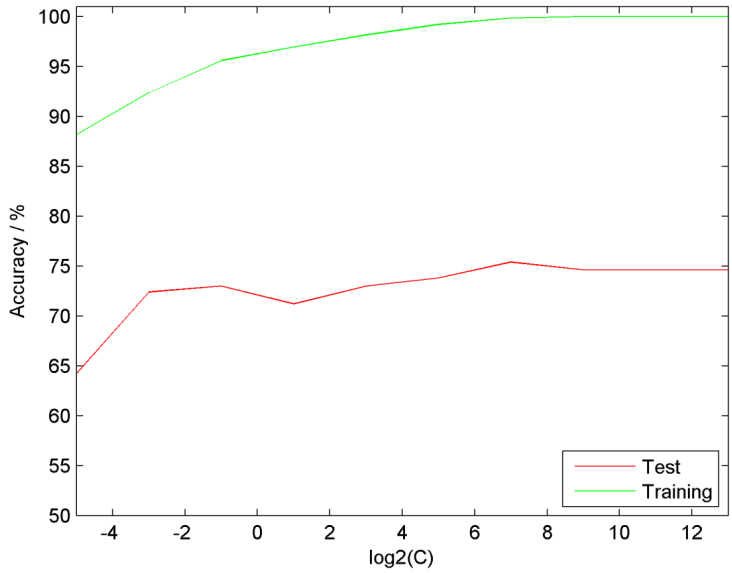


Figure 4.11 Loose parameter estimation for linear SVM (data set B).

A second parameter search was carried out in order to fine tune parameter *C*. For this search the range of *C* was set as $C = [2^8, 2^{8.25},\ldots,2^{10}]$. For the fine-tuning procedure of *C* the whole training set was used and the results showed that a parameter *C* of $2^{6.75}$ is the most suitable setting for the final model. The resulting model correctly classified 97.4% of the training set and 92.4% of the independent test set. When comparing this result with the linear SVM model for data set A it becomes obvious that the spectra selection improves the classification performance in all means.

Table 4.14 Linear SVM results for data set B.

	Training			**Test**		
log2(*C*)	**Sens. %**	**Spec. %**	**Acc. %**	**Sens. %**	**Spec. %**	**Acc. %**
6.75	94.4	98.6	97.4	100	90.1	92.4

Polynomial kernel

For this kernel type the process of model optimisations included searching for the best number of polynomial degrees d and the parameter C. First of all, a loose parameter estimation was carried out and thus the range of polynomial degrees was set as $d = [2,\ldots,8]$ and the range of parameter C was set as $C = [2^{-5}, 2^{-3},\ldots, 2^{15}]$. For the loose grid search only a subset of the training data, consisting of five positive nodes and five negative nodes, was used. Out of the result a colour map was generated as illustrated in Figure 4.12.

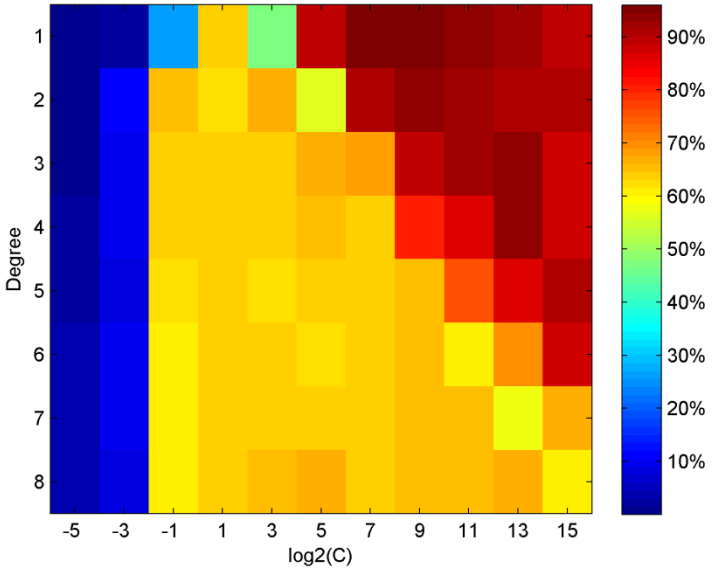


Figure 4.12 Result for loose parameter search for polynomial kernel (data set B).

The colour map in Figure 4.12 shows that a higher C as well as a lower polynomial degrees result in an improved training performance. Thus, less complex models, using a lower polynomial degree seem to be more appropriate for modelling these data. It was found that a polynomial kernel with a degree of three obtained the best classification result in the loose parameter search. The best three classification results and their related kernel parameters are summarised in the Table 4.14.

Table 4.15 Estimated parameters and grid search result for data set B.

Polynomial degree	log2(*C*)	Accuracy %
3	13	93.8
4	13	93.0
2	9	93.0

After identifying good parameter settings a final optimisation was carried out. For this purpose, the range of parameter *C* was extended around the values identified in the pervious step. In contrast, for this procedure the whole training set was used. The best three cross-validation results and the related parameters are summarised in Table 4.16.

Table 4.16 Fine-tuned parameters for polynomial SVM (data set B).

Polynomial degree	log2(*C*)	Accuracy %
3	14	87.2
4	14	85.6
2	10	86.7

The settings as shown in Table 4.16 were used to build the final model, which was then evaluated by classifying the independent test set. All three models correctly classified a minimum of 98.6% of the test set. A polynomial SVM of degree 3 even classified 99.2% of the test set correctly, which is equivalent to three misclassified spectra out of 355. All results are presented in more detail in Table 4.17. It again becomes obvious that the spectral selection method improved drastically the performance of the employed classifier, which is in this case a polynomial SVM. Although the polynomial SVM failed to classify the data of data set A it correctly classified almost 100% of the testing set in this approach. This can be explained by the normalising effect of the spectra selection method.

Table 4.17 Polynomial SVM results for data set B.

Parameters		Training			Test		
Degree	**log2(C)**	**Sens. %**	**Spec. %**	**Acc. %**	**Sens. %**	**Spec. %**	**Acc. %**
3	14	89.8	98.3	95.8	100	98.9	99.2
4	14	85.3	98.6	94.8	98.8	98.9	98.9
2	10	86.2	97.9	94.5	100	98.2	98.6

Radial basis function kernel

SVM using a RBF SVM requires the optimisation of parameters C and γ. Similar as for the other SVM approaches, for the loose grid search only a subset of the training data was used. For this search the range of parameter C was set from $C = [2^{-5}, 2^{-3},\ldots, 2^{15}]$ and $\gamma = [2^{-15}, 2^{-13},\ldots, 2^{3}]$. The obtained cross-validation results were used to generate a heatmap as illustrated in Figure 4.13. Regions coloured in dark red stand for the highest obtained classification rates, which are equivalent to correct classification rates above 90%. The parameters of the top three results were estimated and used for the second optimisation run. The top three classification rates and their related parameters are summarised in Table 4.18.

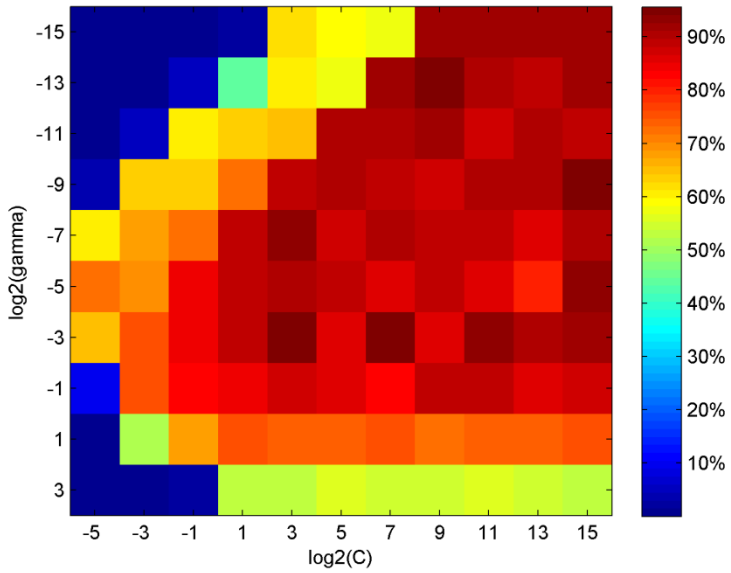


Figure 4.13 Loose grid search for RBF SVM (data set B).

Table 4.18 Estimated parameters for RBF SVM (data set B).

log2(*C*)	**log2(*γ*)**	**Accuracy %**
9	-13	95.6
7	-3	95.0
3	-3	94.8

The approximated parameters were used for a second and final optimisation step. For this purpose, the grid search was extended closely around the approximated parameters. Furthermore, the complete training set was used. The results for the finer grid search and the related parameters are summarised in Table 4.19.

Table 4.19 Fine-tuned parameters for polynomial SVM (data set B).

Range log2(*C*)	Range log2(*γ*)	Best log2(*C*)	Best log2(*γ*)	Acc. %
[8, 8.25, …, 10]	[-14, -13.75, …, -12]	10	-13.25	89.4
[2, 2.25, …, 4]	[-2, -1.75, …, -4]	3	-4	91.5
[6, 6.25, …, 8]	[-6, -5.75, …, - 4]	8	-5	91.3

The fine tuned parameters as shown in Table 4.19 were then used to build the final model, which was tested with the independent test set. The results for all three models are summarised in Table 4.19. All RBF SVM models achieved a testing result close to 100%, and one of the classifiers even 100%. Unexpectedly, the model with the lowest training performance resulted in the highest test performance. Although, this model shows the best testing performance but in comparison a relatively low training sensitivity.

Table 4.20 Polynomial SVM results for data set B.

Parameters		Training			Test		
log2(*C*)	log2(*γ*)	Sens. %	Spec. %	Acc. %	Sens. %	Spec. %	Acc. %
10	-13.25	82.2	98.4	93.7	100.0	100	100
3	-4	95.1	99.2	98.0	98.8	98.9	98.9
8	-5	98.4	99.5	99.2	100	98.5	98.9

4.3.2.4 Summary

The spectra selection method drastically improved the performance of all applied classification techniques. Thus, all applied classification methods yielded a test result over

90%. With exception of the linear SVM all approaches showed a higher testing accuracy than training accuracy.

In this approach the non-linear SVM classified either 100% or almost 100% of the independent test data. This is a drastic improvement in comparison to the models developed for data set A where the RBF and the polynomial SVM did not perform as well. Thus reducing the variance within the spectra of a sample boosts these classifiers. For this reason it must be assumed that a good standard of pre-processing is needed in order to get a good performing non-linear SVM classifier. Accordingly, appropriate pre-processing enables that the data are separable in the higher dimensional feature space, as generated by non-linear kernels. Since this preliminary is ensured, SVMs demonstrated their known potential of separating data, which might not separable in the input space, and consequently performed better than the other applied methods. The best results of all three methods are summarised in Table 4.21.

Table 4.21 Summary of results for data set B.

	Training			**Testing**		
Method	**Sens. %**	**Spec. %**	**Acc. %**	**Sens. %**	**Spec. %**	**Acc. %**
LDA	90.2	89.3	90.3	100	91.9	93.8
PLS-DA	89.1	90.0	94.5	100	93.7	95.2
SVM:						
- linear	94.4	98.6	97.4	100	90.1	92.4
- polynomial	89.8	98.3	95.8	100	98.9	99.2
- RBF	82.2	98.4	93.7	100.0	100	100

In previous work the same data set was used to generate a PC-fed LDA model, which achieved a sensitivity of 88% and a specificity of 80% when tested by LOOCV (Smith, 2005). In comparison all classification models developed in this approach exceeded these results. Furthermore, the results obtained in this work were gathered by testing with an independent test set, which further highlights the excellent performance of these diagnostic models.

In addition, the application of SVM and Raman spectroscopy also performed better than other diagnostic techniques, including imprint cytology, for which an average sensitivity of 63% and an average specificity of 99% was reported (Tew *et al.*, 2005), as well as frozen section analysis, for which sensitivities of 57-87% and specificities greater than 99% were reported (Creager *et al.*, 2002). Thus, the outstanding strength of the combination of SVMs and Raman spectroscopy is the excellent sensitivity in comparison to the other diagnostic methods.

4.3.3 Assessment of model significance

The RBF SVM model built for data set B, as well as the polynomial SVM model, was used to investigate the model fit based on Monte Carlo methods due to the fact that they achieved the highest testing performance. For each of the SVM models, 150 random models were generated and the results used to establish a null distributions as shown in Figure 4.14. The empirical significance was estimated for each of the models at two different levels, 95% and 99%. An observed test accuracy above the estimated significance level of 95% is considered as highly significant and above a significance level of 99% as extremely significant. In order to calculate the empirical significance level of 95%, the null model result was estimated

which 95% of all models do not exceed. For the RBF SVM model this level was found to be 76.9% accuracy and for the polynomial model 78.9% accuracy. Since the observed testing result of both models, 100% for the RBF SVM model and 99.2% for the polynomial SVM model, is significantly above the 95% limit both models are considered as highly significant. Nonetheless, the calculated 95% level for the RBF SVM model is smaller then the 95% level for the polynomial SVM and therefore the RBF SVM is considered to be more significant than the polynomial SVM model. Equally, the 99% level was calculated for both SVM models, which was found to be 79.4% accuracy for the RBF SVM model and 82.5% accuracy for the polynomial SVM model. Both models demonstrate to be extremely significant, whereas the RBF SVM is the most significant between both models.

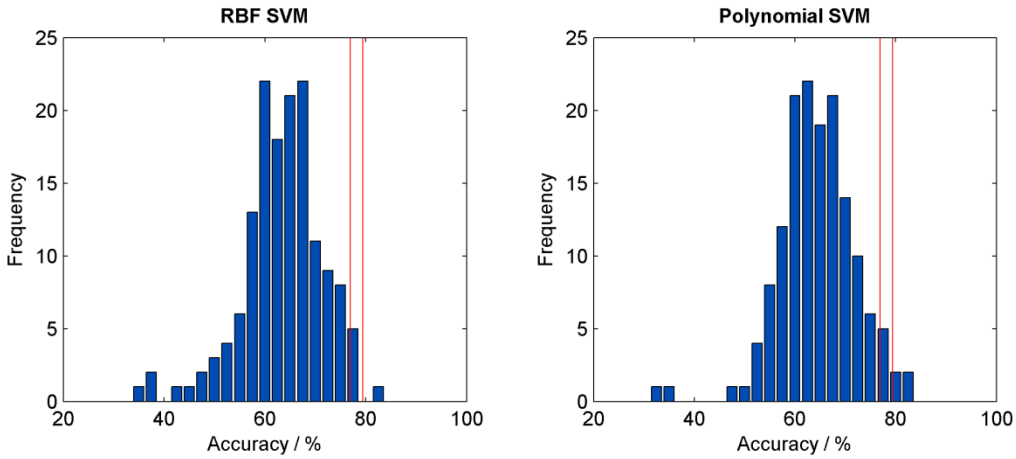


Figure 4.14 Null distributions for RBF and polynomial SVM. The empirical significance level of 95% was estimated to be 76.9% accuracy for the RBF SVM and 78.9% accuracy for the polynomial SVM. The calculated significance level of 99% was 79.4% for the RBF SVM and 82.5% for the polynomial SVM model.

In addition it was investigated if the better performing models in the null distribution have similarity to the class assignment in the original data. This was realised by generating a colour map representing the randomly assigned classes for each null model, sorted according to an increased null model accuracy. As Figure 4.15 shows no trends could be observed.

Therefore it must be assumed that null models achieving a high accuracy, do not have a greater amount of samples randomly assigned to the original sample class.

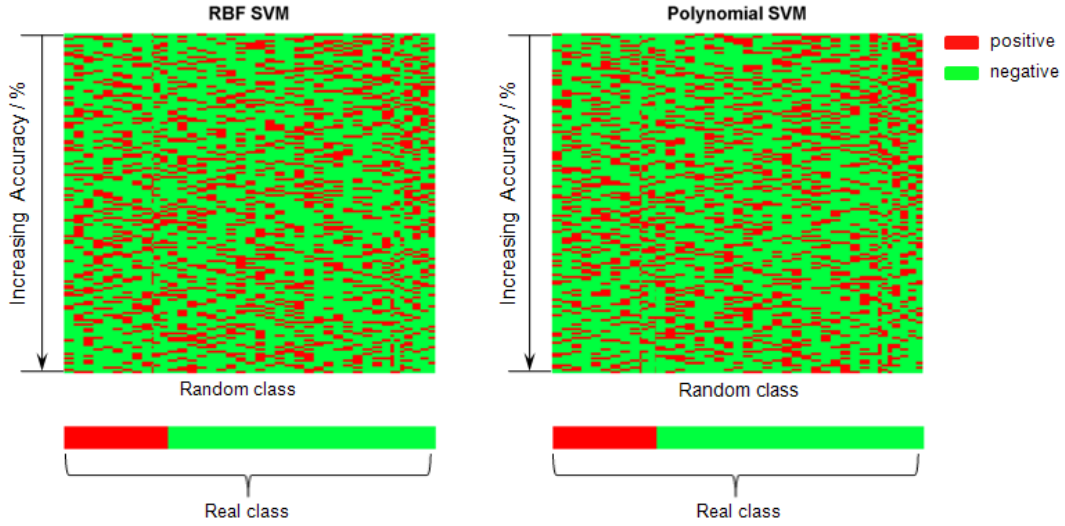


Figure 4.15 Colour map illustrating the random class assignment sorted according to increased null model accuracy. The comparison with the original sample class shows that there are no visible trends.

4.3.4 Investigation of key features

For the best performing model (RBF SVM) built for data set B model the wavenumbers that have the highest impact on the model performance were investigated. For this purpose 100 wavenumbers were eliminated from the data systematically. It showed that out of 15 generated models only two showed a decrease in testing performance. These regions were identified as 451-550cm^{-1} (decreased testing accuracy: 99.2%) and 1151-1250cm^{-1} (decreased testing accuracy: 98.9%). The peaks identified in these regions are related to disulfide bonds (540 cm^{-1}) and cytosine, guanine and adenine (all 1184 cm^{-1}). Nevertheless, the loss in performance is not significant and thus the impact of these spectral features cannot be considered as major. This becomes even more obvious by the fact that when these two regions were eliminated from the data set the generated model still achieved a testing performance of 98.9%. Thus, the remaining spectral features are still sufficient for enabling a distinct differentiation between cancerous and non-cancerous spectra. On the other hand,

creating a model by only using these two regions resulted in a relatively low testing performance of 76.9%. For this reason an assessment of whether the model performance could be improved by adding a third spectral region was performed. The two previously identified spectral regions (450-549cm^{-1} and 1150-1249 cm^{-1}) were combined with one of the remaining ones. Thus, a total of 13 models were generated. It showed that an extract of 300 wavenumbers is sufficient to achieve a 100% testing result and a training result of 90.3%, which is close to the 93.7% testing result of the model built on the whole spectral range. This training and testing result was achieved by the combination of the spectral intervals starting from 450-550cm^{-1} and 1150-1350cm^{-1}. Peaks identified in these regions are 540cm^{-1} (disulfide bonds), 1184cm^{-1} (cytosine, adenine and guanine), 1264cm^{-1}(amide III mode of α-helix and =C-H plane bending in lipids), 1304 cm^{-1} (CH_2 deformation in lipids, adenine and cytosine). Mean spectra for negative and positive lymph nodes are shown in Figure 4.16. Peak assignments for mean spectra and the performance of all individual models are summarised in Table 4.22.

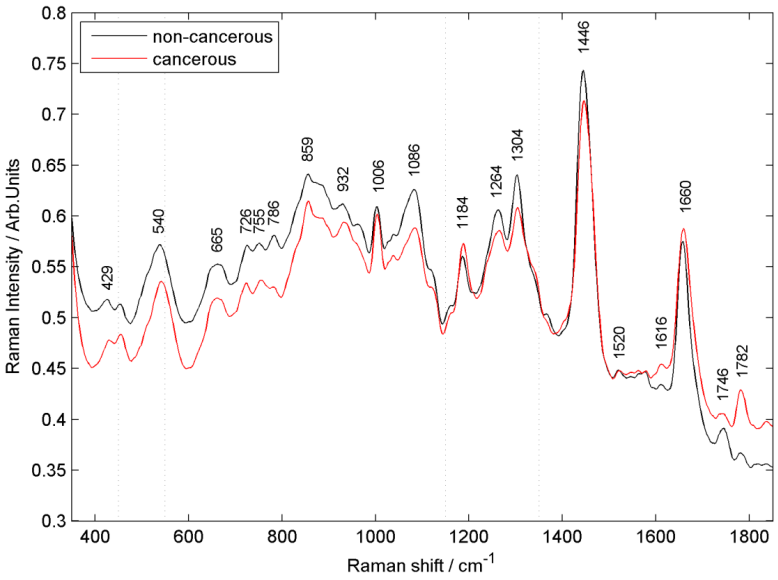


Figure 4.16 The graph illustrates the mean Raman spectra for cancerous and non-cancerous samples.

Table 4.22 Peak assignments for mean Raman spectra for cancerous and non-cancerous samples. The combination of the three intervals highlighted in gray resulted in a 100% testing performance of the RBF SVM model built for data set B.

Interval/cm^{-1}	Peak position/cm^{-1}	Major Assignments
350-450	429	Calcium hydroxyapatite
450-550	540	Disulfide bonds
550-650		
650-750	665	Thiamine
	726	C-S(protein), CH_2 rocking, adenine
750-850	755	Symmetric breathing of tryptophan
	786	DNA: O-P-O, cytosine, uracil, thymine
850-950	859	Tyrosin, collagen
	932	Skeletal C-C: α-helix
950-1050	1006	Phenylalanine, carotenoids
1050-1150	1086	Skeletal C-C stretch
1150-1250	1184	Cytosine, guanine, adenine
1250-1350	1264	Amide III (α-helix), =C-H in plane bending (lipid)
	1304	CH_2 deformation (lipid), adenine, cytosine
1350-1450	1446	CH_2 bending modes of proteins
1450-1550	1520	-C=C-carotenoid
1550-1650	1616	C=C stretching mode of tyrosine and tryptophan
1650-1750	1660	Amide I (protein)
	1746	C=O stretch (lipid)
1750-1850	1782	Unknown assignment

4.4 Conclusion

This work demonstrated that SVM coupled with Raman spectroscopy is a superior approach over traditional methods for the classification of Raman spectral data derived from tissue. Especially, the RBF SVM shows high diagnostic potential for future application due to the

fact that this achieved 100% classification accuracy on previously unseen data. This approach also exceeded other techniques such as touch imprint cytology and frozen section analysis. Thus the high potential of SVM and Raman spectroscopy for diagnostic application in lymph node assessments of breast cancer patients was demonstrated.

As reported the spectra selection and therefore removing potentially corrupting artefacts significantly improved the performance of the employed classifiers. However, in real clinical applications further data variations and corrupting artefacts might occur and for this reason it is important that employed models are robust enough to reliably classify data, which might contain imperfections. In the next chapter the robustness of the developed models for future clinical application is addressed.

5 Robustness assessment of classification models built for Raman spectroscopy

5.1 Introduction

The future application of Raman spectroscopy as a routine technique for cancer diagnosis strongly depends on chemometric pattern recognition techniques. If pattern recognition based on Raman spectroscopy is to be translated from the research laboratory to the clinic, its real world performance and limitations need to be fully understood. This necessitates rigorous model testing, which goes further than testing a classification model with an unseen testing set. In order to fully assess the performance of a diagnostic model it is important to take account of the fact that acquired data might be subject to error from a range of sources such as by system to system variation, working condition and by the operator.

In this chapter a series of methods to simulate the effect of error sources on the data set is presented. This includes various impacts, such as linear spectral shifts - which either shift the whole spectra by a wavenumber at a time or modify the intensity of each spectral point linearly, non-linear spectral shifts - either resulting in a stretching or a bending of a spectrum, and random noise, which decreases the signal to noise ratio. These are all potential errors which are suspected to occur by differing amounts when moving between instruments, modifying the design and changing working conditions or sampling methodologies. In order to assess the robustness of classification models for such errors the classifiers were used to predict the class membership of the corrupted data. For this approach the classification models described in section 4.3.2 were investigated.

5.2 Methods

5.2.1 Simulation of spectral artefacts

In order to assess the model robustness different types of perturbations were simulated on all spectra of the testing set. The training set was left unmodified – this reflects the possible deployment of a method trained in a control laboratory setting into a clinical setting where sources of error are harder to control. The original models were then used to predict the class membership of the corrupted testing set. For this approach three general types of perturbation were investigated: linear shifts, non-linear shifts and random noise. A list of the applied spectral artefacts and their causes is shown in Table 5.1. For each approach the perturbation level was increased systematically. It is expected that increased corruption levels result in a loss of predictive power, allowing the assessment of robustness as a function of spectral quality. This allows the comparison of the different types of classifiers and due to that which type of classifier is more sensitive to a specific spectral artefact for this particular spectral data.

Table 5.1 Summary of all simulated spectral artefacts and potential experimental sources.

