Improving Radiotherapy Using Image Analysis and Machine Learning

Dean Montgomery

THE UNIVERSITY OF EDINBURGH

2016

To Miriam,
my strong helper.

Abstract

With ever increasing advancements in imaging, there is an increasing abundance of images being acquired in the clinical environment. However, this increase in information can be a burden as well as a blessing as it may require significant amounts of time to interpret the information contained in these images. Computer assisted evaluation is one way in which better use could be made of these images. This thesis presents the combination of texture analysis of images acquired during the treatment of cancer with machine learning in order to improve radiotherapy. The first application is to the prediction of radiation induced pneumonitis. In 13-37% of cases, lung cancer patients treated with radiotherapy develop radiation induced lung disease, such as radiation induced pneumonitis. Three dimensional texture analysis, combined with patient-specific clinical parameters, were used to compute unique features. On radiotherapy planning CT data of 57 patients, (14 symptomatic, 43 asymptomatic), a Support Vector Machine (SVM) obtained an area under the receiver operator curve (AUROC) of 0.873 with sensitivity, specificity and accuracy of 92%, 72% and 87% respectively. Furthermore, it was demonstrated that a Decision Tree classifier was capable of a similar level of performance using sub-regions of the lung volume. The second application is related to prostate cancer identification.

T2 MRI scans are used in the diagnosis of prostate cancer and in the identification of the primary cancer within the prostate gland. The manual identification of the cancer relies on the assessment of multiple scans and the integration of clinical information by a clinician. This requires considerable experience and time. As MRI becomes more integrated within the radiotherapy work flow and as adaptive radiotherapy (where the treatment plan is modified based on multi-modality image information acquired during or between RT fractions) develops it is timely to develop automatic segmentation techniques for reliably identifying cancerous regions. In this work a number of texture features were coupled with a supervised learning model for the automatic segmentation of the main cancerous focus in the prostate - the focal lesion. A mean AUROC of 0.713 was demonstrated with 10-fold stratified cross validation strategy on an aggregate data set. On a leave one case out basis a mean AUROC of 0.60 was achieved which resulted in a mean DICE coefficient of 0.710. These results showed that is was possible to delineate the focal lesion in the majority (11) of the 14 cases used in the study.

Acknowledgements

I would like to thank the following people, without whom none of this would have been possible:

My wife, Miriam, for her love, laughter and patience.

My friends and family, and in particular my parents, for their love and encouragement over so many years.

My supervisors Bill Nailon, for his invaluable guidance and kind friendship, and Steve McLaughlin for his assistance and fresh perspective. Also all my colleagues at Covidien, and especially Paul Addison, for his flexibility and heartfelt support.

And finally, and most importantly, God. Everything I have is from Him, and without Him I could not have even reached the starting line.

"I can do all this through Him who gives me strength."
Philippians 4.13

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Dean Montgomery

Contents

Figures and Tables

Figures

Tables

Chapter 1

Introduction

1.1 Introduction

In recent decades continuous technological advancements have made the acquisition of medical images much cheaper and faster. These advances have been accompanied by an increase in the volume of imaging data that is generated by hospitals. This ever increasing amount of data can be time consuming to analyse and may contain an abundance of under exploited information from a great many patients. This phenomenon has been compounded by the prevalence of volumetric data in recent years - it is now common for three dimensional (isotropic) data to be acquired in the normal course of diagnosis and treatment, particularly when considering cancer patients. Alongside these advances in imaging, radiotherapy techniques and treatment options have also developed rapidly to allow higher doses to be administered to tighter constraints, leading to more sophisticated and effective treatment of different cancers.

Both of these advances in imaging and radiotherapy have intertwined the two areas and they are often used in tandem in many clinical work flows. This is particularly true in the treatment of cancer. For example, a typical patient with prostate cancer may undergo an MRI and CT scan for the purpose of diagnosis and radiotherapy planning, respectively.

This thesis seeks to identify approaches that can aid the clinician in their decision making processes, during the clinical work flow, by making use of image analysis and machine learning. In particular, texture analysis will be used to extract features from CT and MR images which are fed into a classification pipeline in order to predict radiation induced pneumonitis and identify focal disease in the prostate.

1.2 Clinical Context and Objectives

In 2011, there were 331,487 new cases of cancer and 161,823 deaths reported in the United Kingdom [30]. There are a wide range of treatments available depending on the location and stage of the cancer and any other co-morbidities. Examples of such treatments are chemotherapy, cryotherapy, surgery and radiotherapy. Many of these rely on imaging for diagnosis and planning, especially radiotherapy. This work presents novel methods for improving the treatment of two different cancers: lung cancer and prostate cancer. Imaging and radiotherapy often play a key role in the treatment of these two cancer types and it is for this reason that there is a significant opportunity for research into the application of advanced image processing and machine learning techniques to improve patient diagnosis and treatment. It is this context that this thesis is presented.