Spectral artefact	Possible sources
X-shift	- Ambient temperature change - Calibration error
Constant y-shift	- Laser intensity variation
Gradient y-shift	- Stray off axis light entering the system - Ambient light - New signals from specimens (fluorescence)
Sine perturbation	- Collected light not fully focussed onto CCD detector
Cosine perturbation	- Optical artefacts caused by vignetting in the spectrometer
Random Noise	- Reduced exposure time - Low laser power

5.2.2 Linear shifts

For the investigation of the linear shift three independent simulations were executed: a constant x-shift, a constant y-shift and a gradient y-shift.

X-shift

Raman shifts can be a result of changes in ambient temperature and poor calibration procedures. In order to evaluate the impact of a varying x-shift on the model performance the first and the last 15 wavenumbers of the training data set were eliminated. The removal of these wavenumbers was necessary in order to gain room for shifting the data set. Thus, the original spectral range of the training set was reduced from 350-1850cm^{-1} to 365-1835cm^{-1}. The reduced data set was finally used to generate the different types of classification models.

The x-shift was simulated on the testing data by extracting an alternating spectral range of the data. For instance, an x-shift of -15cm^{-1} was introduced by extracting the spectral range from 350-1820cm^{-1}. The resulting testing set was then classified by the model. In Figure 5.1 the simulation of a x-shift of 15 wavenumbers is shown.

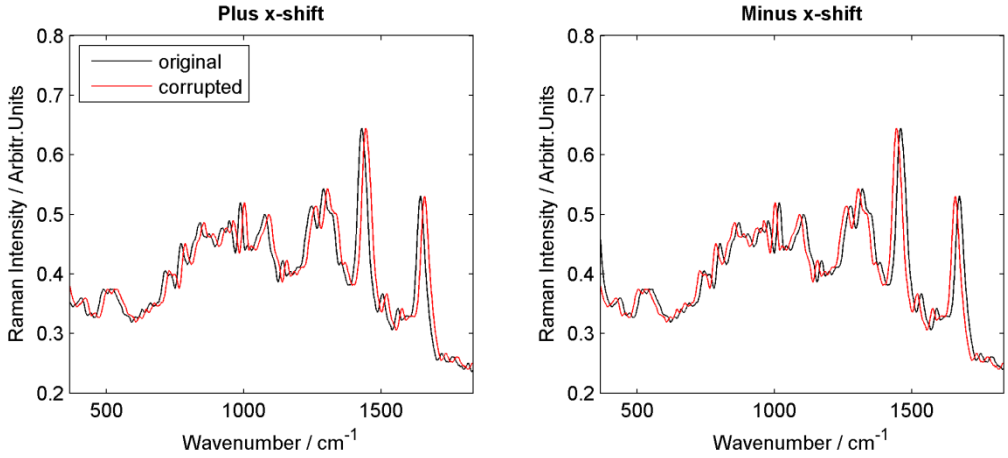


Figure 5.1 Illustration of a plus and a minus x-shift of 15 wavenumbers / cm^{-1}.

Constant y-shift

A constant y-shift was introduced by adding 0.01 arbitrary intensity units at a time to the original measured intensity of all testing spectra. Thus, for each wavenumber the intensity was consequently increased by 0.01 arbitrary units. This was executed 50 times up to an intensity increase of all spectra to a maximum of 0.5 arbitrary units. A spectrum corrupted by a y-shift of 0.15 arbitrary units is illustrated in Figure 5.2.

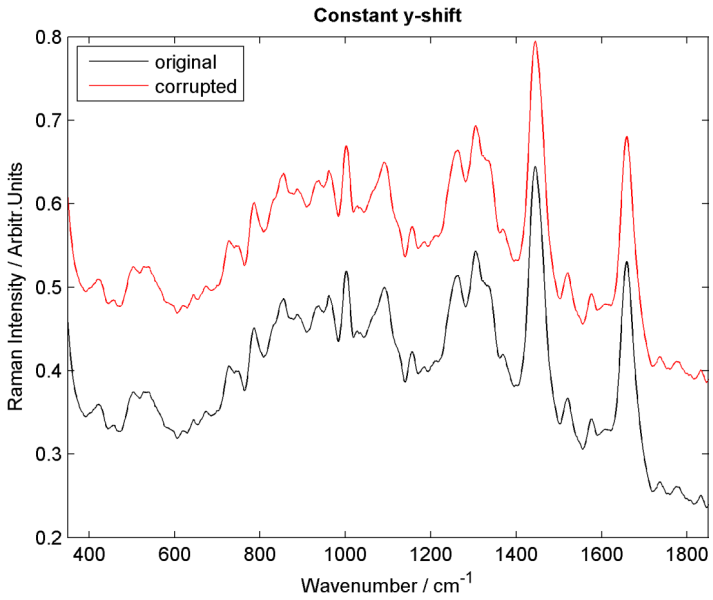


Figure 5.2 Illustration of a y-shift of 0.15 arbitrary units.

Gradient y-shift

In order to simulate a linear gradient a linear function was added to all spectra of the testing data. The gradient of this function was consequently increased by 0.0001 ranging from negative to positive gradients of 0.0013. The impact of a gradient of 0.0001 on a sample spectrum is illustrated in Figure 5.3.

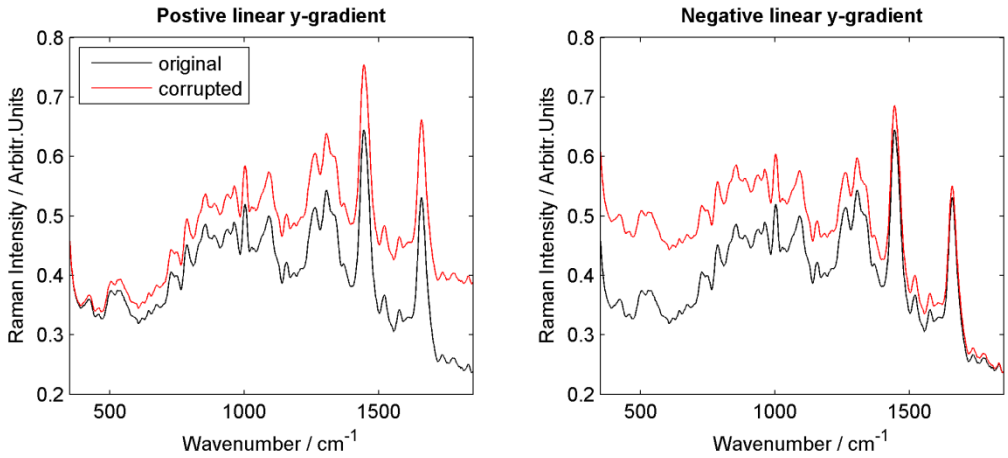


Figure 5.3 Illustration of a positive and negative gradient y-shift of 0.0001.

5.2.3 Non-linear shift

Non linear shifts were simulated in two ways. The first was sine based and the second cosine based. Accordingly, the two following functions were used to manipulate all spectra of the testing set:

Function 1: $$f(x) = a * 0.5(1 + \cos x) \tag{5.1}$$

Function 2: $$f(x) = a * 0.5(1 + \sin(x - \pi/2)) \tag{5.2}$$

The impact of the perturbation function is regulated by the amplitude *a* and due to that for both functions the amplitude was increased in steps of 0.1 starting from 0.1 up to 30. In order to corrupt the data set the resulting base function was interpolated on the testing data. As Figure 5.4 illustrates a cosine perturbation has a strong impact on the peripheral zones of the spectra. In comparison, a sine perturbation, also shown in Figure 5.4, has a higher impact on the centre of a spectrum. However, for Raman measurements a spectral stretching, as simulated by cosine perturbation, is more likely to occur than a bending, which is simulated by a sine perturbation.

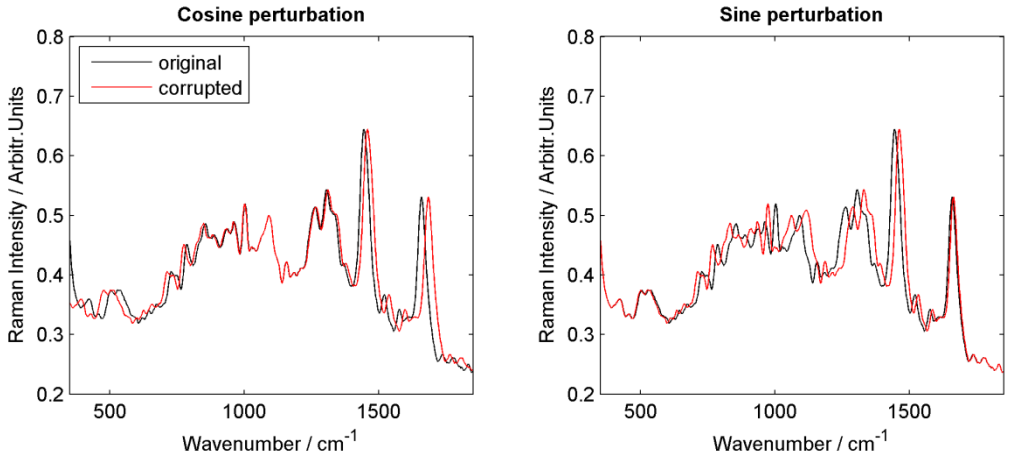


Figure 5.4 Illustration of the impact of a cosine and a sine perturbation using an amplitude of 30. The cosine perturbation results in a stretching of the sample spectra and the sine perturbation of a bending of the sample spectra.

5.2.4 Random noise

Random noise n was computed independently for every individual spectral point $s_{(i,j)}$ in the testing set $x_{(i,j)}$. The noise n can take any value between minus one and one. In order to introduce a gradient only a percentage p of the generated noise n was added to the original measured intensity:

$$x_{(i,j)} = s_{(i,j)} + s_{(i,j)} * n * p \tag{5.3}$$

In Figure 5.5 the impact of the addition of 10% noise on the Raman spectrum is illustrated. For each percentage level 100 models were generated where every time a new noise simulation was made for the testing set. The repetitions were executed due to the fact that a single repetition would not be representative because of the random nature of the perturbation.

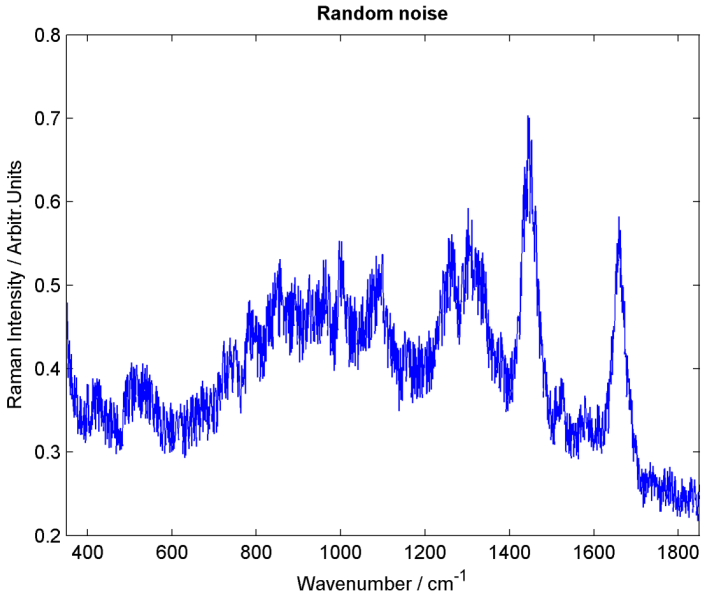


Figure 5.5 Illustration of a sample spectra after adding 10% noise.

5.2.5 Robustness score

In order to provide a summary of the overall robustness of each classification model, a score was calculated. For this purpose, each corruption approach was assigned with a total of 100 points. In the case that there was a two-way perturbation, for example the x-shift can be either positive or negative, each way was independently assigned 50 points. Consequently, a total of 600 points is theoretically achievable. In addition, an assessment level was set on which the robustness should be tested. For this approach, first of all it was set to be the maintenance of 90% accuracy. In this manner, it can be investigated how much the data can be corrupted by maintaining a minimum of 90% performance. For instance, in the case of an x-shift this is the maximum shift in *n* wavenumbers, which allows a classification performance of 90%. In order to calculate the score the proportion of the estimated perturbation limit to the applied maximum perturbation is estimated. This procedure is executed for every single corruption approach and all individual scores were summed. The higher the estimated score the higher is the robustness of a model. Thus, the score facilitates a numerical comparison of the overall robustness of classification models at a predefined performance level.

5.3 Results and Discussion

5.3.1 Linear shift

X-shift

As illustrated in Figure 5.6 a negative spectral wavenumber shift has a significantly higher impact on the classification performance than a shift in positive direction. Among all classification models the PLS-DA model was most badly affected. A shift of 15 wavenumbers in the negative direction resulted in a reduction to around 45% of prediction accuracy. In comparison, the SVM and the LDA model did not lose more than 20% in

prediction accuracy at the maximum negative x-shift. Although all models declined in overall accuracy, the sensitivity, as illustrated in Figure 5.6, was not impacted by a negative x-shift.

An x-shift in the positive direction had a strong impact on the diagnostic sensitivity of the PLS-DA and SVM model. A shift of 12 wavenumbers resulted in a complete loss of sensitivity for the PLS-DA model and a shift of 15 wavenumbers resulted in a diagnostic sensitivity as low as 1.2% for the SVM model. Overall, the LDA model demonstrated to be the most robust model in the presence of an x-shift. Further investigations showed that this is due to the fact that only a minimum number of PCs were fed into the LDA. Increasing the number of PCs resulted in a total loss of sensitivity and thus a similar performance loss as for the other classification models. Thus, the previous application of PCA for data reduction and the optimisation of the number of PCs fed into the LDA has a beneficial effect on this model and its robustness. The PLS-DA model faced the highest performance loss caused by an x-shift and due to that must be considered to be the least robust model for this type of perturbation.

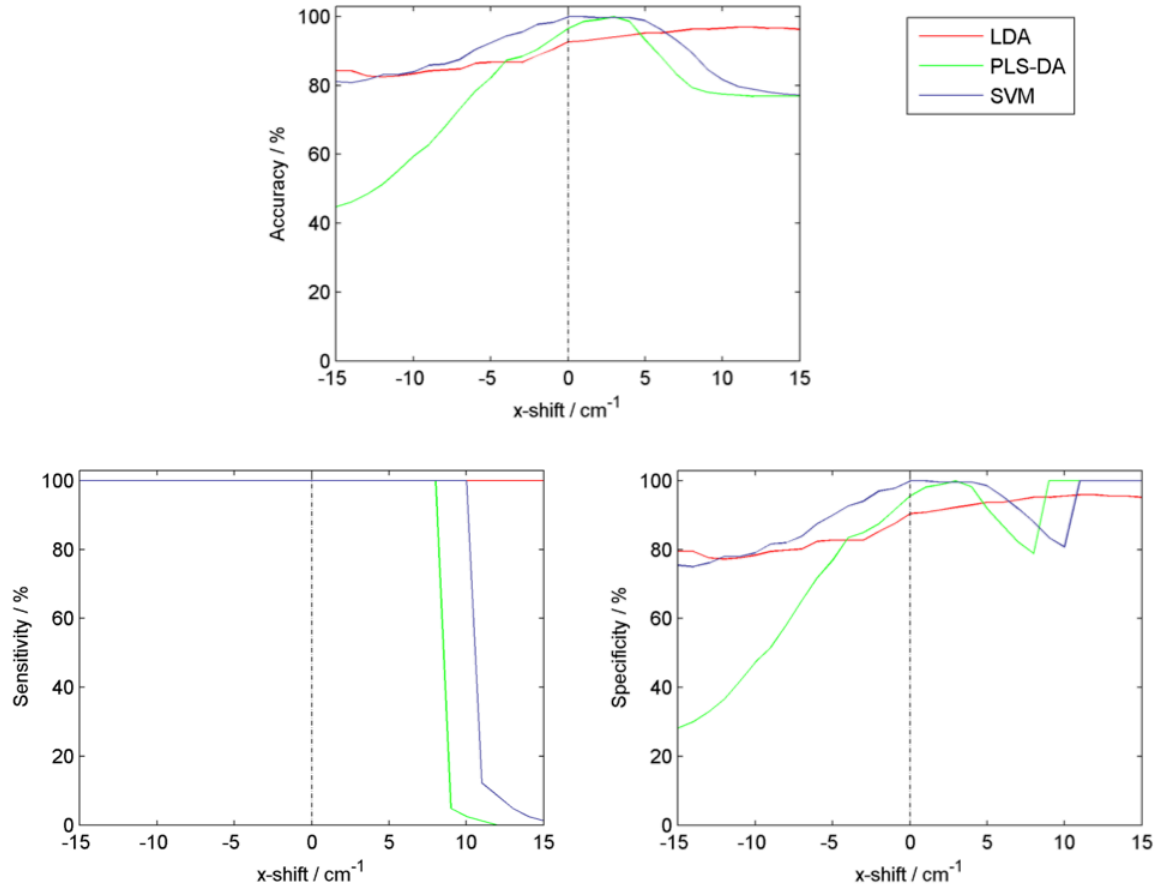


Figure 5.6 Robustness testing results for x-shift perturbation.

Constant y-shift

A constant y-shift had a severe impact on the model performance, shown in Figure 5.7, which mainly resulted in a loss of specificity. This source of perturbation did not impact the sensitivity of any of these models and due to that all performance loss is caused by a reduced specificity. The PLS-DA model achieved a classification accuracy of 39.7%, which is equivalent to a specificity of 21.6%, after increasing the intensity of all test spectra by 0.5 arbitrary units. The SVM model performed similarly by achieving a specificity of 22.7% and classifying 40.6% of the testing set correctly. The decline in specificity is illustrated in Figure 5.7. The major difference in the robustness of these models is that the SVM model loses performance abruptly, whereas the PLS-DA model loses performance more gradually. In

comparison the LDA model only achieved an accuracy of 23.4% and a total loss of specificity. For this reason it must be considered that the LDA model is most affected by the effect of a constant y-shift.

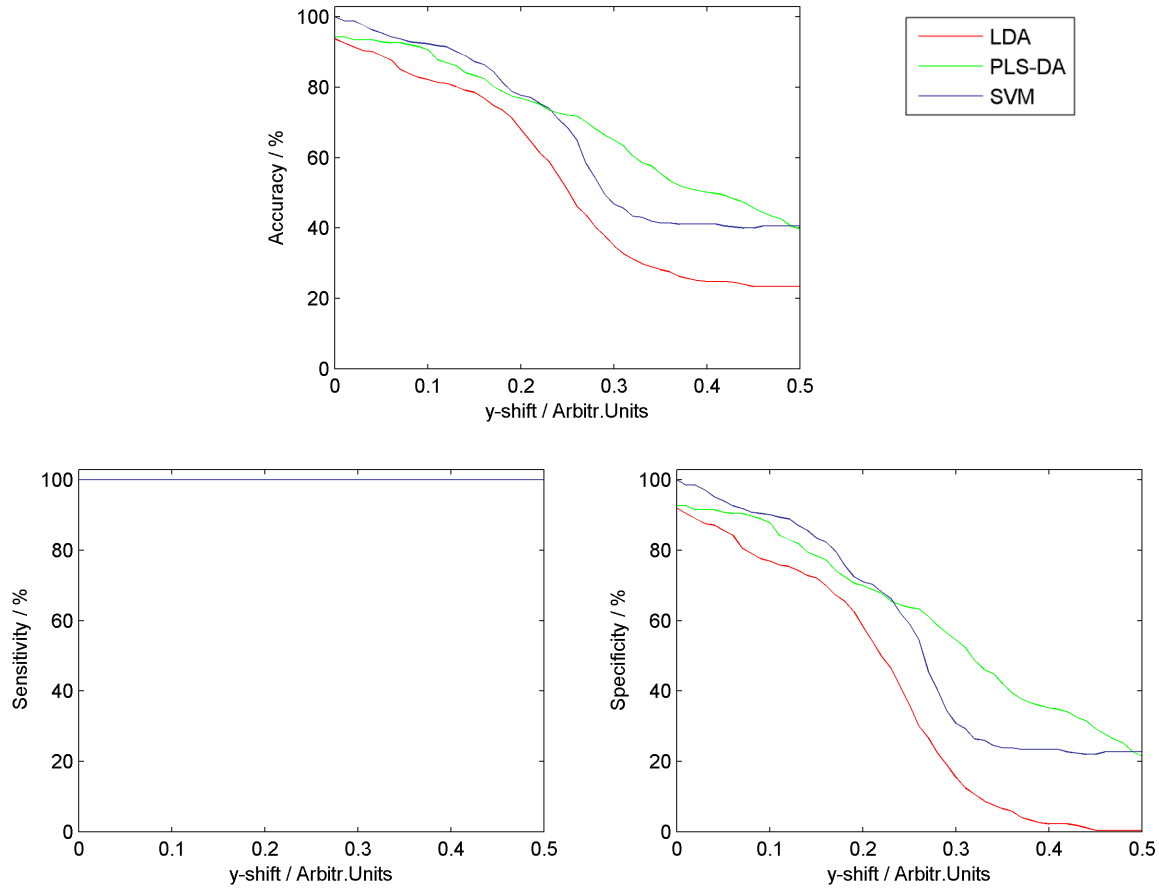


Figure 5.7 Robustness testing results for constant y-shift perturbation.

Gradient y-shift

A positive gradient drastically impacts the performance of all models as illustrated in Figure 5.8. The major reason for the performance decrease is the loss in diagnostic specificity. Thus, all normal lymph nodes were predicted as cancerous lymph nodes. For all models, as shown in Figure 5.8, a sensitivity of 100% can be maintained up to a y-gradient of 0.0013. The most

disturbed models were the LDA and the PLS-DA model, which both only classified 23.1% of all testing spectra correctly with an applied gradient of 0.0013. The relatively low accuracy in comparison to 100% sensitivity can be explained by the fact that there are more negative than positive samples in the testing set. Nevertheless, the SVM model achieved a classification accuracy of 30.7% at the same gradient level.

In comparison, a negative gradient results in a total loss of sensitivity for the LDA and the SVM model. The PLS-DA model is not impacted at all and thus maintains the original model performance up to the maximum gradient of -0.0013. For this reason it is assumed that for the PLS-DA model lower wavenumbers, which are mainly impacted by this perturbation, are of less significance than for the remaining classification models. Therefore, PLS-DA can be considered to be the most stable classification model for this type of perturbation.

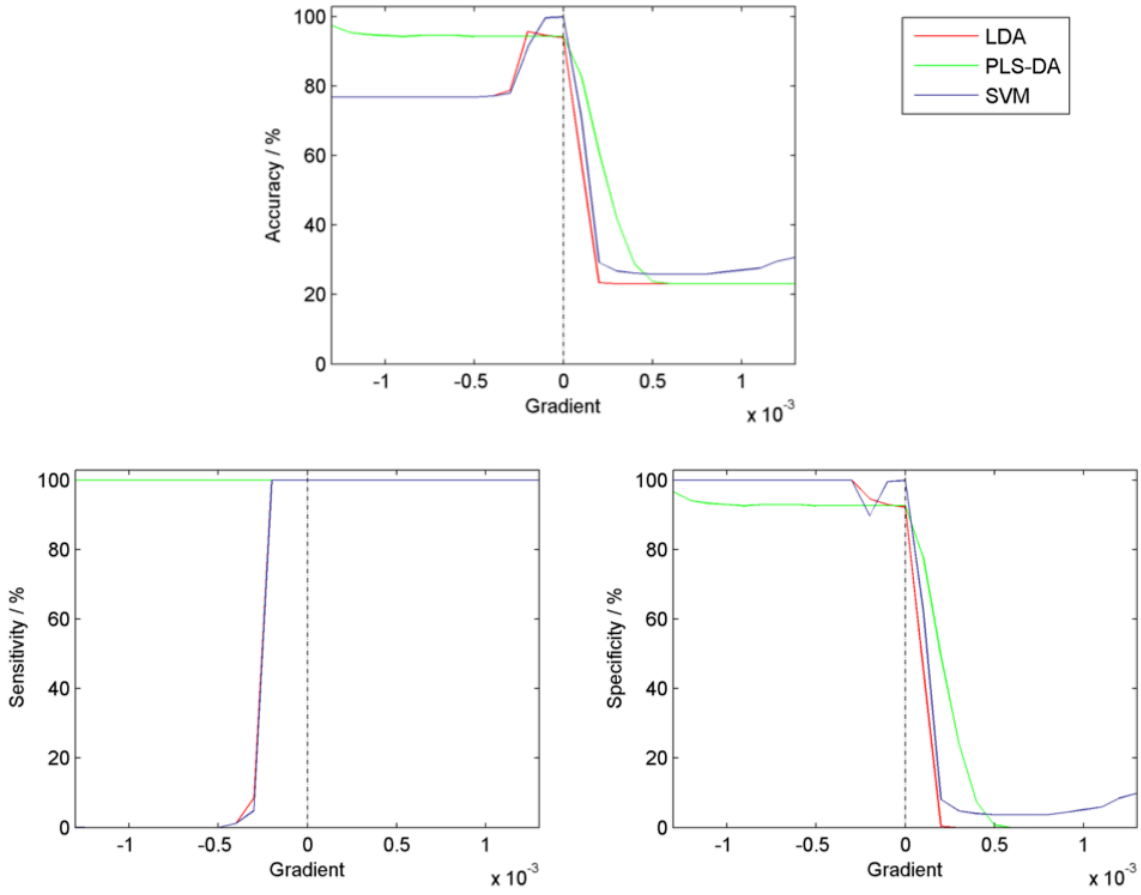


Figure 5.8 Robustness testing results for gradient y-shift perturbation.

5.3.2 Non linear shift

A cosine perturbation, which has a corrupting effect on the peripheral regions of a spectrum, has a major impact on the model performance as illustrated in Figure 5.9. For this perturbation source it was observed that the sensitivity of the LDA and SVM model was affected severely. This source of perturbation can be compensated for, up to a specific level, by each of the classifiers and then drops suddenly from a sensitivity of 100% to lower than 40%. The PLS-DA model is the first to lose sensitivity at a cosine amplitude of 16.8. The SVM model is capable of maintaining sensitivity up to an amplitude of 20.7. The LDA

classifier proved to be the most robust model by maintaining the sensitivity up to a cosine amplitude of 30.

In comparison, the sine perturbation, which impacts the centre of a spectrum, proved to have a minor impact on the performance of all classifiers. It showed that an applied sine amplitude of 30 does not result in a loss of sensitivity and in conclusion a sensitivity of 100% is maintained, as shown in Figure 5.9. The loss of specificity, also illustrated in Figure 5.9, is marginal for all classification models. The SVM model is the least impacted model for the reason that it maintains a specificity of 90.6%, which is equivalent to an accuracy of 92.7%, at the maximum level of sine perturbation. The LDA and the PLS-DA model perform equally and thus both achieve 85.1% accuracy and 80.6% specificity at the maximum sine disruption level. Based on these results the SVM model can be considered as the most robust one for sine perturbation.

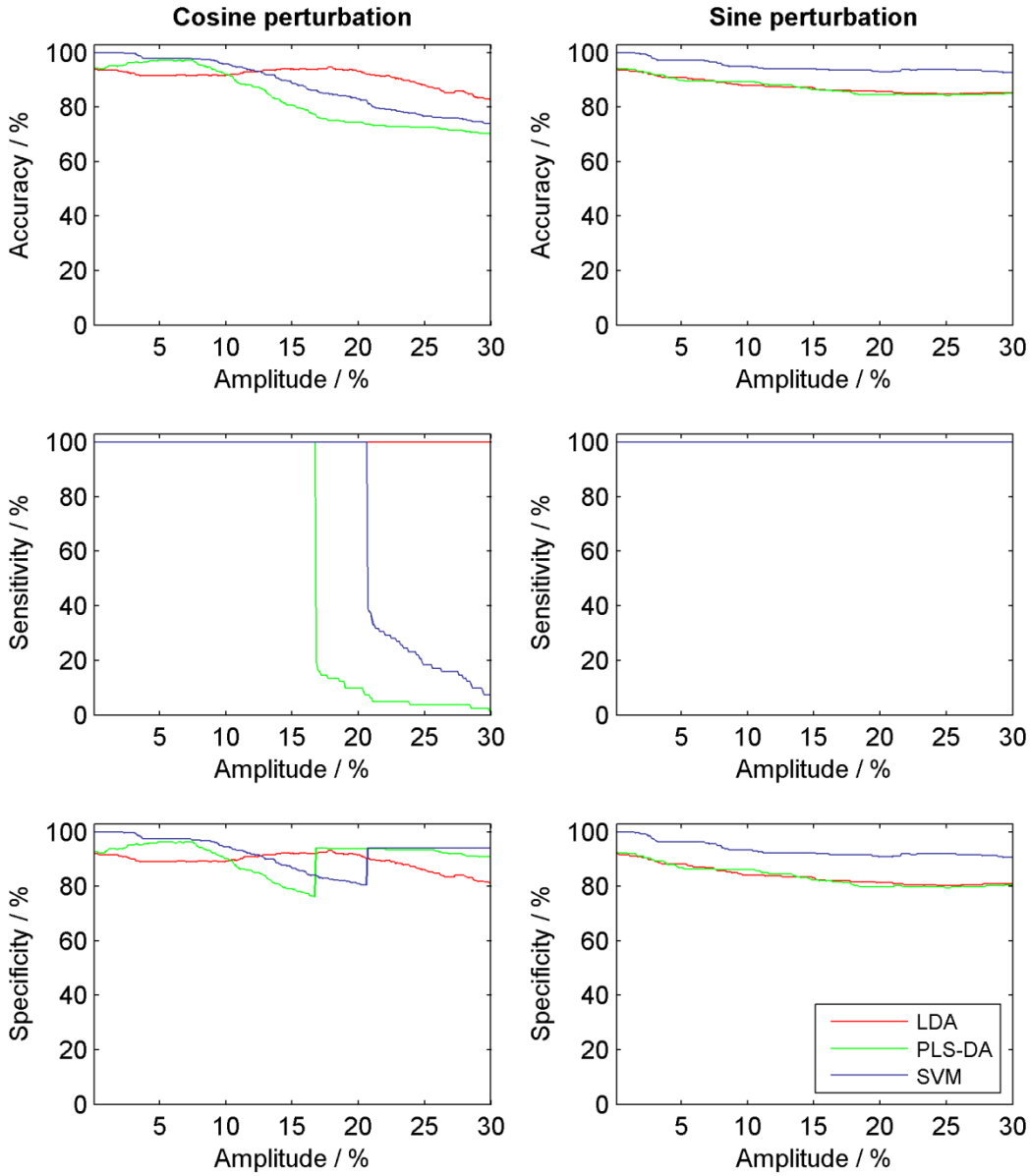


Figure 5.9 Robustness testing results for cosine and sine perturbation.

5.3.3 Random noise

For each percentage level noise was randomly added individually to every spectrum of the testing set. This procedure was repeated 100 times and the class membership was then predicted by the classifier. The average result was calculated for each noise level. The

addition of random noise proved to have only a minor impact on the performance of all classification models, as illustrated in Figure 5.10.

Although the impact of the noise on to the spectra is high no major loss in classification performance could be observed. For all models the sensitivity was not affected at all and thus only a loss of specificity was observed. It is assumed that noise did not severely corrupt the overall spectral patterns, which are crucial for the decision making of the classifiers, and consequently a high accuracy could be maintained.

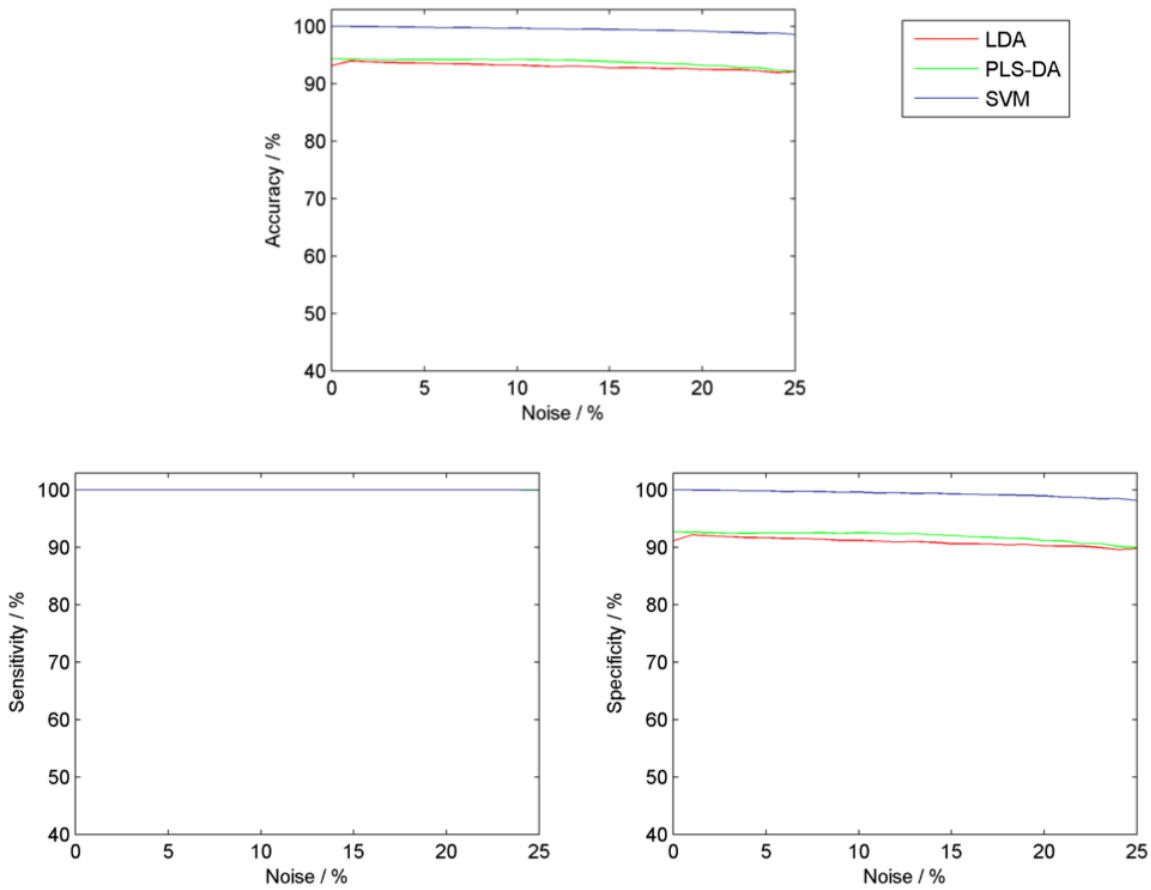


Figure 5.10 Robustness testing results for random noise perturbation.

5.3.4 Overall Robustness

In order to assess the overall robustness for each classification model a score was calculated, representing how much perturbation each model can compensate until the predictive accuracy drops under 90%. In Table 5.2 all scores for each individual perturbation source are summarised. The SVM model achieved the highest total score of all models. Nevertheless it did not demonstrate the highest robustness for each individual perturbation source. The LDA model showed to be more robust towards a positive x-shift and cosine perturbation than the other classifiers. The PLS-DA demonstrated superiority in tolerating an increasing negative y-gradient. All three models showed almost no tolerance for a positive y-gradient. Summarising, the SVM model can be considered as the most robust model since it can cope with a high level of perturbation before dropping under 90% predictive accuracy.

Table 5.2 Robustness scores for all classification models. Each single score was calculated based on the fact which maximum perturbation can be tolerated by maintaining 90% of predictive accuracy.

		LDA		**PLS-DA**		**SVM**	
	Max	**Tolerance**	**Score**	**Tolerance**	**Score**	**Tolerance**	**Score**
Pos. X-shift	15	15	50	6	20	8	27
Neg. X-shift	-15	-1	3	-2	7	-6	20
Const. Y-shift	0.5	0.05	10	0.11	22	0.14	28
Pos. Y-gradient	0.0013	$1.2*10^{-5}$	0	$3.8*10^{-5}$	2	$3.5*10^{-5}$	1
Neg. Y-gradient	-0.0013	$-2.34*10^{-4}$	9	-0.0013	50	$2.09*10^{-4}$	8
Cosine perturbation	30	23.3	78	10.7	36	14.1	47
Sine perturbation	30	7.0	23	4.7	16	30.0	100
Noise	25	25	100	25	100	25	100
Total Score			273		253		331

5.4 Conclusion

In this work it was shown the extent to which various perturbations to Raman spectra would compromise diagnostic systems built around multivariate classification models. Linear perturbations were found to be the most disruptive. Among this group it was found that a positive linear y-gradient had the strongest impact on the model performance. It was observed that even an extremely low positive linear gradient results in a drastic performance loss. Therefore, unexpected spectral features, such as stray light, fluorescence signals in new samples and ambient light signals might have the highest impact on the performance of classification models and in conclusion must be considered to be the most disrupting error source when applying Raman spectroscopy for routine diagnostics. Conversely, non linear perturbations were found to have negligible impact on the performance of the models. The same was observed for random noise. Since the major cause of random noise is reduced exposure times, these results demonstrate that a reduced exposure time would not impact on the model performance when constructed with high quality data. This demonstrates that faster spectral measurements are feasible, which is of specific importance for *in vivo* measurements where the minimisation of acquisition times is desirable.

The overall robustness does not vary drastically between the different types of classification methods. Nevertheless, it was shown that each classification method had specific strengths. In relation to the other methods, LDA is less impacted by a positive x-shift or cosine perturbation. In comparison, PLS-DA copes better with a linear negative y-gradient and SVM with a sine perturbation and random noise. In real clinical use the most likely differences between newly collected data and data used for training models would be small linear x-shifts and cosine shifts. The intensity related changes can be corrected for using normalisation methods and/or baseline subtraction. With this in mind, the most robust method would be LDA. However, since these corruptions are expected to be small in real applications the most

suitable classification method is SVM. This is due to the fact that it not only achieved the best classification performance on the original data set it is also not impacted by small x-shifts and cosine perturbation. SVM loses predictive power only at very high x-shifts and under substantial cosine perturbation. In order to further increase the robustness of the SVM model it would be required to incorporate imperfect spectra (ideally from different instruments) into the training data, such that the expected variance is captured in the model. Finally, it would be advisable to apply noise reduction methods that, for example, remove fluorescence background, prior to attempting to classify spectra.

6 Breast cancer diagnostics using ensemble support vector machines and infrared spectroscopy

6.1 Introduction

In the previous chapters support vector machines demonstrated great potential for employment as diagnostic models for Raman spectroscopic data. Thus, this chapter presents the investigation SVMs of for breast disease diagnostics using mid-FTIR micro-spectroscopy. The aim was to develop a classification model, which is capable to predict different breast pathologies reliably, including benign breast disease, ductal carcinoma *in situ* (DCIS) and invasive breast cancer. In this approach the diagnosis and staging of breast cancer is focused on micro-calcifications, which are commonly found in breast tissue and related to malignant development.

Mammography is the most important tool in breast cancer screening, as it enables detection of small masses, ill-defined densities, areas of distortion and microcalcifications. In a significant number of cancer cases microcalcification are the only indicator for the presence of malignant development (Morgan *et al.*, 2005). Due to the fact that microcalcifications are very important features in order to predict abnormalities in breast tissue. Currently, morphological features such as size, shape, clustering and branching are the only parameters to correlate microcalcifications with malignancy (Tse *et al.*, 2008). Nonetheless, this does not allow a distinctive differentiation between malignant and benign lesions. Under the circumstances that suspicious calcifications are found in mammograms typically biopsies are taken from the area concerned. Further details on the link of micro-calcifications and breast disease are provided in chapter 7.

A known problem in machine learning is that classifiers that achieve a good training performance frequently exhibit a low generalisation performance on unseen data. Exactly this circumstance was encountered when employing a traditional single SVM for the infrared data set. A possibility to overcome these limitations is to create an ensemble consisting of several classifiers and combine the output of all independent classifiers (Polikar, 2006). This procedure can be compared with the process of consulting several experts for their opinion before making a final decision. For a medical problem this would mean seeking the opinion of several doctors. However, it must be taken into account that an ensemble of classifiers might not beat the performance of the best classifier within the ensemble. Nonetheless, it certainly reduces the risk of making a poor predictive decision based on a single unfortunately selected classifier.

Another major advantage of ensemble classifiers is that they are able to successfully address the problem of data sets consisting of only a small number of samples. This is a common case in diagnostic approaches using vibrational spectroscopy, where tissue samples are obtained by invasive biopsy. Frequently, acquired data sets consist of less than 100 samples. From a typical tissue sample a modern spectrometer easily acquires hundreds of data points per spectrum. Thus, there are always more data points in a spectrum than samples measured. Data sets exhibiting these features are likely to result in unstable classifiers with low predictive power. Classification ensembles can address model instability, which results in an increased predictive power (Beleites *et al.*, 2008).

Since a traditional single SVM was not able to reliably predict an independent testing set different types of ensemble SVMs were implemented in order to overcome the poor performance. The same training data, as used to develop a single classifier were used to

implement SVM ensembles, permitting direct comparison of the two approaches. Finally, both systems were assessed with the same independent test samples.

6.2 Materials and Methods

6.2.1 FT-IR data

A total of 71 breast tissue samples were collected under ethical approval of Glouestershire LREC. The tissue samples represented three different pathology types, benign, ductal *in-situ* carcinoma (DCIS) and invasive cancer, as confirmed by histopathology. DCIS is a non-invasive form of breast cancer. In all specimens micro-calcifications were present, which can be used to stage breast cancer as demonstrated in previous work (Baker *et al.*, 2010a).

All samples were measured in paraffinsed condition using a Perkin Elmer Spotlight 300 FT-IR system in transmission mode over the spectral range 720 to 4000 cm^{-1}. Each image was generated using a pixel size of 6.25 μm and a 2 cm^{-1} spectral resolution, with 64 scans per pixel. Spectra representing tissue calcifications, which contain a distinctive phosphate peak at 1026 cm^{-1} were extracted from the generated 99 maps. The resulting data set consisted of 1628 spectra, which were all baseline corrected, smoothed (by a Savitzky-Golay filter) and normalised. For the development of classification models only the fingerprint region, ranging from 800 to 2000 cm^{-1}, was used. Mean spectra of all pathology groups are shown in Figure 6.1.

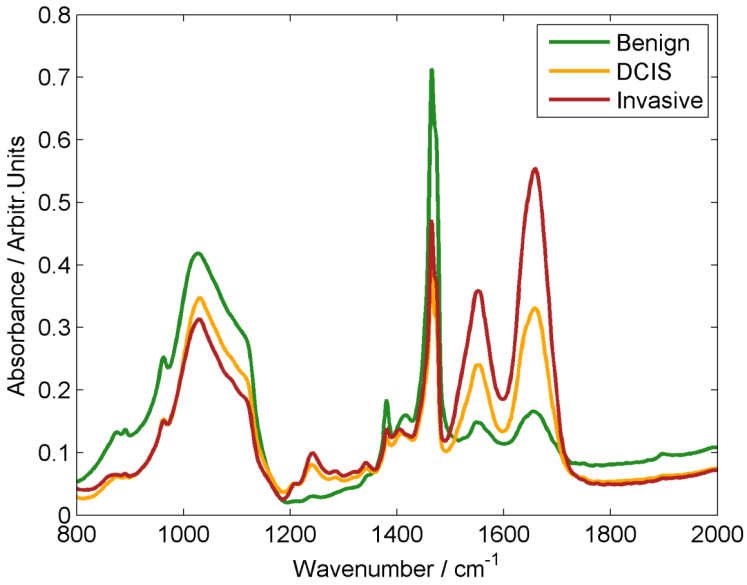


Figure 6.1 Mean spectra of the three breast disease pathologies. For the benign mean spectrum 11 samples (227 spectra) were averaged, for the DCIS mean spectrum 17 samples (332 spectra) and for the invasive mean spectrum 25 samples (567 spectra).

6.2.2 Ensemble-based systems

For building an ensemble-based system various classifiers are generated and trained independently. The resulting models are then combined in some way in order to predict the class-membership of an unseen test set (Dietterich, 2000). The general architecture of a classification ensemble is illustrated in Figure 6.2.

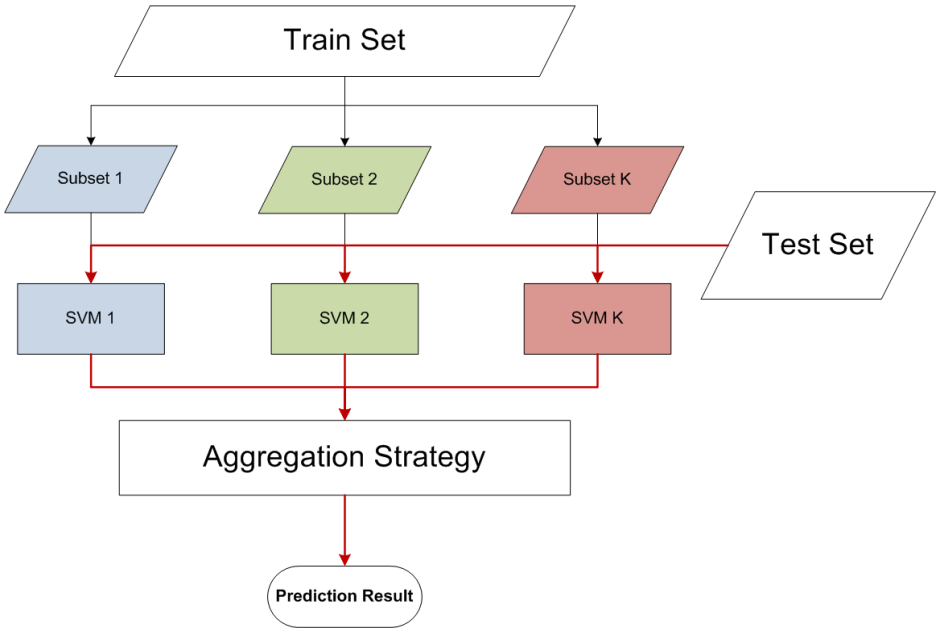


Figure 6.2 Support vector machine ensemble architecture. The train set is used to generate K smaller data subsets, for instance by bootstrapping. All subsets K are used to generate an independent SVM. Thus K SVMs are built, which are all used to predict the sample class membership of the independent test set. Consequently this results in K individual predictions for the test set, which are combined by an aggregation strategy (e.g. majority voting), in order to obtain the final prediction result.

In the past ensembles were shown to perform better than single classifiers. Hansen *et. al.* (1990) demonstrated why an ensemble of L classifiers $\{f_1, f_2, \ldots, f_l\}$ is capable of achieving a higher prediction accuracy of the test data $\boldsymbol{x}$ than a single classifier. This can be explained by the fact that the individual classifiers are different and their errors are uncorrelated. Thus, when the prediction of $f_l(\mathbf{x})$ is wrong the prediction of the majority of the remaining classifiers may be correct and subsequently the majority voting is correct. Additionally, providing the error of each independent classifier is $p < ½$ then the probability that the majority vote is incorrect decreases with the number of classifiers.

Several methods for generating ensemble-based classifiers have been developed. All of them have one general aim in common: all single classifiers should differ from each other as much as possible. In order to achieve this requirement, the different classifiers are built

by using varying training sets. For the generation of different training sets various methods are available, including bagging (Breiman, 1996), boosting (Schapire, 1990), stacked generalisation (Wolpert, 1992) and mixture–of-experts (Jacobs *et al.*, 1991). This work focused on the most popular ones – bagging and boosting – and in addition a tree-based system (Schwenker, 2000).

Once all independent classifiers are built their outputs must be combined. This can be done in various ways, for example majority vote, weighted majority vote, naïve Bayes combination or combining continuous outputs. In this work we focused on majority vote, weighted vote and naïve Bayes combination were investigated.

6.2.2.1 Bagging

Boostrap aggregating or, for short, bagging, was one of the first successfully applied ensemble-based techniques (Breiman, 1996). In a bagging approach bootstrapping is used to generate multiple data sub-sets N by randomly re-sampling from a learning set $\boldsymbol{Z} = \{(\mathbf{x}_i, y_i) | \ i = 1,2,3,...,N)\}$. Each data sub-set n (consisting of train subset n and test subset n) is used to train a SVM independently. The resulting K SVMs are then combined using an appropriate aggregation method.

Usually, bagging performs better than a single classifier under the circumstances that the single predictor is unstable. This is caused by the fact that bootstrapping assures that the sub-sets differ as much as possible and thus results in a greater improvement of the classification result. In the case that a single predictor is stable, bagging will not result in a performance improvement.

6.2.2.2 Boosting

The underlying aim of boosting is to enhance the performance from a weak classifier to a strong classifier. This idea was derived from the 'probably approximately correct' (PAC) framework (Valiant, 1984). In general, boosting is based on building a classifier ensemble through an iterative reweighting procedure. Sequentially higher weights are set on to training samples which showed to be hard to classify. Thus, a weak learning algorithm, which only might perform slightly better than random guessing can be transformed into a strong learning algorithm (Cao *et al.*, 2010).

Among all boosting algorithms, AdaBoost (Freund *et al.*, 1997) is probably the best known as well as the most successful approach. AdaBoost appears in many variations, however the most frequently used are AdaBoost.M1, capable of solving multiclass problems, and AdaBoost.R, for regression problems (Polikar, 2006).