1.2.1 Prediction of Radiation Induced Pneumonitis

Lung cancer has the second highest incidence rate and the highest mortality rate in the world. In men it is the most commonly diagnosed cancer and also the leading cause of cancer death, in women it is the fourth most diagnosed and second leading cause of cancer death [58]. In 2011 43,463 people in the UK were diagnosed with lung cancer and there were 35,184 reported deaths attributed to lung cancer [31]. Treatment often involves surgery, chemotherapy, radiotherapy or a combination of these options. However radiotherapy, in a significant minority of cases, may induce pneumonitis.

Radiation induced pneumonitis is a specific instance of radiation induced lung injury and occurs in 13-37% of cases [92] after radiotherapy for lung cancer (See Section 2.5.4 for further details). Existing research into this problem has tended to approach the topic from a clinical angle and focus on the use of a wide range of clinical parameters or dose information to attempt to predict its occurrence. However many of these parameters can require considerable effort to collect and process and so an automated approach would be able to reduce additional burdens on clinical staff.

The objective of the approach presented in this thesis (Chapter 5) is to use a small amount of clinical data that is already collected as standard practice in the care of a lung cancer patient coupled with a number of texture features derived from a CT scan of a patient's lungs to predict the likelihood of radiation induced pneumonitis occurring. Three dimensional (3D) texture features are calculated from the lung parenchyma and passed to a classifier after preprocessing with principal component analysis (PCA). A binary prediction is then made as to whether this patient will develop pneumonitis. These predictions are then compared with a gold standard (follow up data) to assess the viability of the method.

A significant benefit of this approach is that the process may be run automatically and off line upon acquisition of a CT scan. Even if the texture analysis and classification requires several

hours to complete, the predicted outcome would almost certainly be ready before a clinician requires it. This would potentially allow for the modification of the dose plan so that treatment may proceed in a manner which minimises the risk of radiation induced pneumonitis. That is, the ability to predict the occurrence of radiation induced pneumonitis would assist clinicians in planning the most effective therapy for each patient with lung cancer.

1.2.2 Segmentation of Prostate Focal Lesion

Prostate cancer is the second most commonly diagnosed cancer in men and the sixth leading cause of cancer deaths in men worldwide, making up 14% of the total new cancer cases and 6% of male cancer deaths in 2008 [58]. In 2011, 41,736 men in the UK were diagnosed with prostate cancer and there were 10,793 reported deaths attributed to prostate cancer [32]. MRI is often used for diagnosis and radiotherapy is a popular treatment choice when intervention is required.

When radiotherapy is used in the treatment of prostate cancer the whole gland is almost always treated - the whole of the prostate is within the clinical target volume [18]. However, in a significant number of prostate cancer cases, the disease is contained within a much smaller part of the prostate (the “focal lesion” or “focal disease”), usually near the periphery. If this region can be reliably identified then avenues for sparing the remaining healthy prostate may open up. This would enable the reduction of the side effects of irradiating the whole prostate with a therapeutic dose. One potential treatment plan would be to reduce the average dose to the whole prostate and “boost” the focal disease with a much higher dose. Currently, one of the main issues with this approach is in identifying the focal disease in a reliable and timely fashion within the normal radiotherapy planning time line (See Section 2.6.4 for further details).

A first step towards a potential solution is presented in Chapter 6 where the focal disease is automatically delineated using texture analysis on MRI data. A set of three dimensional texture descriptors are calculated for every point in the prostate volume and an AdaBoost classifier is used to predict whether a pixel is healthy prostate tissue or cancerous tissue. These labels are then mapped back onto the MRI image for evaluation and comparison with a contour outlined by an expert clinician. Transfer of these contours to planning CT scans by registration (as demonstrated in [41]) could allow for the creation of radiotherapy plans that boost the focal disease.

1.3 Contributions

The contributions made by the author towards lung cancer treatment are:

- the development of a novel image analysis approach for the prediction of radiation induced pneumonitis using radiotherapy planning CT images,
- the combination of 3D texture features with clinical features (such as smoking history) to create a classification pipeline,
- the demonstration of the efficacy of the approach for prediction of pneumonitis using features derived from the:
  - whole lung volume
  - region surrounding the GTV
  - V_{20}.

The contributions made by the author towards prostate cancer treatment are:

- the development of a novel pathway for the automatic delineation of focal disease in the prostate volume using 3D texture features,
- the demonstration of the ability of 3D texture features (derived from diagnostic MRI images) to correctly classify focal disease when considering an aggregate data set made up from multiple cases,
- the demonstration of the efficacy of the same texture features when applied in a leave one case out fashion and combined with a reconstruction and morphological cleaning step to create a clinically viable approach.

1.4 Outline of thesis

The rest of this thesis is structured as follows: Chapter 2 provides a more substantial medical background and context, focussing on the imaging modalities and radiotherapy relevant to Chapters 5 and 6 and also lung and prostate cancer. Chapter 3 explicates texture analysis theory and applications with a particular focus on the methods used in later this work. Chapter 4 follows a similar approach for machine learning, examining the supervised learning methods used in Chapters 5 and 6.