The practical implementation of AdaBoost.M1 for a multiclass problem can be described in the following steps:

1. Input

A sequence of M training samples $\boldsymbol{Z} = [(\mathbf{x}_j, y_j)]$, $i=1,\ldots,M]$ with labels $y_j \in \Omega$, $\Omega=\{\omega_1,\ldots,\omega_C\}$

Select learning algorithm

Pick L, the number of interactions and thus the numbers of classifiers to train

Initialise the weights for the training set: $\mathbf{w}_i^1 = 1/M$, $i=1,\ldots M$

2. Training

For iterations $k = 1,\ldots,K$

Select a training subset $\boldsymbol{S}_k$ from Z using distribution $\boldsymbol{w}^k$

Train classifier D_k with $\boldsymbol{S}_k$

Calculate the weighted prediction error at iteration k:

$$D_k : e_k = \sum_{i:D_k(x_i)\neq y_i} w_i^k \tag{6.1}$$

if $\varepsilon_k = 0$ or $\varepsilon_k > 0.5$, abort

Set: $\beta_k = \varepsilon_k/(1-\varepsilon_k)$

Update and normalise weights:

$$w_i^{k+1} = \frac{w_i^k}{\sum_i w_i^k} \times \begin{cases} b_k & \text{if } D_k(x_i) = y_i \\ 1 & \text{otherwise} \end{cases} \tag{6.2}$$

3. Test

Set of unseen example **x**

Calculate the support for each class ω_i:

$$\mu_j = \sum_{k:D_k(\mathbf{x})=w_j} \log\frac{1}{\beta_k} \quad j = 1,\ldots,C \tag{6.3}$$

Select class that achieved highest vote

6.2.2.3 Tree-based ensemble

This approach aggregates several SVMs into a tree-like structure similar to a classification tree, where each node represents an independent binary SVM (Schwenker, 2000). Each SVM can either be trained for separating one single class from the remaining classes as well as for separating groups of classes. By doing so the multi-class problem is split into a

series of binary problems, which are organised in a hierarchical structure. For a three-class problem there are three distinct trees possible as illustrated in Figure 6.3.

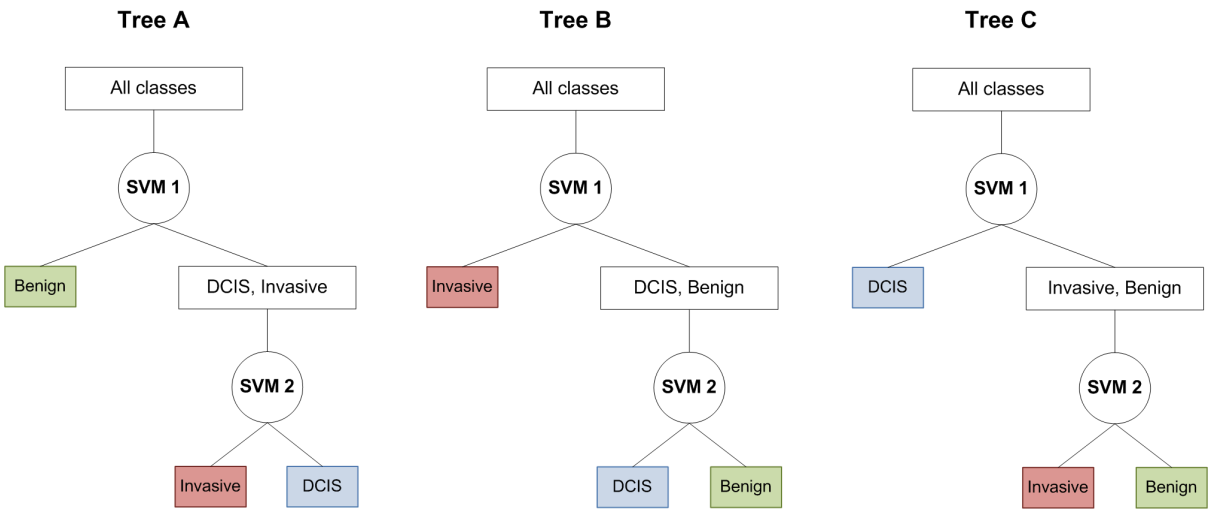


Figure 6.3 Architecture of the tree-structured ensembles for a three-class problem. These tree-structured ensembles are designed for predicting the pathologies of three different types of breast disease: benign, ductal carcinoma *in situ* (DCIS) and invasive cancer.

In order to find the optimum tree structure for a given training data, all possible trees are grown and all the integrated SVMs optimised. The most suitable tree structure can then be identified based on the training performance.

6.2.3 Aggregation methods

6.2.3.1 Majority vote

Majority voting is considered to be the simplest way of combining the predicted class labels of multiple SVMs. This system, which is also called plurality voting, selects the class which achieved the highest number of votes. Let the output of the SVM D_i be defined as $d_{i,j} \in \{0,1\}$, $i = 1,\ldots,L$ and $j = 1,\ldots,C$, where L is the number of SVMs and C the number of class labels. Thus, the output of D_i is given as a C-dimensional binary vector

$[d_{i,1},\ldots,d_{i,C}]^T$. Under the circumstances that the D_i selects class ω_j $d_{t,j} = 1$, otherwise $d_{t,j} = 0$ (Kuncheva, 2004). Accordingly the majority system will select class ω_k if:

$$\sum_{t=1}^{T} \mathrm{d}_{t,j} = \max_{j=1}^{C} \sum_{t=1}^{T} \mathrm{d}_{t,j} \tag{6.4}$$

6.2.3.2 Weighted majority vote

In an ensemble classifiers might have different accuracies and therefore some perform better than others. This evidence can be used to give a higher power to classifiers, which demonstrated to be more competent. Practically, this can be implemented by rewarding more accurate classifiers with a higher weight. The weight w_i for classifier D_i can be calculated by using the predictive error of the training performance or instead, as in this work, the predictive error of the bootstrap test set. The bootstrap test set is an internal test set and is fully independent from the final unseen test set. The weights are introduced in a similar way to the boosting procedure and thus the training error ε_i is used to calculate the normalised error β_i:

$$\beta_i = \varepsilon_i / (1 - \varepsilon_i) \tag{6.5}$$

The reciprocal of β_i is then used as the weight. However, since the training error is frequently close to zero and therefore $1/\beta_i$ can be very large, which can be a potential source of instability, it is advisable to use the logarithm of $1/\beta_i$ (Polikar, 2006).

$$\mathrm{w}_i = \log \frac{1}{\beta_i} \tag{6.6}$$

6.2.3.3 Naïve Bayes combination

This approach is called naïve Bayes combination because it is assumed that all classifiers used to derive a prediction are mutually independent (conditional independence) (Domingos *et al.*, 1997). Taking into account that the conditional independence is valid, the probability that classifier D_i labels $\boldsymbol{x}$ in class $s_j \in \Omega$, where Ω is the set of class labels $\{\omega_k, \ldots, \omega_c\}$, can be defined by:

$$\mathrm{P}(\mathrm{s} \mid \mathrm{w}_\mathrm{k}) = \mathrm{P}(\mathrm{s}_1, \mathrm{s}_2, \ldots, \mathrm{s}_\mathrm{L} \mid \omega_k) = \prod_{i=1}^{L} \mathrm{P}(\mathrm{s}_\mathrm{i} \mid \omega_k) \tag{6.7}$$

Then the posterior probability required to label x can be described according to Bayes' theorem, where $k = 1, \ldots, C$ represents the classes:

$$\mathrm{P}(\mathrm{w}_\mathrm{k} \mid \mathrm{s}) = \frac{\mathrm{P}(\omega_k)\mathrm{P}(\mathrm{s} \mid \omega_k)}{\mathrm{P}(\mathrm{s})} = \frac{\mathrm{P}(\omega_k)\prod_{i=1}^{L} \mathrm{P}(\mathrm{s}_\mathrm{i} \mid \omega_k)}{\mathrm{P}(\mathrm{s})} \tag{6.8}$$

P(s) can be ignored for the reason that it is the same for each class and therefore does not have an effect on their relative probabilities. The support for each class is estimated as

$$\mu_k(\mathbf{x}) \propto \mathrm{P}(\omega_k) \prod_{i=1}^{L} \mathrm{P}(\mathrm{s}_\mathrm{i} \mid \omega_k) \tag{6.9}$$

In order to implement the naïve Bayes combination built for a data set **Z** with N samples, for each SVM D_i in the ensemble, a $C \times C$ confusion matrix CM^i is calculated based on the achieved training result. The entry $cm^i_{k,s}$ represents the number of samples that have been correctly classified in the course of the training procedure. Therefore, these samples were assigned with the true class ω_k to class ω_s by the classifier D_i. The probability $\mathrm{P}(\mathrm{s_i}|\omega_\mathrm{k})$ is given by $cm^i_{k,s}/N_k$ and the prior probability for class for class ω_k by N_k/N and thus Equation. 6.9 can be written as:

$$\mu_k(\mathbf{x}) \propto \frac{1}{\mathrm{N}_k^{L-1}} \prod_{i=1}^{L} \mathrm{cm}_{k,s_i}^{i} \tag{6.10}$$

As a matter of fact the estimation of $P(s|\omega_k)$ can be zero, which automatically nullifies $\mu_k(\mathbf{x})$ without taking the remaining estimates into account. Titterington *et al.* (Titterington *et al.*, 1981), suggested modifications in order to address the problem of a zero estimates. For the naïve Bayes combination this can be applied as (Kuncheva, 2004):

$$\mu_k(\mathbf{x}) \propto \frac{\mathrm{N}_k}{\mathrm{N}_k^{L-1}} \prod_{i=1}^{L} \frac{\mathrm{cm}_{k,s_i}^{i} + 1/\mathrm{c}}{\mathrm{N}_k + 1} \tag{6.11}$$

6.2.4 Support vector machine implementation

The FTIR data set was randomly split into a training and a test set, within the constraint that the test set contained one third of each pathology group and the remaining samples were integrated into the training set. Thus the test set contained four benign samples, six DCIS samples and eight invasive samples. Consequently the training set contained spectra from 11 benign samples, 17 DCIS samples and 25 invasive samples.

The resulting training set was used to generate 200 subsets by bootstrapping. Those nine samples of each pathology group were randomly selected for the train subset and two samples of each pathology group for the test subset. In addition to that the spectra were balanced for each sample by randomly choosing ten spectra respectively. Thus, each sample is represented by varying spectra throughout the bootstrap sets. This allows capturing a greater variance of samples and balancing the data throughout at the same time. The resulting 200 bootstrap sets were used for the development of ensemble classifiers. For the optimisation of the RBF SVM parameters a grid search was applied by

setting the parameter ranges for $\gamma = [2^{-19}, 2^{-17}, 2^{-15}, 2^{-13}, 2^{-11}, 2^{-9}, 2^{-7}, 2^{-5}, 2^{-3}, 2^{-1}, 2, 2^{3}]$ and for $C = [2^{-5}, 2^{-3}, 2^{-1}, 2, 2^{3}, 2^{5}, 2^{7}, 2^{9}, 2^{11}, 2^{13}]$. All data analysis was performed using Matlab 2008a (Mathworks Inc., Natick, MA) and LIBSVM toolbox (Chang *et al.*, 2001).

6.2.4.1 Single SVM classifier

Two different RBF SVMs were implemented, one was optimised by leave one-sample-out cross-validation (LOOCV), the other was optimised by bootstrapping. For the first single RBF SVM the complete training set, as described in the previous section, was cross-validated by leaving one patient sample out. This LOOCV procedure was executed in a grid search and according to the mean LOOCV result the best parameters were estimated ($C = 2^{13}$, $\gamma = 2^{-7}$). The optimised parameters were then used to build the final SVM.

In a second approach a single SVM was optimised by bootstrapping. For this purpose the 200 bootstrap datasets, as described in earlier, were individually optimised in a grid search. For each parameter combination in the grid search the bootstrap train set was used to build a RBF SVM, which was then tested with the bootstrap test set. According to the mean test results of all 200 bootstrap sets the optimum parameters were estimated ($C = 2^{3}$, $\gamma = 2^{-1}$). The resulting parameters and the complete training set were then used to build the final RBF SVM.

6.2.4.2 Ensemble classifier

For the bagging approach each of the 200 bootstrap sets was individually optimised in a grid search. The optimum parameters were estimated according to the highest test performance and used to build the final model. All optimised models were then integrated

into the ensemble. For the combination of the predictions derived from the various ensemble members, three different methods were implemented: majority vote, weighted majority and naïve Bayes. The performance of the resulting ensemble was evaluated with the independent test set.

For the boosting ensemble all individual SVMs were built according to the SVM parameters as previously optimised in the bagging approach. Each of these models was individually subjected to the AdaBoosting procedure by executing a total of 25 boosting iterations. Since 25 iterations is a generalised setting, this number was optimised afterwards for each SVM separately. This was achieved by stopping the boosting at the iteration where the maximum predictive accuracy for the bootstrap test set was achieved. It showed that AdaBoosting could improve the predictive accuracy of only 50 out of 200 models. For the remaining 150 models the boosting procedure resulted in an overfitting and consequently decreased the predictive accuracy of the bootstrap test set. For this reason it must be assumed that most of the SVM models were already strong classifiers, which explains why the performance could not be improved any further. The 50 boosted models were finally integrated in the ensemble. The other 150 models were integrated in their original, not boosted, form. For the aggregation of the outputs of the 200 individual SVMs different aggregation methods, including majority vote, weighted majority vote and naïve Bayes, were implemented.

Since the investigated data set consisted of three classes, three different types of tree-structured systems were implemented as shown in Figure 6.3. Each of the SVM bootstrap data sets was used to generate a tree-structured SVM approach. For this purpose each node in the tree, consisting of an RBF SVM, was optimised in a grid search. In this manner, for

each of the three tree-structured ensembles 400 binary SVMs were optimised. In order to combine the multiple prediction outputs of the tree-structured ensemble different aggregation methods were implemented, including majority vote, weighted majority vote and naïve Bayes.

6.3 Results and Discussion

6.3.1 Single classifiers

The RBF SVM model, optimised by LOOCV was assessed with the independent test set and correctly classified 66.9% of all test spectra and 66.7% of all patient samples.

The other single RBF SVM model, which was optimised by bootstrapping, correctly predicted the class membership of 74.3% of all spectra, which is equivalent to 77.8% patient samples. A patient sample was assigned to be correctly classified when the majority of spectra are assigned with the correct class membership. The testing accuracies for all individual classes are summarised in Table 6.1.

Table 6.1 Prediction accuracies for individual classes achieved by a single SVM, optimised by leave one patient sample out cross-validation or bootstrapping.

	Benign		**DCIS**		**Invasive**	
	Spectra	**Patients**	**Spectra**	**Patients**	**Spectra**	**Patients**
Cross-validation	53.0%	50.0%	61.3%	66.7%	83.2%	75.0%
Bootstrapping	62.9%	75.0%	69.6%	66.7%	87.7%	87.5%

In comparison to the SVM optimised by cross-validation, the SVM model optimised by bootstrapping predicted 11.1% more samples correctly. An increase in performance was

observed for benign and invasive samples. No improvement could be seen for DCIS samples. This result demonstrates that bootstrapping results in more accurate classifiers than optimisation by cross-validation.

In previous work calcification spectra derived from all three pathology groups were used to develop a PC-fed LDA model, which was tested by LOOCV. It was reported that the developed model correctly classified 71.3% of the benign spectra, 56.6% of DCIS spectra and 81.1% correctly (Baker, 2009). The comparison of these results with the ones yielded by the single SVM shows that the application of SVMs improved the predictive performance. The LDA model only classified 56.6% of the DCIS samples correctly whereas the SVM model predicted the class-membership of 69.6% of DCIS spectra correctly. Since the SVM model was tested by an independent, which is a more rigid assessment than LOOCV, this further demonstrates the superiority of the SVM model over the LDA model.

Although, the single RBF SVM model performed better than the LDA model it did not achieve an as good predictive accuracy as the RBF SVM model developed for the Raman data set. The most likely reason for this is that the Raman SVM model only separated between two classes, cancerous and non-cancerous. The development of a multiclass model is much more difficult due to the fact that the progression from one disease state into another one might not always be homogenous. This implies that for instance a DCIS sample might be already progressing into the next higher stage, which is invasive cancer. Thus, a classifier trained on a data set, which contains misleadingly assigned samples results in a decreased performance. In a similar way samples might be differentially staged by different histopathologist (intra-observer disagreement), which also negatively impacts

the model development. Summarising, uncertainty in the class-membership assignment can negatively impact the classification performance. In contrast this circumstance were not present during the development of the classification model built for lymph node diagnostics based on Raman spectroscopy.

6.3.2 Ensemble classifiers

The performance of the generated bagging ensemble was evaluated with the independent test set. Among the different aggregation methods, the majority vote proved to be the best choice by classifying 80.1% of all spectra correctly. Furthermore, this bagging approach predicted the pathology of 88.9% of all patient samples correctly. In comparison, the weighted vote aggregation also classified 88.9% of the patient samples (78.9% of spectra) correctly and the naïve Bayes combination predicted 88.9% of the patient samples (79.9% spectra) correctly. These two combination methods resulted in a higher misclassification rate of invasive cancer and DCIS samples than the majority vote combination. Contrary to expectations, the majority vote did not improve the classification result. More details on prediction accuracies for all aggregation methods are provided in Table 6.2.

Table 6.2 Prediction accuracies for bagging and boosting SVM ensemble.

	Benign		**DCIS**		**Invasive**	
	Spectra	**Patients**	**Spectra**	**Patients**	**Spectra**	**Patients**
BAGGING:						
Majority Vote	70.5%	75.0%	81.7%	100%	85.5%	87.5%
Weighted Vote	68.9%	75.0%	80.1%	100%	84.9%	87.5%
Naïve Bayes	72.0%	75.0%	80.6%	100%	84.9%	87.5%
BOOSTING:						
Majority Vote	69.7%	75.0%	81.2%	100%	86.0%	87.5%
Weighted Vote	70.5%	75.0%	82.2%	100%	86.0%	87.5%
Naïve Bayes	70.5%	75.0%	82.2%	100%	83.8%	87.5%

The boosted ensemble was finally tested with the independent test set. As the results in Table 6.2 show, the boosting slightly improves the performance of the ensemble. Since only a small improvement was observed, it is likely to be caused by the fact that merely a quarter of the models could be boosted. Therefore it must be assumed that the boosted models did not have a significant impact on the final decision making.

Among the tree-based classifiers ensemble, type B using a naïve Bayes combination performed the best by classifying 81.5% of all spectra. The increased performance in comparison to the other tree ensembles can be explained due to a higher accuracy for DCIS spectra. This ensemble also classified 88.9% of all patient samples correctly. However, all the other ensembles, with the exception of tree-structured ensemble C, also achieved this accuracy for patient samples. Tree ensemble C achieved a significantly lower accuracy by classifying 77.8% of all patient samples correctly. Therefore, the application of this ensemble type performed equally like a single SVM optimised by bootstrapping. The decreased performance might be explained by the fact that separating the DCIS

samples first, as executed by tree model C, from the remaining classes is more difficult. This might be caused by the fact that DCIS represents the pathology state between benign and invasive cancer. The transformation from one pathology state into another occurs smoothly and thus there are no sharply defined borders between one and the next higher or lower pathology group. In this manner splitting of the most distinct class first, in this approach either the benign or the invasive samples, reflects positively on the classification performance of a tree-ensemble classifier. Detailed results of all tree-based ensembles are illustrated in Table 6.3.

Table 6.3 Prediction accuracies for tree-based SVM ensemble.

	Benign		**DCIS**		**Invasive**	
TREE A:	**Spectra**	**Patients**	**Spectra**	**Patients**	**Spectra**	**Patients**
Majority Vote	72.0%	75.0%	79.6%	100%	86.6%	87.5%
Weighted Vote	72.0%	75.0%	80.6%	100%	86.0%	87.5%
Naïve Bayes	70.5%	75.0%	80.1%	100%	86.0%	87.5%
TREE B:						
Majority Vote	72.0%	75.0%	80.6%	100%	88.8%	87.5%
Weighted Vote	72.0%	75.0%	80.6%	100%	87.7%	87.5%
Naïve Bayes	71.2%	75.0%	82.7%	100%	87.7%	87.5%
TREE C:						
Majority Vote	70.5%	75.0%	61.8%	66.7%	77.7%	87.5%
Weighted Vote	70.5%	75.0%	61.3%	83.3%	77.6%	87.5%
Naïve Bayes	71.2%	75.0%	61.8%	83.3%	78.8%	87.5%

Comparing the performance of the ensemble SVM with the results of the single SVM demonstrates the superiority of this approach. The better performance can be explained by the fact that the generation of an ensemble allows us to capture more sample variance,

which is of specific importance for small data sets or sets where different classes are represented by varying sample numbers. The classification results of the ensemble SVMs show that neither the building method nor the combination method generally have a significant impact on the ensemble SVM performance. The only observed exception was tree-structured ensemble C.

Since tree model B using a naïve Bayes combination achieved the highest correct accuracy for spectra, it is considered to be the most appropriate classifier for this data set. Table 6.4 illustrates a detailed breakdown of the exact prediction for each individual sample in the independent test set.

Table 6.4 Breakdown of results for tree model B using weighted majority vote.

		Pathology predicted per patient sample %		
Pathology	**Sample ID**	**Benign**	**DCIS**	**Invasive**
Benign:	37	**77**	23	0
	47	0	**100**	0
	72	**100**	0	0
	76	**82**	18	0
DCIS:	40	30	**70**	0
	52	30	**70**	0
	60	22	**72**	6
	71	0	**100**	0
	84	0	**92**	8
	91	0	**83**	17
Invasive:	55	0	0	**100**
	68	0	**83**	17
	69	0	8	**92**
	85	0	0	**100**
	93	0	20	**80**
	95	0	17	**83**
	99	0	0	**100**
	105	0	25	**75**

It shows that the majority of all samples were correctly classified with a minimum of a two-thirds majority. Only one benign sample was classified as DCIS and one invasive sample classified as DCIS. Of all benign samples no spectrum was predicted to be invasive. Incorrectly classified spectra were predicted to be DCIS. This suggests that in some of the benign samples development into a higher pathology grade may be present. Similarly, for the invasive samples no spectrum was classified as benign and misclassified

spectra were predicted to be spectra representing DCIS. According to these wrongly classified samples were assigned either to the next higher or the next lower pathology group, which possibly indicates a trend in disease development for each sample.

6.3.3 Number of classifiers

Each ensemble consisted of 200 classifiers, which was found to be a sufficient number in order to stabilise an ensemble. For assessing the impact of the number of models, 200 ensembles were built. The first ensemble consisted of only one SVM and successively one more SVM was added at a time. Each of the 200 ensembles was used to predict the independent test set. Figure 6.4 illustrates how the prediction error of the bagging ensemble decreases with an increased number of models. It further shows that the predictive error of the majority vote combination decreases quicker than the other combination methods. Figure 6.4 also illustrates the stabilisation of the predictive error for tree-structured ensemble B. In comparison to the bagging ensemble the error stabilises slower for this ensemble system. Interestingly these results suggest that the best predictive performance could be achieved by an ensemble consisting of only 11 SVMs. Nonetheless, a higher number of SVMs stabilises the predictive error and thus it is more likely that further unseen data is predicted with higher accuracy.

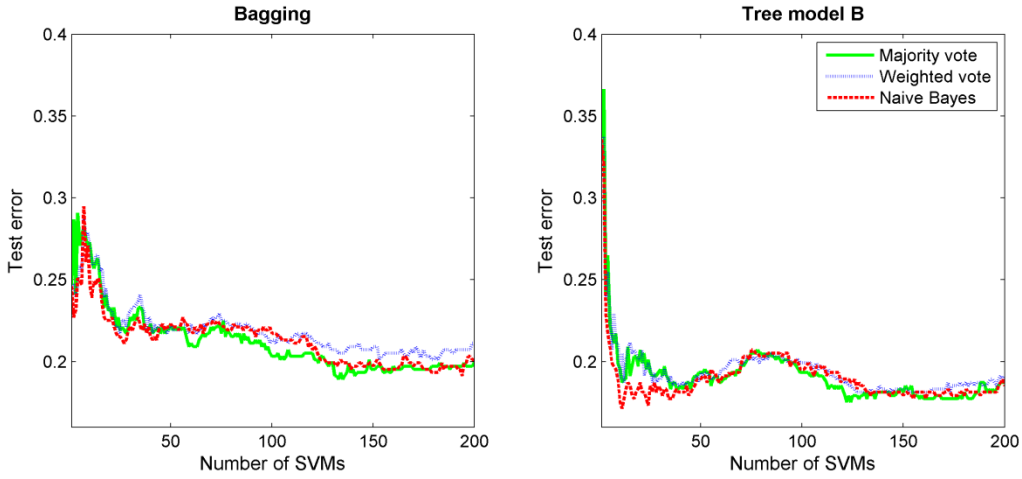


Figure 6.4 Impact of the number of SVMs on the predictive error of ensemble systems.

6.4 Conclusion

The application of SVM ensembles brings a drastic improvement for classification approaches for data that previously seemed to be difficult to classify. Thus, a diagnostic model was developed, which reliably predicts the different stages of breast disease. In addition, ensembles allow building stable classifiers for unbalanced data or data sets consisting only of a small number of samples. This brings great benefits for applications where data availability is restricted, such as biomedical research. In particular, the results of this work demonstrate the high potential of FTIR spectroscopy for diagnosing and staging breast cancer based on micro-calcification found in breast tissue.

7 Analysis of breast tissue calcifications in infrared image

7.1.1 Introduction

Naturally, calcifications are present in many different biological tissues, for instance in teeth and bones. Conversely the occurrence of calcification in soft tissue can be often a result of disease and therefore is associated with several medical conditions including crystal-associated osteoarthritis, diabetes and breast cancer (Sun *et al.*, 2003).

In general, microcalcification found in breast tissue can be divided into type I, consisting of calcium oxalate dihydrate (COS), and type II, consisting of calcium hydroxyapetite (CHAP) (Matousek *et al.*, 2007). The presence of type I calcification is related to benign lesions, in contrast type II calcifications occur in benign as well as in malignant breast lesions (Morgan *et al.*, 2005).

Although it is known that calcifications are of great diagnostic importance, the mechanism of calcification formation of breast tissue is not clear. However, it is assumed that there exist two general types of calcification mechanisms in the breast, a secretory type and a necrotic type. In the first type calcifications are built by secretion accumulation (Tse *et al.*, 2008). In this process of mineralisation carbonated hydroxiapatite crystals are decomposed in an extracellular matrix, which consists of type I collagen and other non-collagenous proteins (Morgan *et al.*, 2001). Beside this it was found that three bone matrix proteins are increasingly expressed in breast cancer, osteonectin, ostepontin and bone sialoprotein. These proteins might possibly create appropriate environment for initiating hydroxyapatite formation (Bellahcene *et al.*, 1994). The second calcification mechanism, necrotic calcification as found in comedo necrosis, is a result of rapidly proliferating tumour cells

cutting off the vascular supply, which consequently leads to tumour cell death. This type of calcification is particularly found in high grade DCIS (Tse *et al.*, 2008).

Since the calcification process in breast tissue is poorly understood it is of significant interest to gain a deeper understanding of the relation between calcification and breast disease progression. For this purpose an image analysis algorithm was developed, which builds up on the developed SVM ensemble presented in chapter 6. The aim of this method is to facilitate monitoring of the chemical transformation of calcifications during the progression of malignancy development in breast tissue. The developed and findings of an imaging method for analysis of calcifications in infrared images is presented in this chapter.

7.1.2 Materials and Methods

7.1.2.1 FT-IR data

For this approach the same data set a presented in section 6.2.1 was used. This data set consisted of 99 infrared maps obtained from 71 patient samples representing benign, DCIS and invasive breast cancer and different grades respectively. A summary of the data set including the different grades is provided in Table 7.1.

Table 7.1 FT-IR data set representing the number of available samples for each grade. There is no actual grading system for benign breast disease.

Pathology	Total of samples	Grade 1	Grade 2	Grade 3	Grade unknown
Benign	15	-	-	-	-
DCIS	23	2	8	10	3
Invasive Cancer	33	3	14	9	7

In order to allow an independent assessment of each single patient sample a total of four train and four test sets was generated randomly. It was ensured that each patient sample was only once represented in either one of the four independent test sets.

7.1.2.2 Algorithm development

A two-step imaging algorithm was developed that allows identifying calcification in infrared maps taken from breast tissue samples. In the first step calcification spectra are identified and separated from tissue spectra. Following this procedure the pathology of the identified calcifications is predicted, which can be either benign, ductal carcinoma *in situ* (DCIS) or invasive. Based on the classification result an image is generated, which presents calcifications in colour and remaining tissue in greyscale. As a colour-coding for the calcifications traffic light colours were chosen. Thus green representing benign calcifications, yellow DCIS and red for invasive calcifications. A general layout of the classifier system is shown in Figure 7.1.

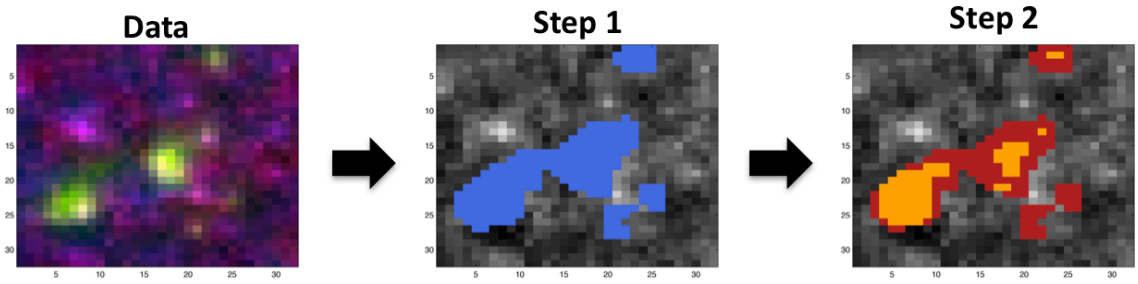


Figure 7.1 Image analysing workflow: First of all the classifier identifies potential calcifications in the image. In a second step the calcification are assigned with a pathology.

For the implementation of this system a single RBF SVM was trained to differ between calcification spectra and all other remaining spectra as commonly found in infrared maps

taken from breast tissue samples. In order to enable a faster image analysis only the spectral range from 800 to 1200cm^{-1}, representing phosphate bands, was used. As shown in figure 7.2 this allows a distinct differentiation between calcification spectra and other spectra present in infrared maps.

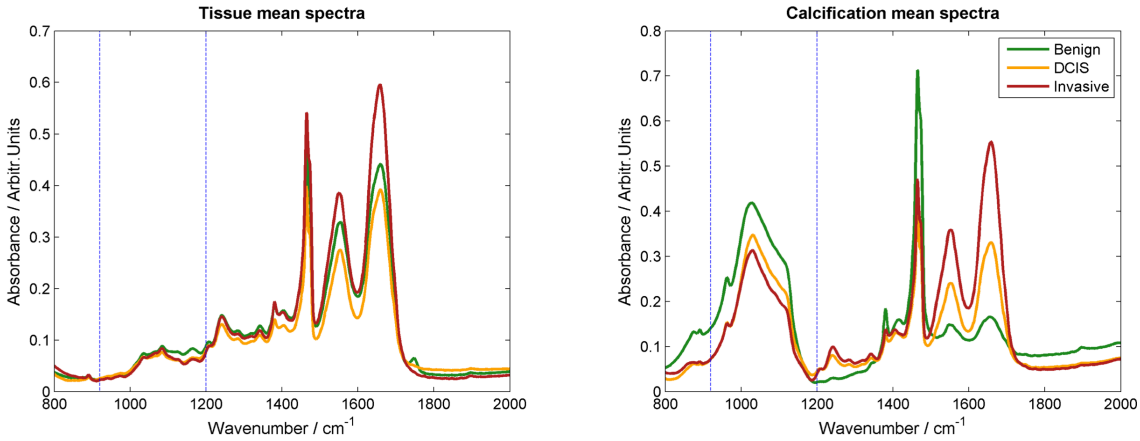


Figure 7.2 Tissue and calcification mean spectra. Calcification spectra can be distinctly differed from tissue spectra based on the phosphate peak between 1026cm^{-1}.

For the prediction of the pathology of calcifications the RBF SVM ensemble developed in the previous chapter was used. The implemented ensemble uses a tree-structured system and combined the output by a weighted vote. This classifier was selected to due the fact that it achieved the highest accuracy in predicting the independent test set in the classification approach presented in the previous chapter.

In order to analyse all individual patient samples, four different image analysis approaches were executed. For each approach the data were split into train and test set. It was ensured that each patient sample was presented once in the independent test set, which facilitated gathering colour-coded maps for each patient sample as an independent prediction. Consequently the train data was used to build the classification system, which was then used to generate colour-coded images.

7.1.3 Results and Discussion

7.1.3.1 Visualisation of calcifications

The developed imaging method was used to analyse each of the 99 infrared maps individually. As shown in Figure 7.3 this image analysis approach successfully identified micro-calcification in tissue samples. Further investigation of the calcification spectra, which were identified in the Raman maps, showed that they coincide with the spectral features as observed in calcification representing the chemical composition of different pathologies. A comprehensive summary of all generated maps is provided in Appendix A.

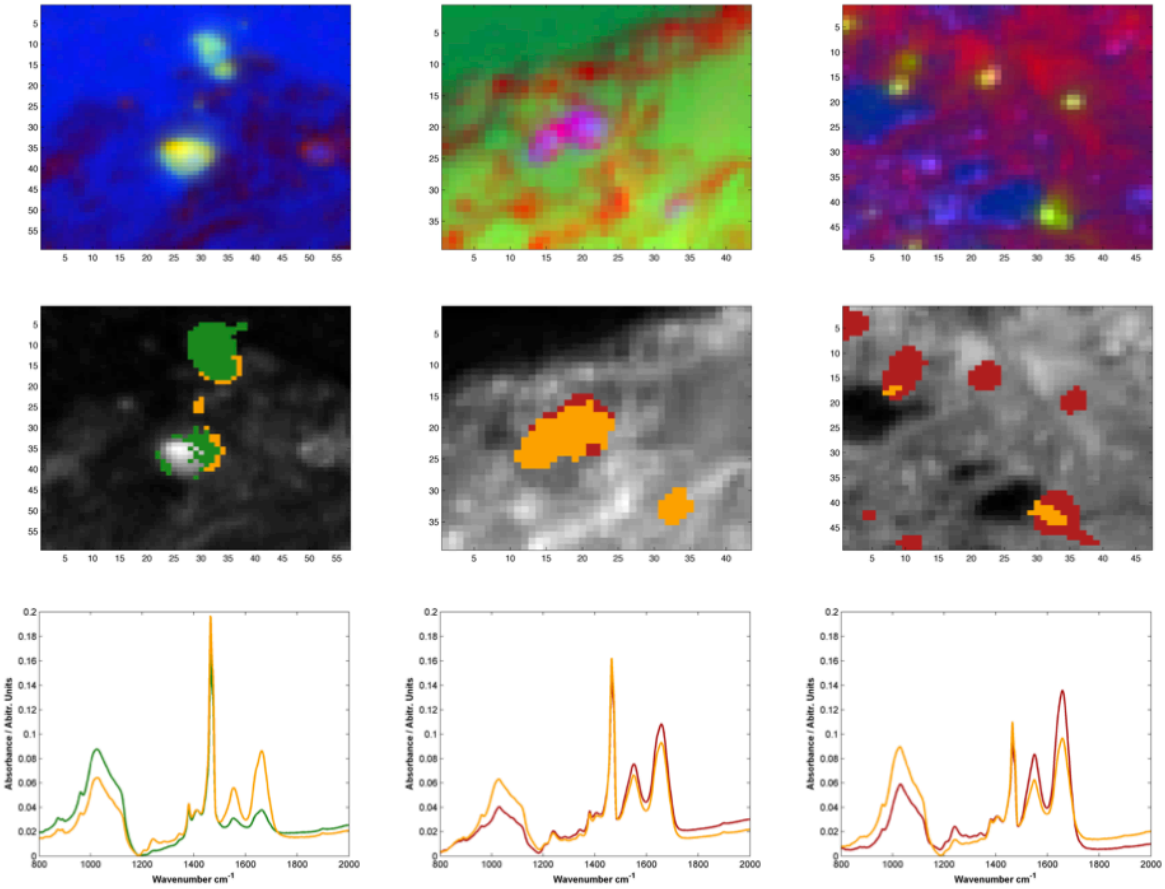


Figure 7.3 Three examples of image analysis results. The first row illustrates false colour images generated by using the first three PCs. Different colour represent different areas based on predominant spectral features, nonetheless there is no colour coding for representing different areas. The second row represents the images generated by the classifier. Coloured areas indicate calcifications, where green predicts benign, yellow DCIS and red invasive calcifications. The third row shows the mean spectra representing the identified calcification areas.

7.1.3.2 Average composition of breast calcifications

For assessing the average composition of calcifications within samples, all calcification spectra classified by the algorithm were extracted for each individual infrared map. As illustrated in Figure 7.4 it showed that calcifications found in benign breast tissue consist of an average of 62.5% benign calcification spectra, 27.2% DCIS spectra and 10.3% invasive spectra. Calcifications found in tissue samples diagnosed by histopathology as DCIS were found to be composite in the average of 31.0% benign spectra, 39.8% DCIS spectra and 29.2% invasive spectra. For invasive tissue samples the average calcification composition was found to be 9.4% benign spectra, 15.1% DCIS spectra and 75.5% invasive spectra. This result clearly indicates a trend in which the amount of invasive spectra increases with progression of malignancy. On the other hand the amount of benign spectra decreases with the progression of malignancy. This further demonstrates that DCIS is the state between pathologies since 31.0% of calcification spectra were identified to be benign and almost the same amount, 29.2%, to be invasive. In this manner the transformation from benign to malignant could be observed based on the content of spectra.

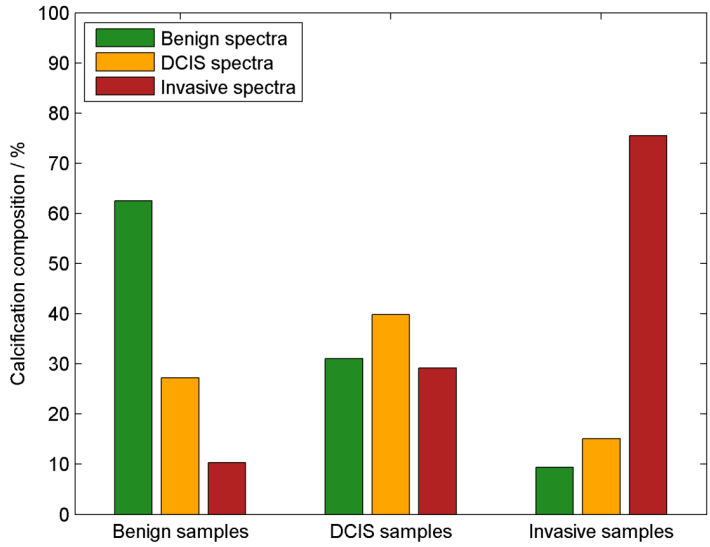


Figure 7.4 Histogram illustrating the mean composition of calcifications of the three different breast pathologies

In a similar manner it was investigated if a trend can be observed in the different disease grades. As the results in Figure 7.5 show, the amount of benign spectra continuously declines with disease progression. *Vice versa* the amount of invasive spectra increases with higher gradings. It showed that DCIS grade 2 samples contained almost the sample amount of benign, DCIS and invasive samples. This suggests that DCIS grade 2 represents a turning point in disease development. Summarising, the grading of disease development can be followed based on the composition of calcification found in tissue samples.

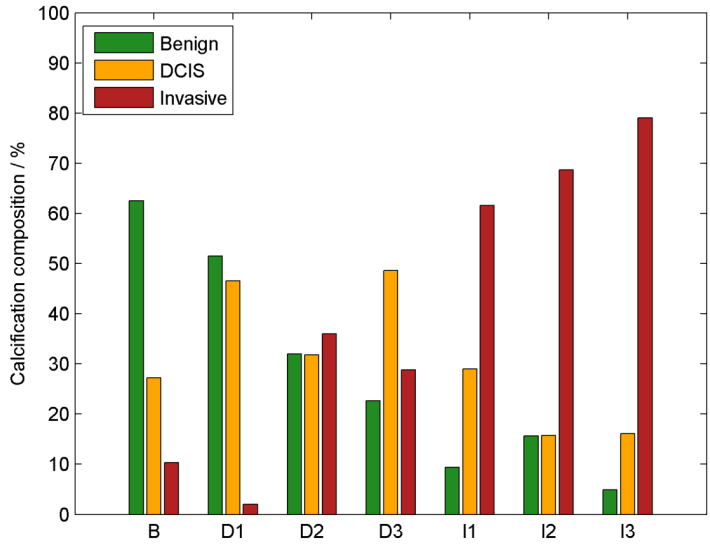


Figure 7.5 Histogram illustrating the mean composition of calcifications sorted according to pathology (B = benign, D = DCIS, I = invasive) and grade (1,2,3). This clearly indicates that the amount of invasive spectra increases with higher grade pathologies. On the other hand the average amount of benign spectra decreases with increasing pathology grade.

7.1.3.3 Transformation in calcification composition during disease progression

A highly interesting finding was that the imaging method allows observing the progression of calcifications from one pathology grade into the next higher one. It was found in 73% percent of images (72 infrared maps), that this transformation starts at the peripheral areas of calcifications. Example images are illustrated in figure 7.6, a comprehensive summary of all images sorted according to pathology grade is provided in Appendix A.

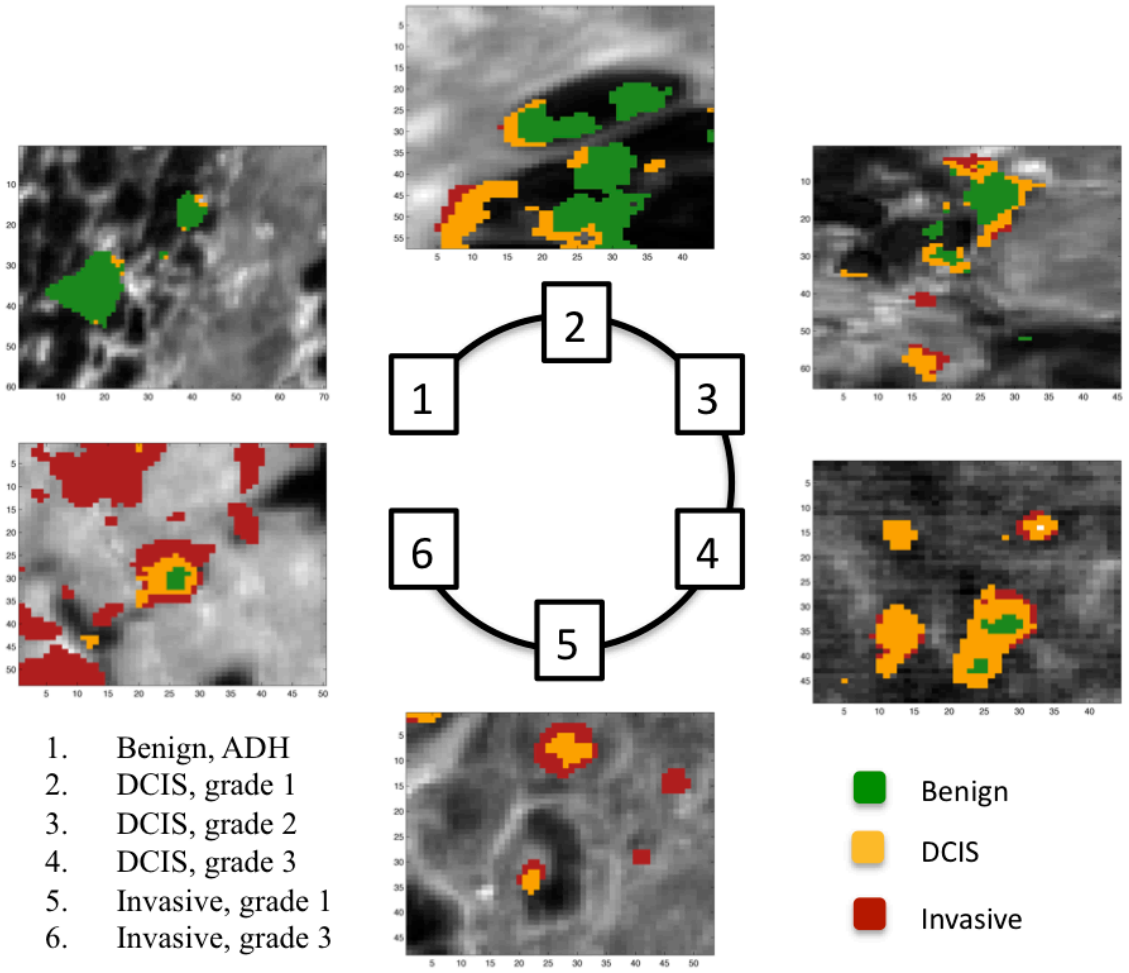


Figure 7.6 Sample illustration of transformation of calcifications during disease progression. This also shows that the transformation from one pathology grade into the next higher one starts at the peripheral areas of calcifications.

Invasive spectra show significantly higher absorption bands at wavenumbers 1549 and 1658cm^{-1}, representing amide II and amide I, than DCIS and benign spectra. This suggests the presence of a greater amount of proteins in the as red labeled invasive calcification areas. A possible explanation for this might be the presence of bone matrix proteins. Evidence of over expressed bone matrix proteins, including osteonectin (OSN), osteopontin (OPN) and bone sialoprotein (BSP), in human breast has been demonstrated in the past (Bellahcene *et al.*, 1995, Bellahcene *et al.*, 1994). Especially the three mentioned proteins have been extensively studied and it showed that they are involved in the onset of bone matrix (Young *et al.*, 1993). OSN features the ability to bind to calcium, hydroxyapetite and collagen I and is

typically localised in mineralised tissue, such as bone and dentine (Termine *et al.*, 1981a, Termine *et al.*, 1981b). The significantly increased amide content in invasive spectra as well as a notably higher collagen content (band 1286 cm^{-1}) in combination with the fact that type II calcifications consist of hydroxyapatite might indicate the presence of OSN and other bone matrix proteins. Regarding this, calcification areas shown as red (invasive) could be further interpreted as areas of actively ongoing mineralisation process due to an increased protein activity. For this reason image areas illustrated as red might not only indicate the presence of invasive cancer, they also might imply an active ongoing calcification process.

7.1.3.4 Diagnostic prediction

The resulting maps were further used for a diagnostic approach based on the identified calcifications in infrared maps. For each individual map the number of calcification spectra was estimated and subdivided into pathology groups. The final diagnostic prediction was estimated based on the most frequently observed pathology group. It was found that 70.1% of all benign maps, 44.2% of all DCIS maps and 84.1% of all invasive maps were correctly classified. As the confusion matrix presented in Table 7.2 shows 15 benign maps, 15 DCIS maps and 37 infrared maps were correctly predicted.

Table 7.2 Confusion matrix for classification results based on infrared maps

	Benign	**DCIS**	**Invasive**
Benign	15	6	0
DCIS	8	15	11
Invasive	6	1	37

Hence, for several patient samples multiple maps were measured. For this reason the predictive result for each individual patient samples was estimated. Under the circumstances that for one patient sample multiple maps were available the number of identified calcification spectra within all maps were summed. Based on the total amount of calcification spectra the final prediction was taken using a majority vote. It showed that 80.0% benign samples, 43.4% DCIS samples and 87.9% invasive samples were correctly classified. Overall the predictive accuracy for invasive samples and benign samples is very good. Nonetheless, the DCIS predictive rate is disappointing. Table 7.3 illustrates the confusion matrix established for the achieved results.

Table 7.3 Confusion matrix for classification results based on patient samples

	Benign	**DCIS**	**Invasive**
Benign	12	3	0
DCIS	5	10	8
Invasive	2	2	29

7.1.4 Conclusion

In this chapter the successful development of an imaging algorithm for visualisation of type II calcifications in breast tissue was presented. This method, utilising the ensemble SVM approach developed in the previous chapter, allows the generation of colour-coded images representing different breast disease pathologies. Furthermore, the resulting images revealed a deeper insight in the dynamics of breast calcification development. Thus, for the first time the visualisation of breast calcifications including the transformation of calcifications with increasing pathology grade has been reported.

8 Final remarks

8.1 General conclusion

The major objective of this work was to investigate new methods and algorithms for classification of vibrational spectroscopic data, including Raman and infrared data, derived from human tissue samples. The target was to achieve better performance than the LDA models which were so far commonly applied by the Biophotonics group at Gloucester. For this reason PLS-DA and SVMs have been investigated for their diagnostic potential to predict metastases in lymph nodes based on Raman spectroscopy and it was shown that SVMs are very strong classifiers yielding an unbeatable result of 100% correct prediction of an unseen data set and thus SVMs performed clearly better than LDA. Hence, the aim to develop classification models employing various machine learning methods in order to improve the predictive accuracy was achieved.

Employing a single SVM was not sufficient for the more complex multi-class infrared data set. A sophisticated ensemble SVM approach was taken, which successfully addressed this problem and resulted in significantly improved classification performance. Thus, for this data set is was possible to produce a classification model whose performance exceeded LDA models developed in pervious work. The resulting SVM ensemble was further extended for analysing infrared images. This novel approach revealed new insight into calcification and consequently is of great importance for gaining better understanding of breast cancer progression.

In summary, the objectives of this study on the optimisation of machine learning methods for cancer diagnostics using vibrational spectroscopy have been met. However, further work is required in order to make the combination of machine learning and vibrational spectroscopy

applicable as a routine clinical tool. Required future work towards clinical application is addressed in the next section.

8.2 Recommendations for future work

8.2.1 Diagnostic models for Raman spectroscopy

Extensive Raman mapping was carried out for generating the data used for the development of lymph node diagnostics. This type of measurement is too time consuming for an intra-operative application since it can take between four to 12 hours depending on the sample size. A possibility to overcome this is to execute point measurement over the sample surface. Thus, for instance gathering five to ten spectra evenly distributed over the sample surface can reduce the measurement time to nine to 18 minutes respectively. A full assessment of point measurements in combination with the developed models is required to see if lymph node pathology can still be predicted reliably.

An intra-operative assessment through point measurements is less time consuming but also can bear the risk of missing micrometastases. Consequently, a more extensive mapping in greater detail could be applied after intra-operative assessment for samples identified as negative. This would ensure that no micrometastases are present in the lymph node. For this purpose the SVM model developed in this work can be applied. However, some modifications are required for clinical routine application. First of all a visualisation of the results of the analysis of the whole map instead of extracted areas is required. The findings should then be presented in an image highlighting identified micrometastasis. This can be realised in a similar manner as the imaging approach employed for the IR breast cancer data.

In order to make this approach better applicable for future diagnostics it would be necessary to keep the overall analysis as simple as possible. Thus, once the actual Raman measurement is completed the data analysis should be executed automatically and provide a final output to the user. This output ideally contains an image highlighting metastasis, if present, as well a diagnostic result stating if a lymph node is cancerous or not. Finally, this result should be assigned with a confidence reflecting the possible inter-observer disagreement of histopathologists.