Chapter 5 lays out the methodology used to predict radiation induced pneumonitis using radiotherapy planning CT scans and a set of clinical features. Results are then presented and analysed. Additional approaches that use a subset of the information from the lung and results related to the dose volume are also presented.

Chapter 6 presents the work related to the identification of the prostate cancer focal disease using texture. Initial classification results are presented, morphological cleaning applied and the resulting contours are analysed and discussed.

The final chapter presents conclusions and potential avenues for further work.

Chapter 2

Medical Context and Background

2.1 Introduction

In this chapter a short introduction to the roles of CT imaging, MR imaging and radiotherapy in a modern clinical environment are introduced to provide a basic working knowledge of the images and clinical parameters used in Chapters 5 and 6. This is followed by an overview of lung cancer and radiation induced pneumonitis and finally prostate cancer with an emphasis on the treatment of focal disease where this work seeks to contribute.

2.2 CT Imaging

X-ray Computed Tomography (CT) [95] combines the measurement of x-ray attenuation through an object with a back propagation algorithm to create a series of slices of a patient which then form a 3D image. Two example slices from a CT scan can be seen in Figure 2.1. The main advantages of this modality are:

- high contrast between bone and soft tissue,
- high spatial resolution (1mm or better),
- no superimposition of 3D structures onto a 2D plane,
- quick acquisition times,
- consistent image geometry.

The main disadvantages are:

- x-rays are ionising which results in a dose of radiation being delivered to a patient each time a CT scan is acquired[i],
- poor soft tissue contrast,
- contrast agents may be required for some diagnostic applications,

i. The dose from a CT scan (on the order of mGy) is much less than the dose delivered during radiotherapy treatment (on the order of tens of Gy) and is less of a concern during cancer treatment than when used for other purposes. See the Ionising Radiation (Medical Exposure) Regulations 2000 (IRMER). Downloadable from `gov.uk/government/publications/the-ionising-radiation-medical-exposure-regulations-2000`. Although it should be noted that the dose can become more of an issue for 4DCT and 4DCBCT acquisitions.

- some patients may be incapable of maintaining required breath holding (though this is less problematic with the increased development and availability of 4DCT),
- artefacts when implanted metal objects are present, such as an artificial hip.

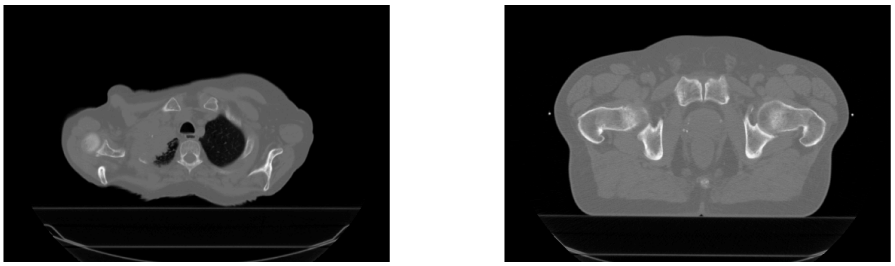


Figure 2.1: Example CT images. Left: an axial slice from a CT scan used for RT planning of a patient with lung cancer. Right: an axial slice from a CT scan used for RT planning of a patient with prostate cancer.

2.2.1 History

X-ray tomography was first developed at the beginning of the 20th century [72]. Initially a purely mechanical approach was used to rotate an x-ray source and film joined by a connecting rod, which allowed for the imaging of a single slice of the object of interest (a human body, for example). New and improved methods for conventional tomography were developed over the course of the next 60 years, resulting in sharper images.

With the advancements in computing in the 1960s, computed tomography became a possibility and the first commercial scanner was produced by EMI (Electric and Musical Industries Ltd) and installed at Atkinson Morley Hospital in 1971. This first machine took a little over five minutes to acquire a scan using a single x-ray tube and detector with a translate-rotate motion. By 1973 the second generation of scanners had been developed, still using a translate-rotate motion but now with 30 detectors and a single tube, which decreased the scan time to approximately one minute. The third generation of scanners, introduced in 1976, had a rotate only motion and a single tube. Using approximately 800 detectors, a single image could be acquired in one to five seconds. The next major advancement (around 1980) was the expansion to approximately 2000 fixed detectors and a rotating x-ray tube. Images still required between one to five seconds to acquire.

Until this time all scanners used axial scanning in which the patient must be moved in discrete steps between the acquisition of each slice. This process is repeated until the desired volume has been captured. Around 1990 helical scanning was introduced which allowed for the smooth motion of tube and couch, leading to reduced movement artefacts and higher throughput. Modern scanners are equipped with multiple parallel rows of detectors to enable simultaneous acquisition of multiple slices. They operate in axial or helical mode and have tube rotation times of 0.3s. These features allow for fast scanning of large volumes and higher resolution along the z-axis for true 3D imaging.

Examples of recent innovations in the field include *a*) 4DCT which over samples the patient and tags each scan with a time or breathing signal to allow for the creation of 3D CT data sets at set points in the breathing cycle and *b*) dual source CT (for example, Siemens) which offers very high temporal resolution to freeze heart motion by using two x-ray tubes at right angles in order to increase the temporal resolution of the scan to a quarter of the tube rotation time.

Currently ongoing research may result in *a*) the use of compressive sensing techniques in the reconstruction of the projection data for better image quality at lower doses [100] or *b*) X-ray phase contrast imaging to reduce dose by measuring phase changes instead of absorption, however, this is still at the research stage because miniaturisation of narrow band x-ray sources is required.

2.2.2 Physics and Image Acquisition

Figure 2.2: A CT scanner at the Western General Hospital.

A CT scanner, Figure 2.2, generates x-rays by using a high voltage electric field to accelerate electrons emitted by a hot cathode. These high energy electrons strike a metal target (typically Tungsten) which causes the generation of a continuous spectrum of x-rays due to Bremsstrahlung and characteristic peaks due to electron transitions in the outer shell of the

target atoms. Bremsstrahlung (braking radiation) is radiation generated by the deceleration of one charged particle by another; in this case the electrons are slowed by the Tungsten nuclei. Each electron typically releases many x-rays as it moves through the metal target generating the curve shown in Figure 2.3. If an electron emits a single x-ray then the x-ray will be of the same energy as the kinetic energy that the electron obtained from the electric field. This is the maximum energy x-ray that can be emitted and is the reason for the sharp cut off in Figure 2.3. The characteristic peaks are produced when the incident electrons cause a bound electron to be ejected from its orbital shell, subsequently allowing an electron in a higher energy shell to transition into the empty lower energy state and emit a photon with energy equal to the difference of the two orbital states. This process is called fluorescent radiation.

It is normal for the beam to then be filtered (to "harden" the beam) in order to remove lower energy x-rays, as they do not contribute to image quality but increase the patient's received dose. The final step is to pass the beam through a set of collimators to produce parallel rays for imaging of the patient.

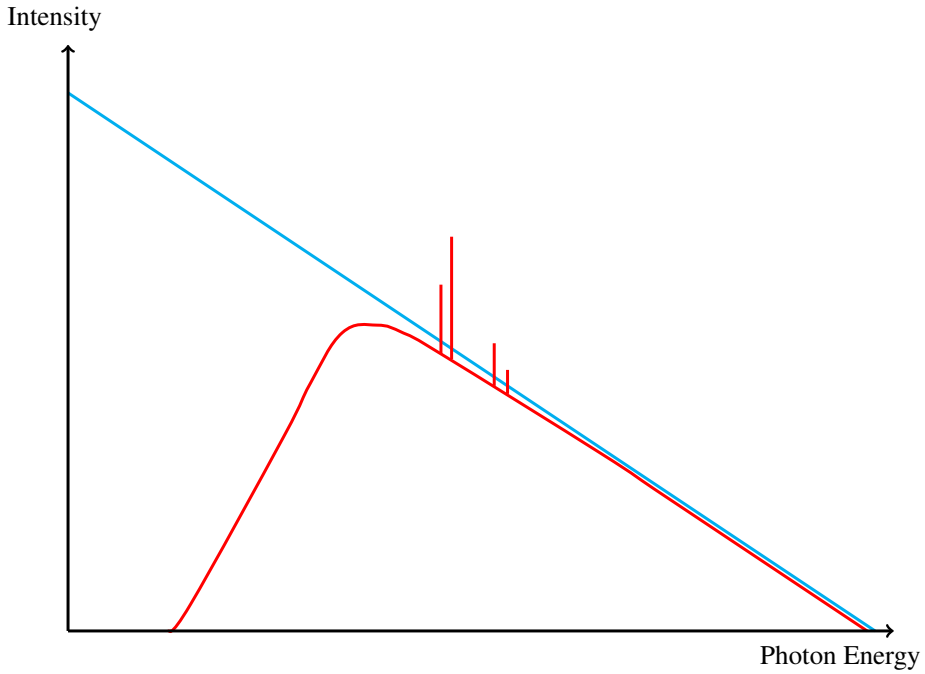


Figure 2.3: A typical x-ray spectrum (red) generated by the crashing of electrons into a heavy metal target. The characteristic lines due to electron transitions can be seen in addition to the Bremsstrahlung curve with a peak at a third of the maximum energy. The blue line is the equivalent unfiltered spectrum in a vacuum.

The transition of these x-rays are measured at many angles and an attenuation map is constructed by back projection. This attenuation map is an array of Hounsfield Units (HU) which

are assigned a grey level scale to generate the final image.

$$\mathrm{HU} = 1000\frac{\mu_{tissue} - \mu_{water}}{\mu_{water}} \tag{2.1}$$

μ_x is the attenuation of the x-ray beam (I) passing through a distance (d) of material x:

$$I_x = Ie^{-\mu_x d} \tag{2.2}$$

The mapping of CT numbers to the grey level scale can be varied to allow for improved visualisation by choosing mappings that "stretch out" the Hounsfield Units of interest across a greater range of grey levels. In radiotherapy planning the Hounsfield Units are used for dose calculation.