8.2.2 Diagnostic models for infrared spectroscopy

The ensemble classification model developed on manually selected calcifications performed exceptionally well, yet when applied for diagnosis of whole infrared maps this performance could not be maintained. One reason for this might be that assigned class is based on the findings of a single histopathologist. Thus, there is the chance that the staging is not consistent throughout the data set since it is known that the opinions of histopathologist on the different grades might vary (inter-observer disagreement). Consequently, a developed classification model can only be accurate if the classes are correctly assigned. In order to allow the development of an improved model it is suggested to get the opinions of several histopathologists, which would take the inter-observer disagreement into account. For instance the different opinions could then be integrated by setting a confidence on the output of the model.

A further possibility to improve the performance of the diagnostic model would be to use a regression approach instead of an absolute classification result. This would not only allow addressing the inter-observer disagreement it also enables more continuous staging and diagnostic result. So far the model only predicted diseases stages, benign, ductul carcinoma *in*

situ or invasive cancer. Applying a regression model would further allow diagnosis of different grades within the stages, especially since this work showed that there is a gradual change in chemical composition of calcifications correlated with disease progression.

The developed imaging algorithm enabled the visualisation of microcalcifications in infrared maps. Further improvement would include presenting the tissue sections, excluding the calcified areas, in a standardised colouring. Ideally this colour coding would be adapted to represent the different morphological features, such as ducts or epithelial cells. Additionally a final diagnostic result based on a regression model and including a confidence interval would be a step further towards the automated application of IR-spectroscopy for breast cancer diagnostics.

References

Adar, F., Lebourdon, G., Reffner, J. & Whitley, A. 2003. FT-IR and Raman Microscopy on a United Platform: A technology whos time has come. *Spectroscopy Magazine,* 18**,** 34-40.

Alfano, R. R., Liu, C. H., Shaw, W. L., Zhu, H. R., Akins, D. L., Cleary, J., Prudente, R. & Cellmer, E. 1991. Human breast tissues studied by IR Fourier transform Raman spectroscopy. *Lasers in the life sciences,* 4**,** 23-28.

Altman, D. G. & Bland, J. M. 1994. Diagnostic tests. 1: Sensitivity and specificity. *BMJ,* 308**,** 1552.

Arnaud, S., Houvenaeghel, G., Moutardier, V., Butarelli, M., Martino, M., Tallet, A., Braud, A. C., Jacquemier, J., Julian-Reynier, C. & Brenot-Rossi, I. 2004. Patients' and surgeons' perspectives on axillary surgery for breast cancer. *Eur J Surg Oncol,* 30**,** 735-43.

Babrah, J., Mccarthy, K., Lush, R. J., Rye, A. D., Bessant, C. & Stone, N. 2009. Fourier transform infrared spectroscopic studies of T-cell lymphoma, B-cell lymphoid and myeloid leukaemia cell lines. *Analyst,* 134**,** 763-8.

Baker, M. J., Gazi, E., Brown, M. D., Shanks, J. H., Gardner, P. & Clarke, N. W. 2008. FTIR-based spectroscopic analysis in the identification of clinically aggressive prostate cancer. *Br J Cancer,* 99**,** 1859-66.

Baker, R., Rogers, K. D., Shepherd, N. & Stone, N. 2010a. New relationships between breast microcalcifications and cancer. *Br J Cancer*.

Baker, R., Rogers, K. D., Shepherd, N. & Stone, N. 2010b. New relationships between breast microcalcifications and cancer. *British Journal of Cancer,* 103**,** 1034-1039.

Baker, R. N. 2009. *Spectroscopic analysis of breast tissue microcalcifications.* PhD, Cranfield University.

Beleites, C. & Salzer, R. 2008. Assessing and improving the stability of chemometric models in small sample size situations. *Anal Bioanal Chem,* 390**,** 1261-71.

Beleites, C., Steiner, G., Sowa, M. G., Baumgartner, R., Sobottka, S., Schackert, G. & Salzer, R. 2005. Classification of human gliomas by infrared imaging spectroscopy and chemometric image processing. *Vibrational Spectroscopy,* 38**,** 143-149.

Bellahcene, A. & Castronovo, V. 1995. Increased expression of osteonectin and osteopontin, two bone matrix proteins, in human breast cancer. *Am J Pathol,* 146**,** 95-100.

Bellahcene, A., Merville, M. P. & Castronovo, V. 1994. Expression of bone sialoprotein, a bone matrix protein, in human breast cancer. *Cancer Res,* 54**,** 2823-6.

Bhargava, R. & Levin, I. W. (eds.) 2003. *Recent Developments in Fourier Transform Infrared (FTIR) Microspectrospic Methods for Biomedical Analyses: From Single-Point Detection to Two-Dimensional Imaging,* Boca Raton, Fla. ; London: CRC Press.

Bigio, I. J. & Mourant, J. R. 1997. Ultraviolet and visible spectroscopies for tissue diagnostics: fluorescence spectroscopy and elastic-scattering spectroscopy. *Phys Med Biol,* 42**,** 803-14.

Bird, B., Miljkovic, M., Romeo, M. J., Smith, J., Stone, N., George, M. W. & Diem, M. 2008. Infrared micro-spectral imaging: distinction of tissue types in axillary lymph node histology. *BMC Clin Pathol,* 8**,** 8.

Bird, B., Romeo, M., Laver, N. & Diem, M. 2009. Spectral detection of micro-metastases in lymph node histo-pathology. *J Biophotonics,* 2**,** 37-46.

Blamey, R. W., Wilson, A. R. & Patnick, J. 2000. ABC of breast diseases: screening for breast cancer. *BMJ,* 321**,** 689-93.

Breiman, L. 1996. Bagging predictors. *Machine Learning,* 24**,** 123-140.

Breiman, L. 2001. Random forests. *Machine Learning,* 45**,** 5-32.

Brereton, R. G. 2007. *Applied chemometrics for scientists,* Hoboken, N.J., Wiley ; Chichester : John Wiley [distributor].

Brereton, R. G. 2009. *Chemometrics for pattern recognition,* Oxford, Wiley-Blackwell.

Burkitt, H. G., Young, B., Heath, J. W. & Wheater, P. R. 1993. *Wheater's functional histology : a text and colour atlas*, Edinburgh : Churchill Livinstone, 1993 (1994 [printing]).

Cancer Research Uk. 2010a. *Cancer in the UK : July 2010* [Online]. [Accessed].

Cancer Research Uk 2010b. CancerStats Key Facts breast cancer.

Cancer Research Uk 2010c. Latest UK Cancer Incidence and Mortality Summary - numbers.

Cao, D.-S., Xu, Q.-S., Liang, Y.-Z., Zhang, L.-X. & Li, H.-D. 2010. The boosting: A new idea of building models. *Chemometrics and Intelligent Laboratory Systems,* 100**,** 1-11.

Chan, J. W., Taylor, D. S., Zwerdling, T., Lane, S. M., Ihara, K. & Huser, T. 2006. Micro-Raman spectroscopy detects individual neoplastic and normal hematopoietic cells. *Biophys J,* 90**,** 648-56.

Chang, C.-C. & Lin, C.-J. 2001. LIBSVM : a library for support vector machines.

Chatterjee, S. J., Hawes, D., Taylor, C. R., Neville, A. M. & Cote, R. J. 2002. Occult Metastases. *In:* Silberman, H. & Silbermann, A. W. (eds.) *Surgical Oncology.*

Creager, A. J. & Geisinger, K. R. 2002. Intraoperative evaluation of sentinel lymph nodes for breast carcinoma: current methodologies. *Adv Anat Pathol,* 9**,** 233-43.

Crow, P., Molckovsky, A., Stone, N., Uff, J., Wilson, B. & Wongkeesong, L. M. 2005. Assessment of fiberoptic near-infrared raman spectroscopy for diagnosis of bladder and prostate cancer. *Urology,* 65**,** 1126-30.

Crow, P., Stone, N., Kendall, C. A., Persad, R. A. & Wright, M. P. 2003. Optical diagnostics in urology: current applications and future prospects. *BJU Int,* 92**,** 400-7.

Cserni, G., Bianchi, S., Boecker, W., Decker, T., Lacerda, M., Rank, F. & Wells, C. A. 2005. Improving the reproducibility of diagnosing micrometastases and isolated tumor cells. *Cancer,* 103**,** 358-67.

De Jong, B. W., Schut, T. C., Maquelin, K., Van Der Kwast, T., Bangma, C. H., Kok, D. J. & Puppels, G. J. 2006. Discrimination between nontumor bladder tissue and tumor by Raman spectroscopy. *Anal Chem,* 78**,** 7761-9.

De Veld, D. C., Witjes, M. J., Sterenborg, H. J. & Roodenburg, J. L. 2005. The status of in vivo autofluorescence spectroscopy and imaging for oral oncology. *Oral Oncol,* 41**,** 117-31.

Dhamelincourt, P. 2002. Raman Microscopy. *In:* Chalmers, J. & Griffiths, P. R. (eds.) *Handbook of Vibrational Spectroscopy.* Chischester: Wilez & Son.

Dietterich, T. G. 2000. Ensemble methods in machine learning. *In:* Kittler, J. & Roli, F. (eds.) *Multiple Classifier Systems.* Berlin: Springer-Verlag Berlin.

Domingos, P. & Pazzani, M. 1997. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning,* 29**,** 103-130.

Dukor, R. K., Liebman, M. N. & Johnson, B. L. 1998. A new, non-destructive method for analysis of clinical samples with FT-IR microspectroscopy. Breast cancer tissue as an example. *Cellular and Molecular Biology,* 44**,** 211-217.

Efron, B. 1979. Bootstrap methods: antoher look at the jackknife. *Annals of Statistics,* 7**,** 1-26.

Ellis, D. I. & Goodacre, R. 2006. Metabolic fingerprinting in disease diagnosis: biomedical applications of infrared and Raman spectroscopy. *Analyst,* 131**,** 875-85.

Erahimovitch, V., Talyshinsky, M., Souprun, Y. & Huleihel, M. 2006. FTIR spectroscopy examination of leukemia patients plasma. *Vibrational Spectroscopy,* 40**,** 40-46.

Fabian, H., Lasch, P., Boese, M. & Haensch, W. 2003. Infrared microspectroscopic imaging of benign breast tumor tissue sections. *Journal of Molecular Structure,* 661**,** 411-417.

Fabian, H., Thi, N. A., Eiden, M., Lasch, P., Schmitt, J. & Naumann, D. 2006. Diagnosing benign and malignant lesions in breast tissue sections by using IR-microspectroscopy. *Biochim Biophys Acta,* 1758**,** 874-82.

Ferlay J, Bray F, Pisani P & Dm, P. 2004. GLOBOCAN 2002: Cancer Incidence, Mortality and Prevalence Worldwide. Lyon, France: IARC CancerBase No.5.

Ferlay, J., Autier, P., Boniol, M., Heanue, M., Colombet, M. & Boyle, P. 2007. Estimates of the cancer incidence and mortality in Europe in 2006. *Ann Oncol,* 18**,** 581-92.

Ferraro, J. R., Nakamoto, K. & Brown, C. W. 2003. *Introductory Raman spectroscopy,* Amsterdam ; London, Academic Press.

Fielding, A. 2007. *Cluster and classification techniques for the biosciences,* Cambridge, Cambridge University Press.

Freund, Y. & Schapire, R. E. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences,* 55**,** 119-139.

Gazi, E., Baker, M., Dwyer, J., Lockyer, N. P., Gardner, P., Shanks, J. H., Reeve, R. S., Hart, C. A., Clarke, N. W. & Brown, M. D. 2006. A correlation of FTIR spectra derived from prostate cancer biopsies with gleason grade and tumour stage. *Eur Urol,* 50**,** 750-60; discussion 760-1.

Grabau, D. A., Rank, F. & Friis, E. 2005. Intraoperative frozen section examination of axillary sentinel lymph nodes in breast cancer. *APMIS,* 113**,** 7-12.

Green, M., Ekelund, U., Edenbrandt, L., Bjork, J., Forberg, J. L. & Ohlsson, M. 2009. Exploring new possibilities for case-based explanation of artificial neural network ensembles. *Neural Netw,* 22**,** 75-81.

Grimbergen, M. C., Van Swol, C. F., Van Moorselaar, R. J., Uff, J., Mahadevan-Jansen, A. & Stone, N. 2009. Raman spectroscopy of bladder tissue in the presence of 5-aminolevulinic acid. *Journal of Photochemistry and Photobiology B,* 95**,** 170-6.

Günzler, H. & Gremlich, H.-U. 2002. *IR spectroscopy : an introduction,* Weinheim ; Cambridge, Wiley-VCH.

Haka, A. S., Shafer-Peltier, K. E., Fitzmaurice, M., Crowe, J., Dasari, R. R. & Feld, M. S. 2002. Identifying microcalcifications in benign and malignant breast lesions by probing differences in their chemical composition using Raman spectroscopy. *Cancer Research,* 62**,** 5375-5380.

Haka, A. S., Shafer-Peltier, K. E., Fitzmaurice, M., Crowe, J., Dasari, R. R. & Feld, M. S. 2005. Diagnosing breast cancer by using Raman spectroscopy. *Proc Natl Acad Sci U S A,* 102**,** 12371-6.

Haka, A. S., Volynskaya, Z., Gardecki, J. A., Nazemi, J., Shenk, R., Wang, N., Dasari, R. R., Fitzmaurice, M. & Feld, M. S. 2009. Diagnosing breast cancer using Raman spectroscopy: prospective analysis. *Journal of Biomedical Optics,* 14.

Hand, D. J. & Henley, W. E. 1997. Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society, Series B,* 160**,** 523-541.

Hansen, L. K. & Salamon, P. 1990. Neural Network Ensembles. *Ieee Transactions on Pattern Analysis and Machine Intelligence,* 12**,** 993-1001.

Hirschfeld, T. & Chase, D. B. 1986. FT-Raman spectroscopy: development and justification. *Applied Spectroscopy,* 40**,** 133-137.

Hollas, J. M. 2002. *Basic atomic and molecular spectroscopy,* Cambridge, Royal Society of Chemistry.

Horsnell, J., Stonelake, P., Christie-Brown, J., Shetty, G., Hutchings, J., Kendall, C. & Stone, N. 2010. Raman spectroscopy--a new method for the intra-operative assessment of axillary lymph nodes. *Analyst,* 135**,** 3042-7.

Hsu, C.-W., Chang, C.-C. & Lin, C.-J. 2008. A Practical guide to Support Vector Classification.

Huang, Z., Mcwilliams, A., Lui, H., Mclean, D. I., Lam, S. & Zeng, H. 2003. Near-infrared Raman spectroscopy for optical diagnosis of lung cancer. *Int J Cancer,* 107**,** 1047-52.

Ioachim, H. L., Medeiros, L. J. & Ioachim, H. L. I. S. L. N. P. 2009. *Ioachim's lymph node pathology,* Philadelphia, Pa. ; London, Lippincott Williams & Wilkins.

Isabelle, M., Stone, N., Barr, H., Vipond, M., Shepherd, N. & Rogers, K. 2008. Lymph node pathology using optical spectroscopy in cancer diagnostics. *Spectroscopy,* 22**,** 97-104.

Izenman, A. J. 2008. *Modern multivariate statistical techniques : regression, classification, and manifold learning,* New York ; London, Springer.

Jackson, M. & Mantsch, H. H. 2000. *Infrared Spectroscopy, Ex Vivo Tissue Analysis,* Chichester, John Wiley.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J. & Hinton, G. E. 1991. Adaptive Mixtures of Local Experts. *Neural Computation,* 3**,** 79-87.

Jerjes, W., Swinson, B., Johnson, K. S., Thomas, G. J. & Hopper, C. 2005. Assessment of bony resection margins in oral cancer using elastic scattering spectroscopy: a study on archival material. *Arch Oral Biol,* 50**,** 361-6.

Kawabata, T., Mizuno, T., Okazaki, S., Hiramatsu, M., Setoguchi, T., Kikuchi, H., Yamamoto, M., Hiramatsu, Y., Kondo, K., Baba, M., Ohta, M., Kamiya, K., Tanaka, T., Suzuki, S. & Konno, H. 2008. Optical diagnosis of gastric cancer using near-infrared multichannel Raman spectroscopy with a 1064-nm excitation wavelength. *J Gastroenterol,* 43**,** 283-90.

Kendall, C., Day, J., Hutchings, J., Smith, B., Shepherd, N., Barr, H. & Stone, N. 2010. Evaluation of Raman probe for oesophageal cancer diagnostics. *Analyst,* 135**,** 3038-3041.

Kendall, C., Isabelle, M., Bazant-Hegemark, F., Hutchings, J., Orr, L., Babrah, J., Baker, R. & Stone, N. 2009. Vibrational spectroscopy: a clinical tool for cancer diagnostics. *Analyst,* 134**,** 1029-45.

Kendall, C., Stone, N., Shepherd, N., Geboes, K., Warren, B., Bennett, R. & Barr, H. 2003. Raman spectroscopy, a potential tool for the objective identification and classification of neoplasia in Barrett's oesophagus. *J Pathol,* 200**,** 602-9.

Koljenovic, S., Bakker Schut, T. C., Van Meerbeeck, J. P., Maat, A. P., Burgers, S. A., Zondervan, P. E., Kros, J. M. & Puppels, G. J. 2004. Raman microspectroscopic mapping studies of human bronchial tissue. *J Biomed Opt,* 9**,** 1187-97.

Koljenovic, S., Schut, T. B., Vincent, A., Kros, J. M. & Puppels, G. J. 2005. Detection of meningioma in dura mater by Raman spectroscopy. *Anal Chem,* 77**,** 7958-65.

Krafft, C., Codrich, D., Pelizzo, G. & Sergo, V. 2008. Raman and FTIR microscopic imaging of colon tissue: a comparative study. *J Biophotonics,* 1**,** 154-69.

Krafft, C., Ramoji, A. A., Bielecki, C., Vogler, N., Meyer, T., Akimov, D., Rosch, P., Schmitt, M., Dietzek, B., Petersen, I., Stallmach, A. & Popp, J. 2009. A comparative Raman and CARS imaging study of colon tissue. *J Biophotonics,* 2**,** 303-12.

Krishnaa, C. M., Prathimaa, N. B., Malinia, R., Vadhirajab, B. M., Bhattc, R. A., Fernandesb, D. J., Kushtagic, P., Vidyasagarb, M. S. & Karthaa, V. B. 2006. Raman spectroscopy studies for diagnosis of cancers in human uterine cervix. *Vibrational Spectroscopy,* 41**,** 136-141.

Kumar, V., Abbas, A. K., Fausto, N., Robbins, S. L. & Cotran, R. S. R. P. B. O. D. 2005. *Robbins and Cotran pathologic basis of disease,* Philadelphia, Pa. ; [London], Elsevier Saunders.

Kuncheva, L. I. 2004. *Combining pattern classifiers : methods and algorithms,* Hoboken, N.J., Wiley-Interscience.

Lasch, P., Haensch, W., Naumann, D. & Diem, M. 2004. Imaging of colorectal adenocarcinoma using FT-IR microspectroscopy and cluster analysis. *Biochim Biophys Acta,* 1688**,** 176-86.

Lasch, P. & Naumann, D. 1998. FT-IR microspectroscopic imaging of human carcinoma thin sections based on pattern recognition techniques. *Cell Mol Biol (Noisy-le-grand),* 44**,** 189-202.

Laserna, J. 2001. *An indroduction to Raman spectroscopy: introduction and basic principels* [Online]. Wiley. Available: www.spectroscopyNow.com [Accessed 2nd June 2008].

Lavine, B. K. (ed.) 2000. *Clustering and Classification of Analytical Data,* Chichester: John Wiley.

Li, Q. B., Sun, X. J., Xu, Y. Z., Yang, L. M., Zhang, Y. F., Weng, S. F., Shi, J. S. & Wu, J. G. 2005. Diagnosis of gastric inflammation and malignancy in endoscopic biopsies based on Fourier transform infrared spectroscopy. *Clin Chem,* 51**,** 346-50.

Lieber, C. A., Majumder, S. K., Billheimer, D., Ellis, D. L. & Mahadevan-Jansen, A. 2008a. Raman microspectroscopy for skin cancer detection in vitro. *J Biomed Opt,* 13**,** 024013.

Lieber, C. A., Majumder, S. K., Ellis, D. L., Billheimer, D. D. & Mahadevan-Jansen, A. 2008b. In vivo nonmelanoma skin cancer diagnosis using Raman microspectroscopy. *Lasers in Surgery and Medicine,* 40**,** 461-467.

Llora, X., Priya, A. & Bhargava, R. 2009. Observer-invariant histopathology using genetics-based machine learning. *Natural Computing,* 8**,** 101-120.

Lovat, L. B., Johnson, K., Mackenzie, G. D., Clark, B. R., Novelli, M. R., Davies, S., O'donovan, M., Selvasekar, C., Thorpe, S. M., Pickard, D., Fitzgerald, R., Fearn, T., Bigio, I. & Bown, S. G. 2006. Elastic scattering spectroscopy accurately detects high grade dysplasia and cancer in Barrett's oesophagus. *Gut,* 55**,** 1078-83.

Ly, E., Cardot-Leccia, N., Ortonne, J. P., Benchetrit, M., Michiels, J. F., Manfait, M. & Piot, O. 2010. Histopathological characterization of primary cutaneous melanoma using infrared microimaging: a proof-of-concept study. *Br J Dermatol*.

Ly, E., Piot, O., Wolthuis, R., Durlach, A., Bernard, P. & Manfait, M. 2008. Combination of FTIR spectral imaging and chemometrics for tumour detection from paraffin-embedded biopsies. *Analyst,* 133**,** 197-205.

Magee, N. D., Beattie, J. R., Carland, C., Davis, R., Mcmanus, K., Bradbury, I., Fennell, D. A., Hamilton, P. W., Ennis, M., Mcgarvey, J. J. & Elborn, J. S. 2010. Raman microscopy in the diagnosis and prognosis of surgically resected nonsmall cell lung cancer. *J Biomed Opt,* 15**,** 026015.

Magee, N. D., Villaumie, J. S., Marple, E. T., Ennis, M., Elborn, J. S. & Mcgarvey, J. J. 2009. Ex vivo diagnosis of lung cancer using a Raman miniprobe. *J Phys Chem B,* 113**,** 8137-41.

Mahadevan, A., Mitchell, M. F., Silva, E., Thomsen, S. & Richards-Kortum, R. R. 1993. Study of the fluorescence properties of normal and neoplastic human cervical tissue. *Lasers Surg Med,* 13**,** 647-55.

Mahadevan-Jansen, A. 2003. Raman Spectroscopy: From Benchtop to Bedside. *In:* Vo-Dinh, T. (ed.) *Biomedical photonics handbook.* Boca Raton, Fla. ; London: CRC Press.

Mahadevan-Jansen, A. & Richards-Kortum, R. 1996. Raman spectroscopy for the detection of cancers and precancers. *Journal of Biomedical Optics,* 1**,** 31-70.

Mahadevan-Jansen, A. & Richards-Kortum, R. 1997. Raman spectroscopy for cancer detection: a review. *In:* Ieee/Embs (ed.) *Proceedings.* Chicago, IL. USA.

Marchesini, R., Cascinelli, N., Brambilla, M., Clemente, C., Mascheroni, L., Pignoli, E., Testori, A. & Venturoli, D. R. 1992. In vivo spectrophotometric evaluation of neoplastic and non-neoplastic skin pigmented lesions. II: Discriminant analysis between nevus and melanoma. *Photochem Photobiol,* 55**,** 515-22.

Matousek, P. & Stone, N. 2007. Prospects for the diagnosis of breast cancer by noninvasive probing of calcifications using transmission Raman spectroscopy. *J Biomed Opt,* 12**,** 024008.

Mayinger, B., Jordan, M., Horner, P., Gerlach, C., Muehldorfer, S., Bittorf, B. R., Matzel, K. E., Hohenberger, W., Hahn, E. G. & Guenther, K. 2003. Endoscopic light-induced autofluorescence spectroscopy for the diagnosis of colorectal cancer and adenoma. *J Photochem Photobiol B,* 70**,** 13-20.

Mccreery, R. L. 2000. *Raman spectroscopy for chemical analysis,* New York ; Chichester, Wiley.

Mckevly, M. L. 2000. Infrared Spectroscopy: Introduction. *In:* Meyers, R. A. (ed.) *Encyclopedia of analytical chemistry : applications, theory and instrumentation.* Chichester: John Wiley.

Meier, R. J. 2005. Vibrational spectroscopy: a 'vanishing' discipline? *Chem Soc Rev,* 34**,** 743-52.

Milgram, J., Cheriet, M. & Sabourin, R. 2006. “One against one” or “One against all”: which one is better for handwriting recognition with SVMs? *10th International Workshop on Frontiers in Handwriting Recognition*

Mo, J., Zheng, W., Low, J. J., Ng, J., Ilancheran, A. & Huang, Z. 2009. High wavenumber Raman spectroscopy for in vivo detection of cervical dysplasia. *Anal Chem,* 81**,** 8908-15.