Accelerating potentials are usually in the 50kV to 120kV range and the effective x-ray energy is approximately one third of this. Up to energies of 50keV the photoelectric effect dominates and between 50keV and 100keV both the photoelectric and Compton effects are important. X-ray energies up to 100keV are typical for diagnostic imaging as the photoelectric and Compton effects are dominant at this energy. The photoelectric interactions normalised by density are given by

$$\frac{\tau}{\rho} \propto Z^3 E^{-3} \tag{2.3}$$

(Z = atomic number, E = energy, ρ = density) and Compton interactions are given by

$$\frac{\sigma}{\rho} \propto E^{-1} \tag{2.4}$$

The photoelectric effect is preferred as it allows for the contrast of soft tissue and bone due to the different atomic number (Z) of bones and soft tissue.

Reconstruction

As briefly mentioned above, the CT image is created by back projection [64]. This involves taking a series of measurements at different angles to determine the attenuation coefficient along the line between the source and detector. These measurements are then projected back on top of a matrix: in the naïve implementation the value of the attenuation coefficient is equally divided into the cells of the matrix that fall on the line between the source and detector. By repeating this process for many angles an image is built up. Basic back projection suffers from a blurring effect that can be alleviated by filtering the data, before back projection, with a one dimensional kernel function to correct for unequal sample densities in the spatial domain. These unequal sample densities are caused by the radial sampling of the rectangular grid of the image. In an ideal scenario, with an infinite number of angles and infinite samples along each view, the back projected image will exactly correspond to the "correct" image.

Another method of reconstruction is to make use of the Fourier slice theorem. This states that the Fourier transform of a 2D function projected onto a 1D line is equal to a slice through the origin, parallel to the projection line, of the 2D Fourier transform of the original function. This allows for the Fourier transform of each view to be combined to construct the two dimensional image in the frequency domain. The inverse Fourier transform can then be used to obtain the image in the spatial domain.

2.2.3 Roles in Medical Imaging

CT scans have two main roles in medicine; diagnosis and radiotherapy planning. CT scans are used to diagnose many different diseases, often alongside other modalities. Examples include lung cancer, brain tumours, spinal injuries, complex bone fractures and assessment of coronary arteries.

The other main application is to provide the information required for radiotherapy dose calculation and planning. A planning CT is acquired and the tumour and organs at risk (OAR) are outlined in order to calculate the dose delivered to each organ and the tumour. This enables the correct treatment of the disease with minimal impact on surrounding healthy tissue. The HU that make up the CT image are a measurement of electron density which is required to calculate the dose deposited in radiotherapy (see Section 2.4 for more details).

2.2.4 Cone-Beam CT

CBCT [94] is an imaging modality similar to CT: x-rays and tomography are used to reconstruct an image of a patient. However the x-rays used in CBCT are divergent - forming a cone shape as they are emitted from the x-ray source rather than conventional CT where they pass through a collimator. CBCT is used for patient alignment during radiotherapy delivery, patient follow up and is becoming common in dental applications.

CBCT gives a poorer quality image than conventional CT (primarily due to the divergent beams), data acquisition takes longer and the cone beam reconstruction takes much longer than conventional CT reconstruction (1 minute versus near real time for conventional CT, due to the assumptions that allow quicker reconstruction of conventional CT no longer holding). However CBCT is a useful supplementary modality as a CBCT scanner can be integrated into a radiotherapy linac to allow for pre-treatment and post-treatment patient alignment. CBCT can also be acquired during treatment to monitor tumour response without the need for extra scans on another machine. It should also be noted that CBCT technology is improving with the availability of 4D CBCT and improvements to image contrast and quality.

Figure 2.4: A example of a CBCT scan of a patient with lung cancer. Notice the poorer image quality compared to a CT image - particularly the reconstruction artefacts (Figure 2.1).

2.3 MR Imaging

Magnetic Resonance Imaging (MR/MRI) [114] is an imaging modality that exploits the different relaxation times of nuclear spin states to construct images. The main advantages of MR are:

- good soft tissue contrast,
- image slices can be acquired in any plane,
- non-ionising, there is no radiation dose deposited in the patient.

While the disadvantages include:

- slow acquisition times in the range of 15 to 90 minutes are typical,
- magnetic medical implants can prevent the use of MRI due to the strong magnetic fields used,
- geometry is less consistent than CT imaging,
- the imaging process involves loud noises from the rapid switching of gradient fields which may unsettle some patients in the confined space of an MRI scanner,
- contrast agents can be required for some diagnostic usage.

Figure 2.5: An example slice from an MRI (T2) scan of a patient with prostate cancer.

2.3.1 History

MR Imaging was only possible after the development of Nuclear Magentic Resonance (NMR) as NMR is a fundamental building block of MRI [12]. NMR uses a strong uniform magnetic field and a radio frequency pulse to identify the chemical composition of a liquid by measurement of the relaxation times of the magnetisation of nuclei in a sample.

Electron spin was discovered in 1921 by Compton, and in 1924 Pauli proposed that nuclear spin also existed. In 1946 a joint Stanford and Harvard group were the first to detect an NMR signal. By the 1970s, NMR had become a powerful tool for the structural analysis of molecules - NMR spectra allows the relative position of different atoms within a molecule to be inferred.

With the advent of superconducting magnets in the 1970s the strength of the magnetic field was increased and in 1973 Lauterbur created the first NMR image of sample tubes in a chemical spectrometer. By creating a magnetic gradient, localised NMR measurements can be made and later processed to form an image. The first commercial scanner was produced in 1981 ($<$0.2T) and in 1985 the first 1.5T scanner was manufactured. The next few years brought major advancement in the signal to noise ratios and the resolution of the images. Today MRI scanners are a common tool used in diagnosis and research and are widely available in the western world with magnetic field strengths in the 1.5T to 7T range.

2.3.2 Physics and Image Acquisition

There are a large number of protons in human tissues (water and fat, for example). Each proton has a quantum spin of $\frac{1}{2}$ and thus a magnetic moment. When a person is subjected to a very strong magnetic field inside an MR scanner all of these protons precess around an axis in the direction of the magnetic field in one of two energy eigenstates determined by the nuclear Zeeman effect:

$$\Delta E = \gamma \hbar B_0 \tag{2.5}$$

where $\hbar$ is the reduced Planck constant, B_0 is the magnetic field strength and $\gamma = 42.5 M\text{hz/T}$ for hydrogen. The lower energy state is made up of protons precessing parallel to B_0 and the higher energy state anti-parallel to B_0. Protons can be excited from the lower to the higher energy state by the absorption of a photon of the correct (resonance) energy. For the strength of fields used in MRI, this corresponds to a photon in the radio frequency range and this frequency is called the Larmor frequency.

The quantity that is measured in MRI is the time taken for the spins of the protons to recover after being excited by a radio pulse. T_1 MR measures the time taken for the longitudinal magnetisation to relax. This is a spin-lattice interaction governed by how easily the protons lose heat to the surrounding environment, this is determined by the molecule the proton is a part of: lipids will have a shorter T_1 than water as lipid molecules are much larger than water and will interact with their local environment more. T_2 MR measures the loss of phase coherence between spins, the transverse magnetisation relaxation. Nearby protons can interact and cause each other to flip alignment. These interactions cause a loss of phase coherence between the protons and this results in the transverse magnetisation decaying with time constant T_2. In practice magnetic field inhomogeneities cause this time constant to be shortened: this time constant is called T_2^*. T_2 is shorter for larger molecules as it is dependent upon the presence of nearby protons.

In order to acquire an image additional magnetic fields are applied, using coils, to create linear variations in B_0 in the x, y and z directions. These variations in the magnetic field cause the Larmor frequency to vary with position inside the patient. Slice position can then be chosen by varying the centre frequency of the applied pulse and the slice thickness can be chosen by varying the magnetic gradient strength and the bandwidth of the pulse:

$$B_z = B_0 + G_z z \tag{2.6a}$$

$$f_z = \gamma(B_0 + B_z z) \tag{2.6b}$$

where B_z and is the magnetic field gradient at a z coordinate and f_z is the Larmor frequency at that z same position. As such, the MR data is not acquired in the spatial domain, but rather in k-space (the spatial frequency domain). More detail in the reconstructed image requires more sampling over k-space which requires more time to acquire resonance signals over different

gradients. More contrast is achieved by greater sampling of the centre of k-space.

Other acquisition methods exist [10], including:

- Echo Planar Imaging - very fast, uses multiple phase encodings per excitation but puts stress on MR equipment as very fast gradient switching is required.
- Diffusion Weighted Imaging - measures the rate of perfusion. However, numbers obtained are not true perfusion like PET imaging.
- Fluid attenuation inversion recovery (FLAIR) - the T1 time of the MR pulse sequence is adjusted to remove the effects of fluid on the scan.

2.4 External Beam Radiotherapy

External beam radiotherapy uses ionising radiation to kill human cells within the body. By using radiation beams from multiple directions a dose can be deposited in a region of choice while a much lower dose is delivered to organs at risk of receiving radiation, organs at risk (OAR). This basic principle of building up dose by using multiple beams underlies external beam radiotherapy and allows for the treatment of cancer at many locations within the body.

The main advantages of radiotherapy include:

- non-invasive treatment,
- easy to treat most parts of the body,
- can be used alongside surgery and chemotherapy,
- curative treatment option.

While the disadvantages are:

- side effects include fatigue, nausea, skin soreness and possible infertility,
- chance of secondary cancers developing due to radiation exposure,
- not suitable for curative treatment of non-localised tumours such as bone metastases, though it can be used palliatively in such situations.