Montgomery, E., Bronner, M. P., Goldblum, J. R., Greenson, J. K., Haber, M. M., Hart, J., Lamps, L. W., Lauwers, G. Y., Lazenby, A. J., Lewin, D. N., Robert, M. E., Toledano, A. Y., Shyr, Y. & Washington, K. 2001. Reproducibility of the diagnosis of dysplasia in Barrett esophagus: a reaffirmation. *Hum Pathol,* 32**,** 368-78.

Moreno, M., Raniero, L., Loschiavo Arisawa, E. A., Do Espirito Santo, A. M., Pereira Dos Santos, E. A., Bitar, R. A. & Martin, A. A. 2010. Raman spectroscopy study of breast disease. *Theoretical Chemistry Accounts,* 125**,** 329-334.

Morgan, M. P., Cooke, M. M., Christopherson, P. A., Westfall, P. R. & Mccarthy, G. M. 2001. Calcium hydroxyapatite promotes mitogenesis and matrix metalloproteinase expression in human breast cancer cell lines. *Mol Carcinog,* 32**,** 111-7.

Morgan, M. P., Cooke, M. M. & Mccarthy, G. M. 2005. Microcalcifications associated with breast cancer: an epiphenomenon or biologically significant feature of selected tumors? *J Mammary Gland Biol Neoplasia,* 10**,** 181-7.

Morton, D. L., Wen, D. R., Wong, J. H., Economou, J. S., Cagle, L. A., Storm, F. K., Foshag, L. J. & Cochran, A. J. 1992. Technical details of intraoperative lymphatic mapping for early stage melanoma. *Arch Surg,* 127**,** 392-9.

Mosteller, F. & Tukey, J. W. 1977. *Data analysis and regression : a second course in statistics,* Reading, Mass. ; London, Addison-Wesley.

Mourant, J. R. & Bigio, I. J. 2003. Elastic-Scattering Spectroscopy and Diffuse Reflectance. *In:* Vo-Dinh, T. (ed.) *Biomedical photonics handbook.* Boca Raton, Fla. ; London: CRC Press.

Mourant, J. R., Bigio, I. J., Boyer, J., Conn, R. L., Johnson, T. & Shimada, T. 1995. Spectroscopic diagnosis of bladder cancer with elastic light scattering. *Lasers Surg Med,* 17**,** 350-7.

Mourant, J. R., Bocklage, T. J., Powers, T. M., Greene, H. M., Bullock, K. L., Marr-Lyon, L. R., Dorin, M. H., Waxman, A. G., Zsemlye, M. M. & Smith, H. O. 2007. In vivo light scattering measurements for detection of precancerous conditions of the cervix. *Gynecol Oncol,* 105**,** 439-45.

Nijssen, A., Bakker Schut, T. C., Heule, F., Caspers, P. J., Hayes, D. P., Neumann, M. H. & Puppels, G. J. 2002. Discriminating basal cell carcinoma from its surrounding tissue by Raman spectroscopy. *J Invest Dermatol,* 119**,** 64-9.

Njoroge, E., Alty, S. R., Gani, M. R. & Alkatib, M. 2006. Classification of cervical cancer cells using FTIR data. *Conf Proc IEEE Eng Med Biol Soc,* 1**,** 5338-41.

Norgaard, L., Sölétormos, G., Harrit, N., Albrechtsen, M., Olsen, O., Nielsen, D., Kampmann, K. & Bro, R. 2007. Fluorescence spectroscopy and chemometrics for classification of breast cancer samples-a feasibility study using extended canonical variates analysis. *Journal of Chemometrics,* 21**,** 451-458.

Oesterling, J. E. 1991. Prostate specific antigen: a critical assessment of the most useful tumor marker for adenocarcinoma of the prostate. *J Urol,* 145**,** 907-23.

Orr, L. E., Christie-Brown, J., Hutchings, J. C., Mccarthy, K., Rose, S., Thomas, M. & Stone, N. 2010. Raman spectroscopy as a tool for the identification and differentiation of neoplasias contained within lymph nodes of the head and neck. *Photonic Therapeutics and Diagnostics VI.* San Francisco, California, USA Proceedings of the SPIE.

Park, S. C., Lee, S. J., Namkung, H., Chung, H., Han, S.-H., Yoon, M.-Y., Park, J.-J., Lee, J.-H., Oh, C.-H. & Woo, Y.-A. 2007. Feasibility study for diagnosis of stomach adenoma and cancer using IR spectroscopy. *Vibrational Spectroscopy,* 44**,** 279-285.

Petry, R., Schmitt, M. & Popp, J. 2003. Raman spectroscopy--a prospective tool in the life sciences. *Chemphyschem,* 4**,** 14-30.

Phan, J., Moffitt, R., Dale, J., Petros, J., Young, A. & Wang, M. 2005. Improvement of SVM Algorithm for Microarray Analysis Using Intelligent Parameter Selection. *Conf Proc IEEE Eng Med Biol Soc,* 5**,** 4838-41.

Pistorius, A. M. A. 1995. Biochemical applications of FT-IR spectroscopy. *Spectroscopy Europe,* 7**,** 8-15.

Polikar, R. 2006. Ensemble based systems in decision making. *Circuits and Systems Magazine, IEEE,* 6.

Popp, J. & Kiefer, W. 2003. Raman Scaterring, Fundamentals. *In:* Meyers, R. A. (ed.) *Encyclopedia of analytical chemistry : applications, theory and instrumentation.* Chichester: John Wiley.

Prierce, G. B. & Damjanov, I. 2006. The pathology of cancer. *In:* Mckinnell, R. G., Parchment, R. E., Perantoni, A. O., Damjanov, I. & Prierce, G. B. (eds.) *The biological basis of cancer.* 2nd ed. ed. Cambridge: Cambridge University Press.

Quaroni, L. & Casson, A. G. 2009. Characterization of Barrett esophagus and esophageal adenocarcinoma by Fourier-transform infrared microscopy. *Analyst,* 134**,** 1240-6.

Rabah, R., Webera, R., Serhatkulua, G. K., Caoa, A., Daia, H., Pandyaa, A., Naika, R., Aunera, G., Poulika, J. & Klein, M. 2007. Diagnosis of neuroblastoma and ganglioneuroma using Raman spectroscopy. *Journal of Pediatric Surgery,* 43**,** 171-176.

Raman, C. V. & Krishnan, K. S. 1928. The Optical Analogue of the Compton Effect. *Nature,* 121**,** 711.

Ramanujam, N. 2000. Fluorescence Spectroscopy In Vivo. *In:* Meyers, R. A. (ed.) *Encyclopedia of analytical chemistry : applications, theory and instrumentation.* Chichester: John Wiley.

Romano, S., Balzi, M., Dei, L., Lerman, A. A., Neuberger, W. & Becciolini, A. 1995. Differences in sugar phosphate bands between normal bladder mucosa and tumoral tissue detected by Fourier transform infrared (FTIR) microreflectance spectroscopy. *Optical and Imaging Techniques in Biomedicine* Lille, France SPIE

Salem, A. A., Douglas-Jones, A. G., Sweetland, H. M. & Mansel, R. E. 2003. Intraoperative evaluation of axillary sentinel lymph nodes using touch imprint cytology and immunohistochemistry: I. Protocol of rapid immunostaining of touch imprints. *Eur J Surg Oncol,* 29**,** 25-8.

Salem, A. A., Douglas-Jones, A. G., Sweetland, H. M. & Mansel, R. E. 2006. Intraoperative evaluation of axillary sentinel lymph nodes using touch imprint cytology and immunohistochemistry. Part II. Results. *Eur J Surg Oncol,* 32**,** 484-7.

Schantz, S. P., Kolli, V., Savage, H. E., Yu, G., Shah, J. P., Harris, D. E., Katz, A., Alfano, R. R. & Huvos, A. G. 1998. In vivo native cellular fluorescence and histological characteristics of head and neck cancer. *Clin Cancer Res,* 4**,** 1177-82.

Schapire, R. E. 1990. The strength of weak learnability *Machine Learning,* 5**,** 197-227.

Schwenker, F. Year. Hierarchical support vector machines for multi-class pattern recognition. *In:* Howlett, R. J. J. L. C., ed. Kes'2000: Fourth International Conference on Knowledge-Based Intelligent Engineering Systems & Allied Technologies, Vols 1 and 2, Proceedings, 2000. 561-565.

Shafer-Peltier, K. E., Haka, A. S., Fitzmaurice, M., Crowe, J., Myles, J., Dasari, R. R. & Feld, M. S. 2002. Raman microspectroscopic model of human breast tissue: implications for breast cancer diagnosis in vivo. *Journal of Raman Sepctroscopy,* 33**,** 552-563.

Shaw, R. A. & Mantsch, H. H. 2000. Infrared Spectroscopy in Clinical and Diagnostic Analysis. *In:* Meyers, R. A. (ed.) *Encyclopedia of analytical chemistry : applications, theory and instrumentation.* Chichester: John Wiley.

Sigurdsson, S., Philipsen, P. A., Hansen, L. K., Larsen, J., Gniadecka, M. & Wulf, H. C. 2004. Detection of skin cancer by classification of Raman spectra. *IEEE Trans Biomed Eng,* 51**,** 1784-93.

Skoog, D. A., Holler, F. J. & Nieman, T. A. 1998. *Principles of instrumental analysis,* Philadelphia ; London, Saunders College Publishing.

Smekal, A. 1923. The quantum theory of dispersion *Naturwissenschaften,* 11**,** 873-875.

Smith, E. & Dent, G. 2005. *Modern Raman spectroscopy : a practical approach,* Chichester, John Wiley.

Smith, J. A. 2005. *Raman spectroscopy in the assessment of axillary lymph nodes in breast cancer* DM, Cranfield University.

Sobottka, S. B., Geiger, K. D., Salzer, R., Schackert, G. & Krafft, C. 2009. Suitability of infrared spectroscopic imaging as an intraoperative tool in cerebral glioma surgery. *Analytical and Bioanalytical Chemistry,* 393**,** 187-195.

Steiner, G., Shaw, A., Choo-Smith, L. P., Abuid, M. H., Schackert, G., Sobottka, S., Steller, W., Salzer, R. & Mantsch, H. H. 2003. Distinguishing and grading human gliomas by IR spectroscopy. *Biopolymers,* 72**,** 464-71.

Steller, W., Einenkel, J., Horn, L. C., Braumann, U. D., Binder, H., Salzer, R. & Krafft, C. 2006. Delimitation of squamous cell cervical carcinoma using infrared microspectroscopic imaging. *Anal Bioanal Chem,* 384**,** 145-54.

Stone, M. 1974. Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society, Series B,* 36**,** 111-147.

Stone, N., Baker, R., Rogers, K., Parker, A. W. & Matousek, P. 2007. Subsurface probing of calcifications with spatially offset Raman spectroscopy (SORS): future possibilities for the diagnosis of breast cancer. *Analyst,* 132**,** 899-905.

Stone, N., Kendall, C., Shepherd, N., Crow, P. & Barr, H. 2002. Near-infrared Raman spectroscopy for the classification of epithelial pre-cancers and cancers. *Journal of Raman Spectroscopy,* 33**,** 564-573.

Stone, N., Kendall, C., Smith, J., Crow, P. & Barr, H. 2004. Raman spectroscopy for identification of epithelial cancers. *Faraday Discuss,* 126**,** 141-57; discussion 169-83.

Stratton, M. R. & Rahman, N. 2008. The emerging landscape of breast cancer susceptibility. *Nat Genet,* 40**,** 17-22.

Sun, Y. B., Zeng, X. R., Wenger, L. & Cheung, H. S. 2003. Basic calcium phosphate crystals stimulate the endocytotic activity of cells-inhibition by anti-calcification agents. *Biochemical and Biophysical Research Communications,* 312**,** 1053-1059.

Teh, S. K., Zheng, W., Ho, K. Y., Teh, M., Yeoh, K. G. & Huang, Z. 2008a. Diagnosis of gastric cancer using near-infrared Raman spectroscopy and classification and regression tree techniques. *J Biomed Opt,* 13**,** 034013.

Teh, S. K., Zheng, W., Ho, K. Y., Teh, M., Yeoh, K. G. & Huang, Z. 2008b. Diagnostic potential of near-infrared Raman spectroscopy in the stomach: differentiating dysplasia from normal tissue. *Br J Cancer,* 98**,** 457-65.

Teh, S. K., Zheng, W., Ho, K. Y., Teh, M., Yeoh, K. G. & Huang, Z. 2010. Near-infrared Raman spectroscopy for early diagnosis and typing of adenocarcinoma in the stomach. *Br J Surg,* 97**,** 550-7.

Termine, J. D., Belcourt, A. B., Conn, K. M. & Kleinman, H. K. 1981a. MINERAL AND COLLAGEN-BINDING PROTEINS OF FETAL CALF BONE. *Journal of Biological Chemistry,* 256**,** 403-408.

Termine, J. D., Kleinman, H. K., Whitson, S. W., Conn, K. M., Mcgarvey, M. L. & Martin, G. R. 1981b. OSTEONECTIN, A BONE-SPECIFIC PROTEIN LINKING MINERAL TO COLLAGEN. *Cell,* 26**,** 99-105.

Tew, K., Irwig, L., Matthews, A., Crowe, P. & Macaskill, P. 2005. Meta-analysis of sentinel node imprint cytology in breast cancer. *Br J Surg,* 92**,** 1068-80.

Tfayli, A., Piot, O., Durlach, A., Bernard, P. & Manfait, M. 2005. Discriminating nevus and melanoma on paraffin-embedded skin biopsies using FTIR microspectroscopy. *Biochim Biophys Acta,* 1724**,** 262-9.

Titterington, D. M., Murray, G. D., Murray, L. S., Spiegelhalter, D. J., Skene, A. M., Habbema, J. D. F. & Gelpke, G. J. 1981. Comparison of discrimination techniques applied to a complex data set of head injured patients. *Journal of the Royal Statistical Society Series a-Statistics in Society,* 144**,** 145-175.

Tomellini, S. A. & Finn, J. W. (eds.) 2000. *Vibrational Spectroscopy in Drug Discovery, Development and Production,* Chichester: John Wiley.

Tse, G. M., Tan, P. H., Cheung, H. S., Chu, W. C. & Lam, W. W. 2008. Intermediate to highly suspicious calcification in breast lesions: a radio-pathologic correlation. *Breast Cancer Res Treat,* 110**,** 1-7.

Valiant, L. G. 1984. A theory of the learnable *Communications of the Acm,* 27**,** 1134-1142.

Vapnik, V. 1995. *The nature of statistical learning theory,* New York ; London, Springer.

Vapnik, V. N. 1998. *Statistical learning theory,* New York ; Chichester, Wiley.

Vo-Dinh, T. 2003. Biomedical Photonics: A Revolution at the Interface of Science and Technology. *In:* Vo-Dinh, T. (ed.) *Biomedical photonics handbook.* Boca Raton, Fla.; London: CRC Press.

Vo-Dinh, T. & Cullum, B. M. 2003. Flourescence Spectroscopy for Biomedical Diagnostics. *In:* Vo-Dinh, T. (ed.) *Biomedical photonics handbook.* Boca Raton, Fla. ; London: CRC Press.

Wang, T. D., Triadafilopoulos, G., Crawford, J. M., Dixon, L. R., Bhandari, T., Sahbaie, P., Friedland, S., Soetikno, R. & Contag, C. H. 2007a. Detection of endogenous biomolecules in Barrett's esophagus by Fourier transform infrared spectroscopy. *Proc Natl Acad Sci U S A,* 104**,** 15864-9.

Wang, X.-Y., Garibaldi, J. M., Bird, B. & George, M. W. 2007b. A novel fuzzy clustering algorithm for the analysis of axillary lymph node tissue sections. *Applied Intelligence,* 27**,** 237-248.

Wang, Y. & Mccreery, R. L. 1989. Evaluation of a diode laser/charge coupled device spectrometer for near-infrared Raman spectroscopy. *Analytical Chemistry,* 61**,** 2647-2651.

Wartewig, S. & Neubert, R. H. 2005. Pharmaceutical applications of Mid-IR and Raman spectroscopy. *Adv Drug Deliv Rev,* 57**,** 1144-70.

Weissleder, R. & Pittet, M. J. 2008. Imaging in the era of molecular oncology. *Nature,* 452**,** 580-9.

Widjaja, E., Zheng, W. & Huang, Z. 2008. Classification of colonic tissues using near-infrared Raman spectroscopy and support vector machines. *Int J Oncol,* 32**,** 653-62.

Wills, H., Kast, R., Stewart, C., Rabah, R., Pandya, A., Poulik, J., Auner, G. & Klein, M. D. 2009. Raman spectroscopy detects and distinguishes neuroblastoma and related tissues in fresh and (banked) frozen specimens. *J Pediatr Surg,* 44**,** 386-91.

Wolpert, D. H. 1992. Stacked generalization. *Neural Networks,* 5**,** 241-259.

Wongravee, K., Lloyd, G. R., Hall, J., Holmboe, M. E., Schaefer, M. L., Reed, R. R., Trevejo, J. & Brereton, R. G. 2009. Monte-Carlo methods for determining optimal number of significant variables. Application to mouse urinary profiles *Metabolomics,* 5**,** 387-406.

Wood, B. R., Chiriboga, L., Yee, H., Quinn, M. A., Mcnaughton, D. & Diem, M. 2004. Fourier transform infrared (FTIR) spectral mapping of the cervical transformation zone, and dysplastic squamous epithelium. *Gynecol Oncol,* 93**,** 59-68.

Yano, K., Ohoshima, S., Gotou, Y., Kumaido, K., Moriguchi, T. & Katayama, H. 2000. Direct measurement of human lung cancerous and noncancerous tissues by fourier transform infrared microscopy: can an infrared microscope be used as a clinical tool? *Anal Biochem,* 287**,** 218-25.

Young, M. F., Ibaraki, K., Kerr, J. M. & Heegaard, A.-M. 1993. *Molecular and cellular biology of the major noncollagenous proteins in bone*.

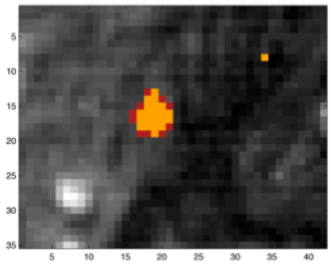
Zeng, H., Mcwillams, A. & Lam, S. 2004. Optical spectroscopy and imaging for early lung cancer detection: a review. *Photodiagnostics and Photodynamic Therapy,* 1**,** 111-122.

Appendix

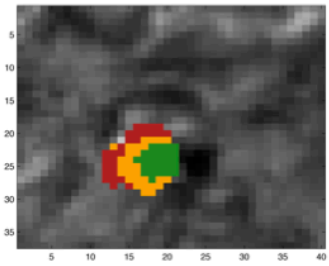
Appendix A

Benign Samples:

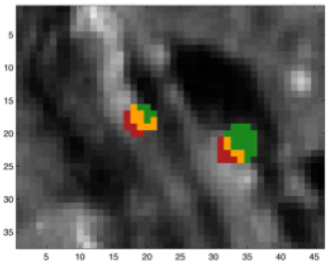
Sample 22



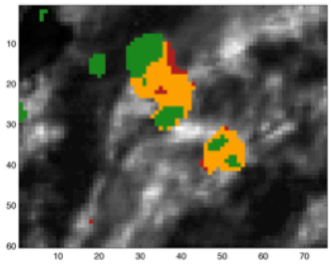
Sample 24, map 1



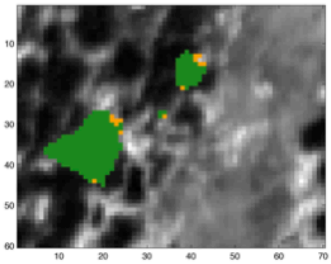
Sample 24, map 2



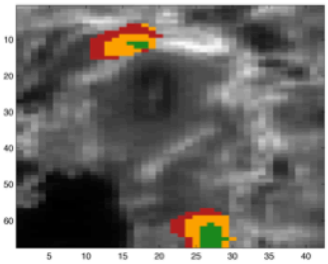
Sample 25



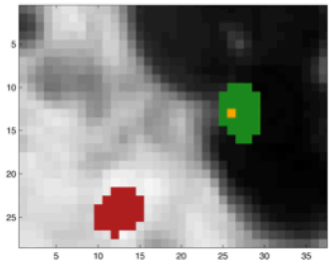
Sample 26, map 1



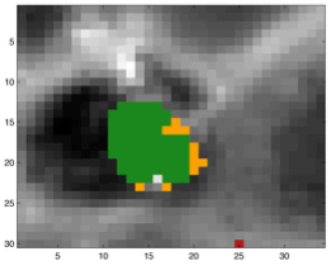
Sample 26, map 2



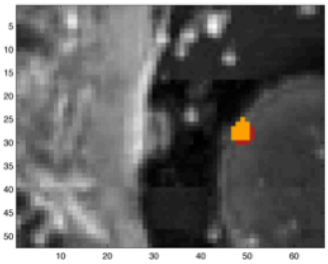
Sample 31



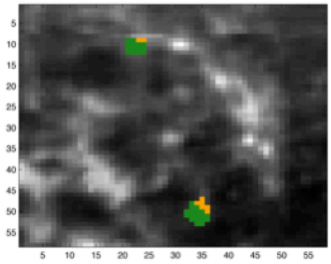
Sample 37, map 1



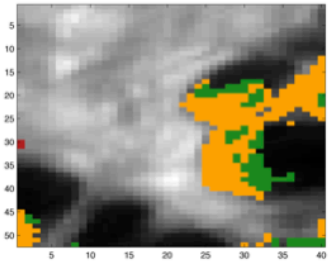
Sample 37, map 2



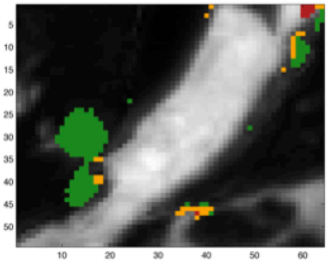
Sample 45



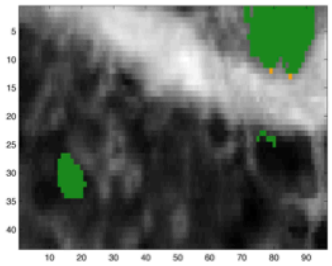
Sample 47



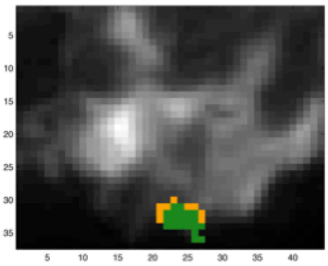
Sample 51



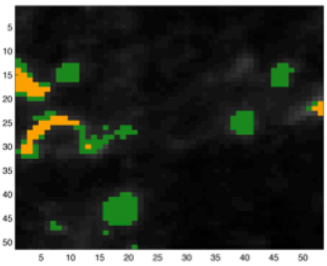
Sample 57



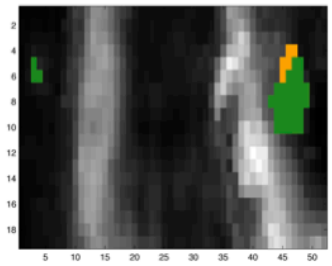
Sample 66, map 2



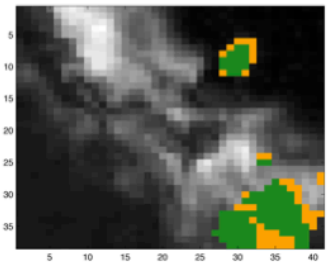
Sample 72



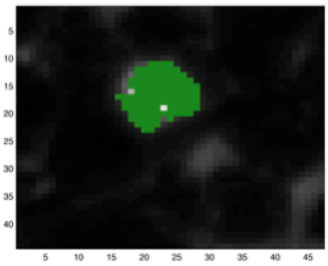
Sample 76, map1



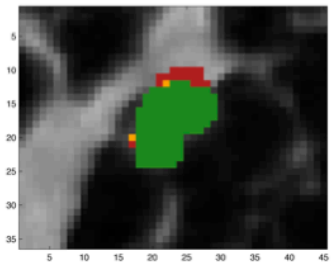
Sample 76, map2



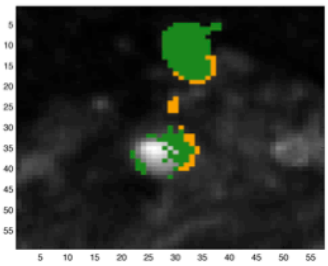
Sample 81, map 1



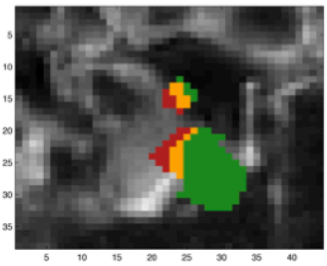
Sample 81, map 2



Sample 110, map 1

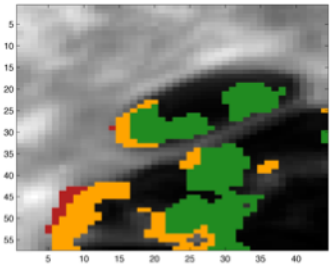


Sample 110, map 2

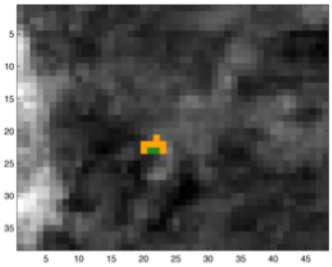


Ductal carcinoma *in situ* grade 1:

Sample 46

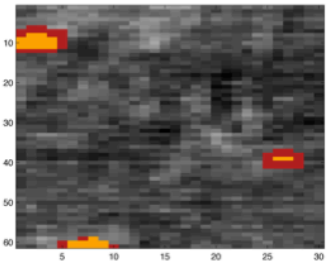


Sample 62, map 2

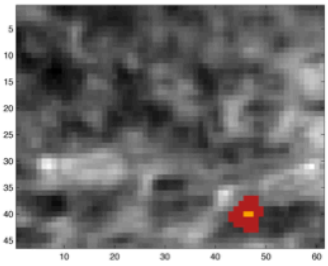


Ductal carcinoma *in situ* grade 2:

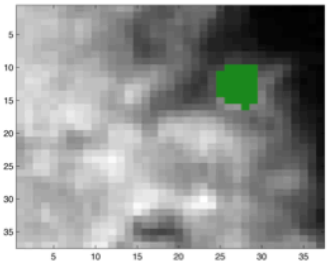
Sample 20, map1



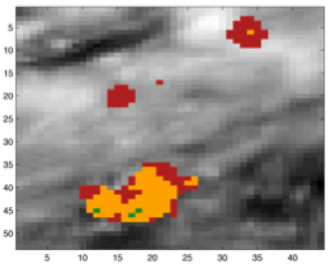
Sample 34, map 1



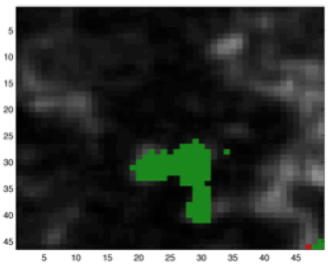
Sample 39, map 1



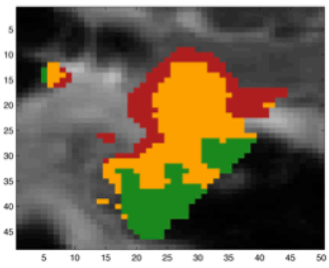
Sample 44, map 1



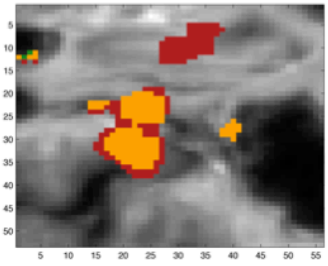
Sample 67



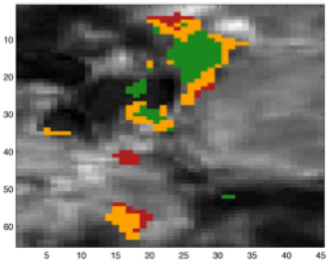
Sample 71, map 1



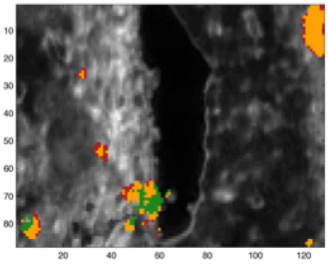
Sample 71, map 3



Sample 79, map 2

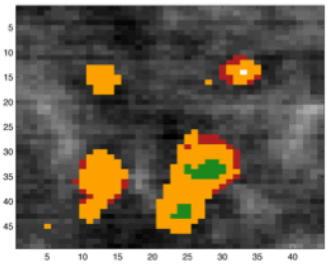


Sample 83, map 1

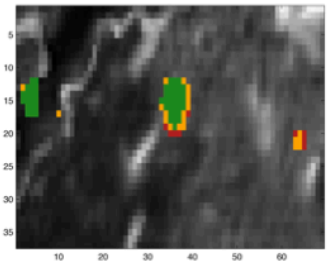


Ductal carcinoma *in situ* grade 3:

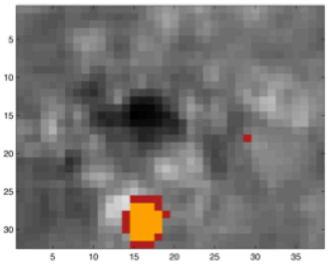
Sample 23, map 1



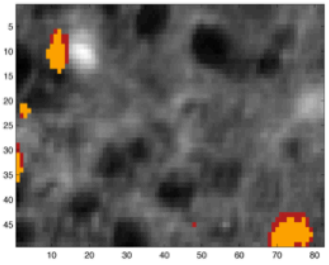
Sample 29, map 1



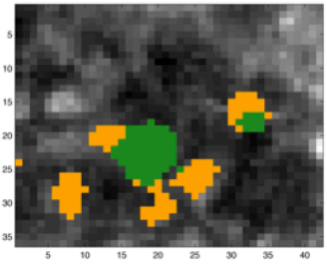
Sample 29, map 2



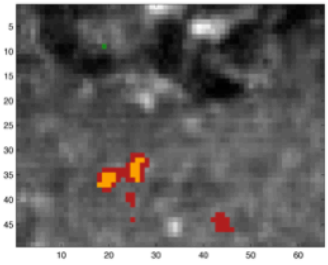
Sample 30, map 1



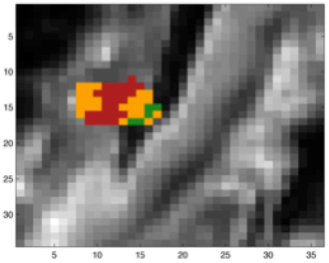
Sample 38 map 1



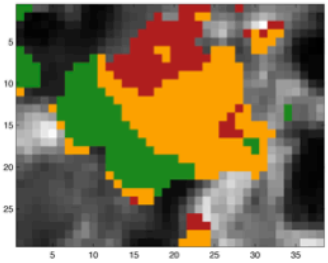
Sample 38, map 2



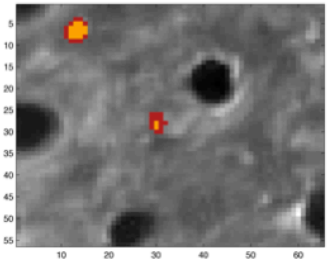
Sample 40, map 1



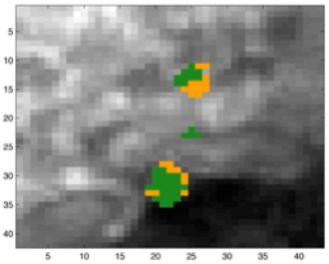
Sample 40, map 2



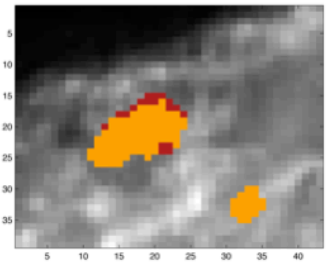
Sample 42, map 1



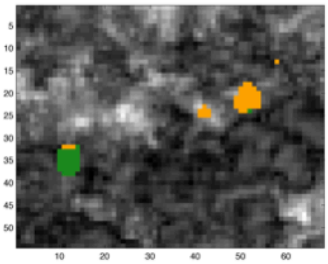
Sample 60, map 1



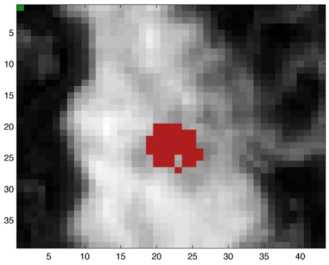
Sample 60, map2



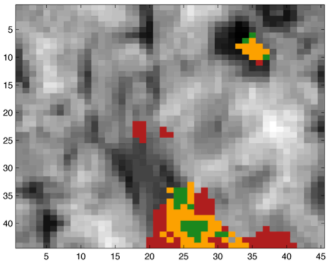
Sample 62, map2



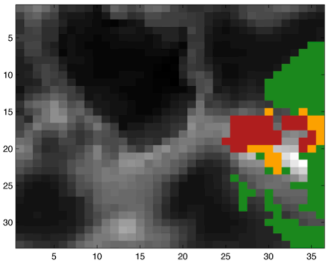
Sample 64, map 1



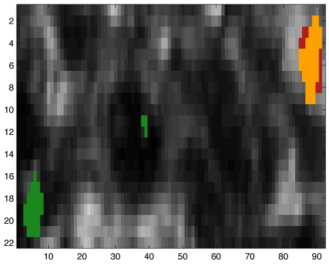
Sample 64, map 2



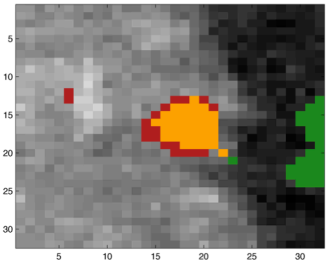
Sample 84A



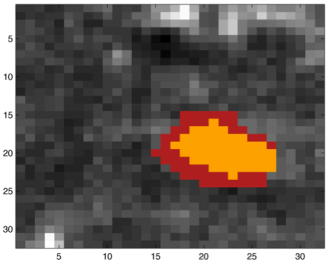
Sample 84B



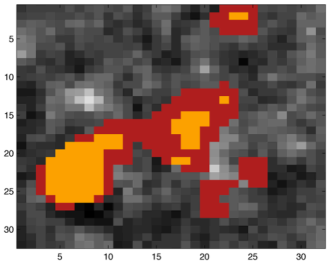
Sample 91, map 1



Sample 91, map 2

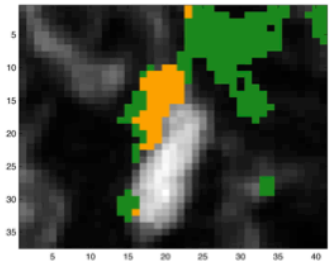


Sample 91, map 3

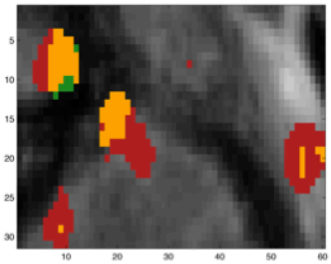


Ductal carcinoma *in situ* unknown grade:

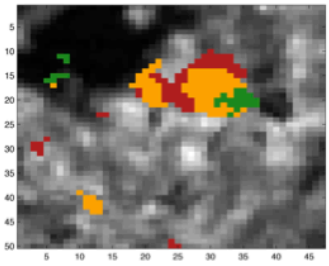
Sample 52



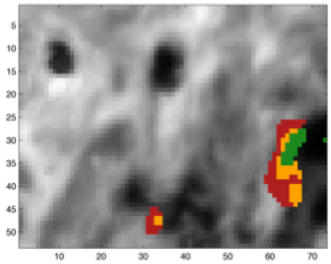
Sample 82, map 1



Sample 82, map 2

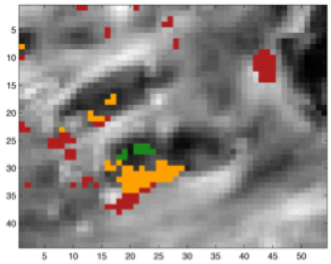


Sample 94

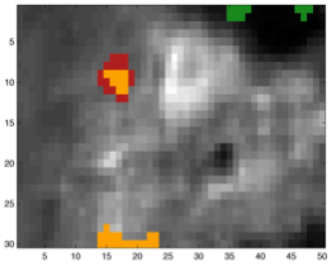


Invasive cancer grade 1:

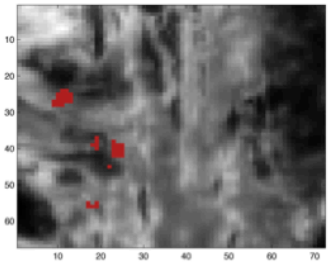
Sample 55



Sample 68

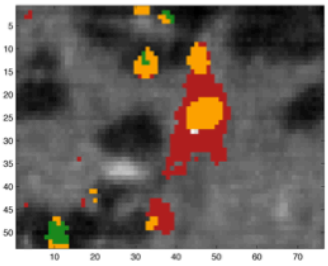


Sample 101

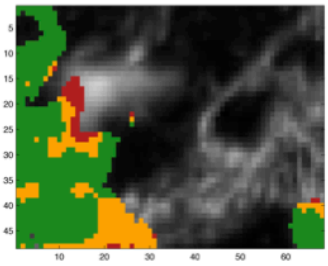


Invasive cancer grade 2:

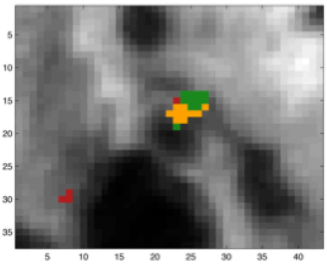
Sample 21, map 1



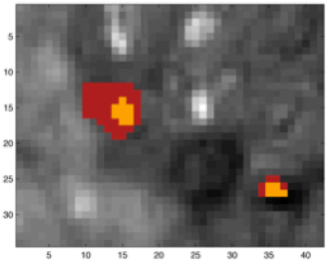
Sample 28, map 1



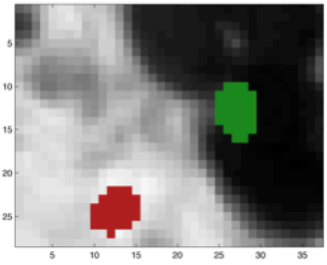
Sample 32, map 1



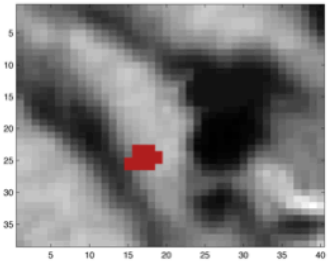
Sample 33, map 1



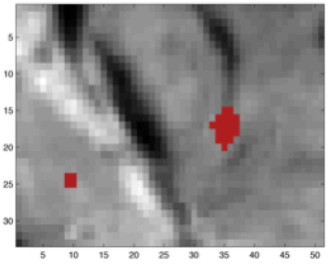
Sample 33 map 2



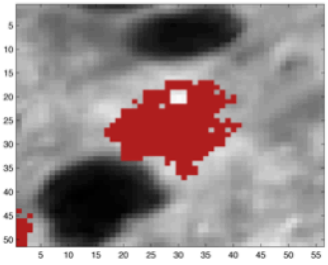
Sample 36, map 1



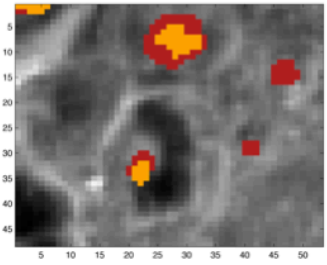
Sample 36, map 2



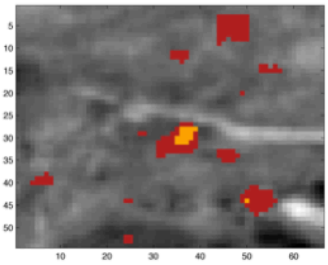
Sample 50



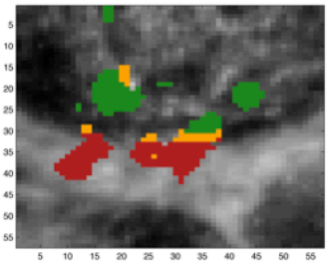
Sample 78, map 1



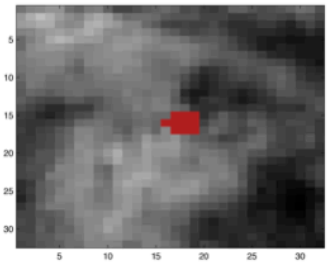
Sample 80, map 1



Sample 85

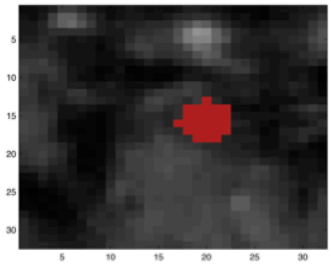


Sample 92, map 1

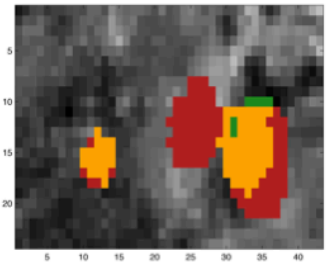


Invasive cancer grade 2:

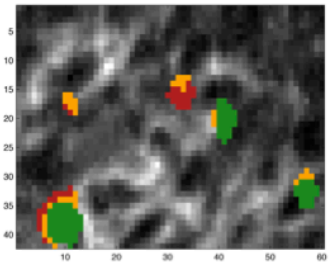
Sample 92, map 2



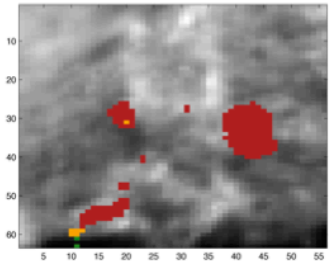
Sample 104, map 1



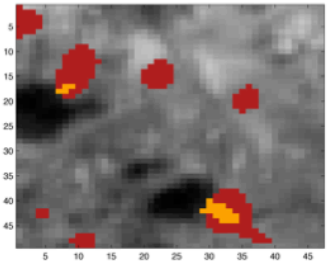
Sample 104, map 3



Sample 108

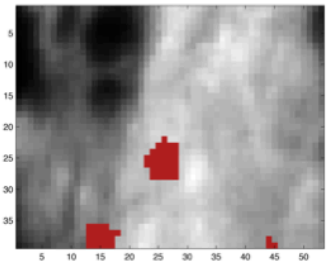


Sample 109

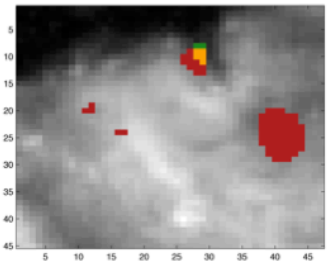


Invasive cancer grade 3:

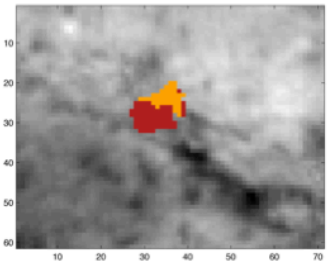
Sample 27, map 1



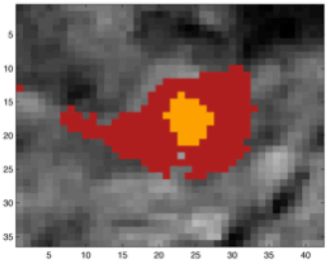
Sample 56



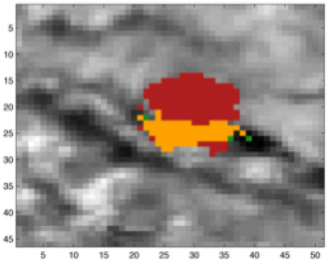
Sample 59, map 1



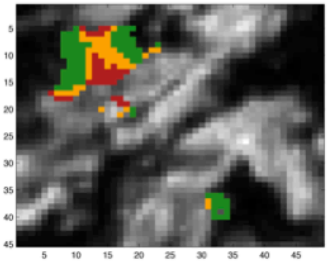
Sample 59, map 2



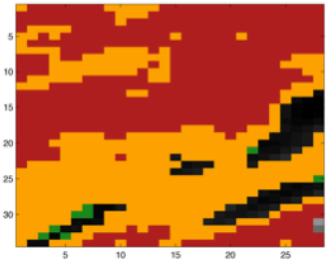
Sample 65 map 1



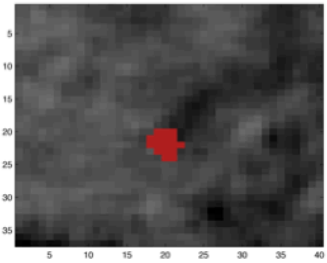
Sample 65, map 2



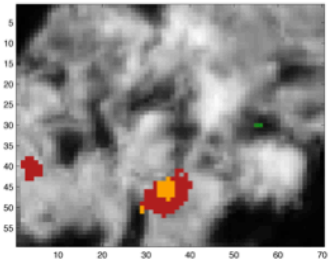
Sample 69



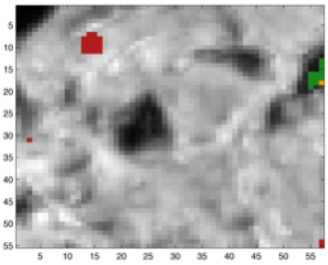
Sample 77, map 1



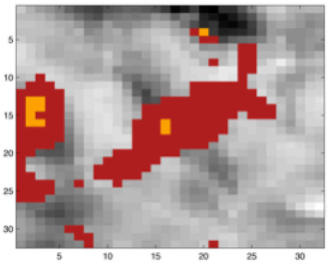
Sample 93, map 1



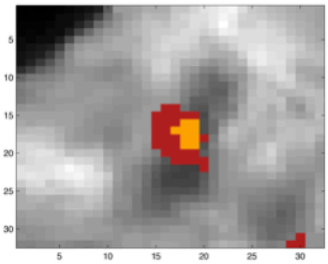
Sample 93, map 2



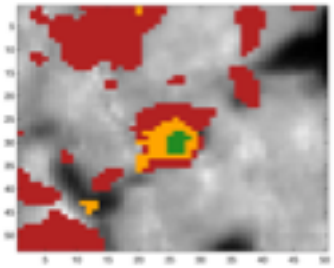
Sample 95, map 1



Sample 95, map 2



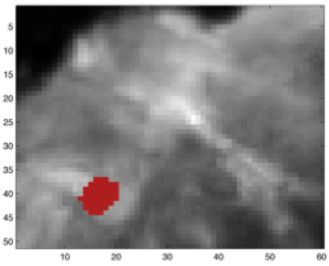
Sample 102



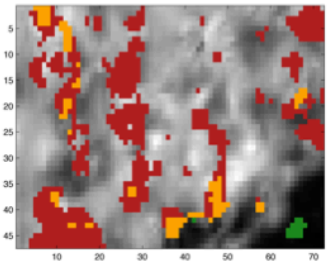
X

Invasive cancer grade unknown:

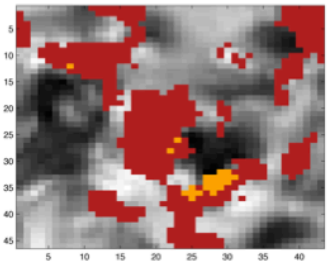
Sample 87



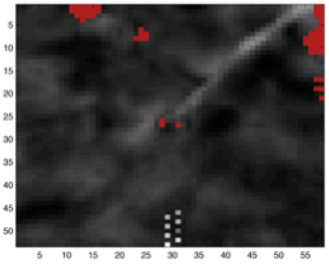
Sample 88, map 1



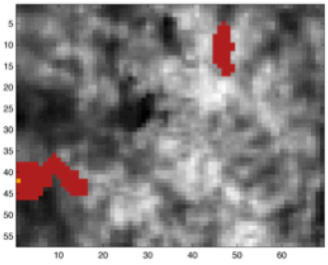
Sample 88, map 2



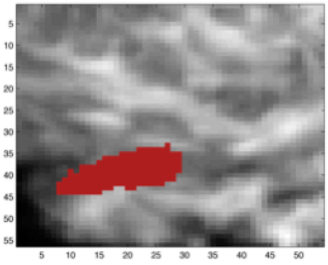
Sample 99



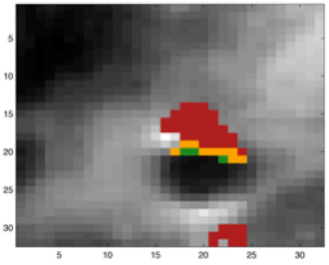
Sample 100



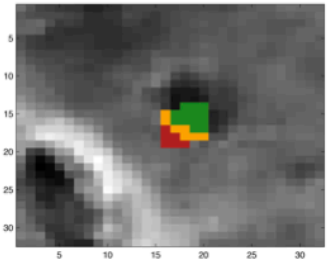
Sample 105, map 1



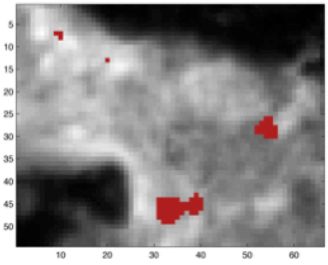
Sample 105, map 1



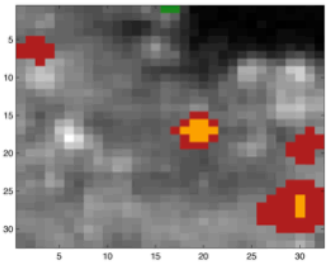
Sample 105, map 2



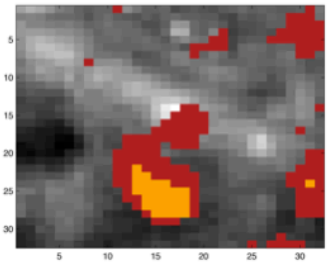
Sample 106



Sample 107, map 1



Sample 107, map 2



Publications