2.4.1 History

X-rays were discovered in 1895 [53] and only a few weeks later they were used to treat a cancer patient. In 1899, two Swedish doctors used x-rays to treat several patients with head and neck cancer. Until 1922 doses were delivered in a single exposure which brought severe side effects, then it was demonstrated that fractionated radiotherapy was equally effective at treating cancer whilst causing fewer side effects. By the 1950s 8MV linear accelerators (linacs),which allow for much higher energy x-rays to be produced, had been developed. These higher energy x-rays allowed for better treatment and increased skin sparing for patients.

Since the 1950s radiotherapy has continued to develop with advances including the integration of CT into the planning process, more sophisticated dose calculation methods, conformal

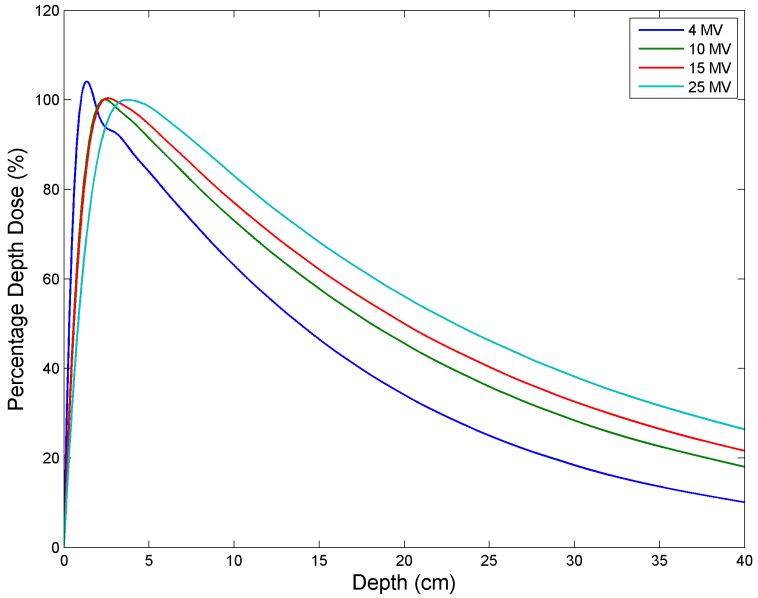


Figure 2.6: Percentage Depth Dose for various photon energies. These curve display the skin sparing effect - note the small amount of dose delivered in the first few centimetres and how the skin sparing effect increases with beam energy. Data taken from [4].

radiotherapy, intensity modulated radiotherapy and stereotactic ablative radiotherapy. There is currently much active research into adaptive (image guided) radiotherapy.

2.4.2 Physics and Treatment Procedure

X-ray Production

X-rays for radiotherapy are generated in a similar fashion to those used for CT imaging (Section 2.2.2) except they are of a higher energy, typically 6-10MeV [13]. These x-rays are produced by accelerating electrons in a linac which uses either a magnetron (lower energy) or a klystron (higher energy) to generate microwave energy. This microwave energy is used to accelerate the electrons in a waveguide and these electrons are shaped by a magnetic field to form a beam. For higher energies this equipment is too long to be mounted vertically and so the beam must be bent downwards towards the patient from the horizontally mounted wave guide. This bending is achieved by using a magnetic field to bend the beams through 270° degrees (rather than 90°) which helps to compensate for the spread of trajectories caused by the range of electron energies. This beam is passed through a slit which only allows electrons of the desired energy to pass (due to the different electron trajectories). In order to generate the x-ray photons the electron beam is directed towards a heavy metal target with which they interact to produce a beam of photons which are passed through a collimator to produce a beam of collinear photons. (In the case of electron treatments, the heavy metal target is replaced with a

scattering foil to spread out the electron beam to evenly cover the treatment field.) This beam of x-rays is passed through a flattening filter to ensure a uniform intensity across the beam profile. An ionisation chamber is the next step in the path of the beam which allows for monitoring of the dose delivered by the beam and to indicate when the beam should be shut off to prevent overexposure. At this point a secondary set of collimators can be used to shape the beam profile in order to conform the beam to a treatment volume. Traditionally this shaping was done with physical wedges, this progressed to "virtual wedges" where one of the collimator jaws moves in a gradual fashion during the radiation delivery. Modern machines are now equipped with multi-leaf collimators which consist of a series of pairs of metal leaves which can be adjusted during delivery to generate complex beam profiles for the detailed conforming of the beam to a target volume.

Dose

X-rays incident on a patient will interact with the matter in the patient by various processes: the photoelectric effect, Compton scattering and pair production. These processes transfer energy from the x-rays to electrons, it is these secondary electrons that cause nearly all of the ionisation in the body.

Absorbed dose is measured in Gray (Gy), 1Gy = 1J/kg. The typical dose prescribed for treatment depends on the type and stage of cancer, patient health, whether the treatment is curative or palliative and other co-morbidities a patient may have. Usually therapeutic doses are between 20 and 80Gy delivered in several fractions. Energies of a few MeV are typical for radiotherapy as this is the region in which Compton scattering dominates. Compton scattering is nearly independent of atomic number which helps reduce the dose received by bones.

When a photon interacts with matter in the body a shower of secondary photons and electrons is initiated causing the dose to spread about the initial direction in a cone. Since the energy is deposited over the range of the electron tracks there is a "build-up" effect: the maximum dose is not delivered until a depth roughly equal to the range of the electrons generated by the Compton scattering interactions. This process is known as the "skin-sparing effect" because it results in the skin being spared a high dose of ionising radiation during treatment. Without this effect the skin would receive a very high dose and radiotherapy may not be a viable treatment for many diseases.

Biological Effect

The primary means by which radiotherapy kills cells is by causing a break in the DNA molecules within the cell nucleus. DNA has a double helix structure: a break in one of the strands of the helix can be repaired by cellular repair enzymes, however a break in both strands cannot be repaired by the cell and will either stop proper cell function or prevent successful cell replication, both of which lead to cell death. Breakages in the DNA are caused by interaction with free radicals produced by the incident ionising radiation, as the dose is increased the number of free radicals increases and a double break in the DNA becomes more likely.

Fractionation

After exposure to radiation the tissue in the body is able to recover. When it was observed that healthy tissues recover quicker than cancer cells it was realised that a fractionated delivery would spare healthy tissue while killing the cancerous tissue. A fractionated scheme involves the administration of a prescribed dose in many smaller "fractions", typically over the course of several weeks. A further benefit of fractionation is that it may allow for the reoxygenation of cells within the tumour (required for free radical production) as the tumour shrinks with the treatment of the periphery - allowing previously radioresistant cells to be killed. Fractionation schemes have been developed using empirical results from clinical trials due to the complexity of biological radiation models [38].

Planning

In order to treat a patient a radiotherapy plan must be created. A plan consists of the arrangement of several beams and their shaping. The first step in the planning sequence is to acquire a CT scan and to outline the disease and any OAR from radiation (the CT numbers (HU) also provide a measure of electron density, allowing for the dose calculation). The Clinical Target Volume (CTV) is the volume which includes the Gross Tumour Volume (GTV) and any subclinical malignant disease that is to be treated. The CTV is then expanded to a Planning Target Volume (PTV). In order to ensure proper treatment of the CTV, the PTV adds a margin to compensate for organ motion, error in the delineation of the CTV and also for any unobservable cancer that may be in the vicinity of the visible tumour. After the CT scan has been outlined appropriately and the prescribed dose and any constraints have been set a suitable plan can be drafted. This once involved manually choosing the number of beams, their orientation and any shaping required to properly treat the PTV while attempting to spare any nearby organs at risk. Now, however, this step in the planning procedure is formulated as a numerical optimisation problem - the beams and their orientations are chosen algorithmically in order to satisfy dose constraints imposed on the tumour and OARs by a clinician.

There are many dose calculation algorithms in clinical use with trade-offs between their speed and accuracy. The convolution/superposition algorithm uses the precomputed probability dis-

tribution of a single photon interaction (pre-calculated using Monte Carlo simulations) and the attenuation coefficient from each point in the patient (from CT numbers) to calculate how much dose is deposited in the body. However the attenuation of the electrons and scattered photons is dependent on the attenuation map preventing the use of Fourier space for this calculation. Being a convolution problem, in the spatial domain, the algorithm is an N^6 problem which would be impossible to solve without a way to reduce the runtime. A common way to achieve this speed up is by using the collapsed cone algorithm which divides space around a point into a series of cones and assumes that all the energy within the cone is deposited along the cone's axis. The complexity of this algorithm is MN^3, where M is the number of cones used, which makes this a viable approach. Monte Carlo approaches are also used where the path of a large number incident photons (or electrons/protons, depending on the therapy) are simulated for a given beam configuration in order to calculate the dose a patient will receive for a given RT plan. Monte Carlo methods are computationally expensive, but can bring great accuracy and flexibility in planning.

2.4.3 Advances in External Beam Radiotherapy

There have been many advances in improving delivery and treatment on top of this basic approach, some of which are briefly described below.

Intensity Modulated RT

IMRT involves the modification of the intensity of radiation across the beam profile to match the shape of the tumour. Beam modulation is achieved by altering the shape of multi-leaf collimators (MLCs). There are several types of IMRT:

- multiple static fields, "step and shoot", where the beam is switched off as the MLCs are changed,
- dynamic MLC, "sliding window", where the beam remains on while the MLCs move, this is faster,
- rotational IMRT (or "volumetric modulated arc therapy" (VMAT)) continuously moves the gantry and the MLCs while the beam stays on,
- tomotherapy is similar to rotational IMRT, but the radiation is delivered in a helical path similar to a CT scan.

IMRT is too complicated to manually place beams until the plan looks acceptable, the delivery protocol must be optimised by a software planning system. In order to optimise a plan there must be a set of criteria for evaluation: usually the Dose and the Dose Volume Histogram (DVH). The Dose Volume Histogram is either differential or cumulative. Differential: what volume receives a particular dose. Cumulative: what volume receives at least a particular dose. An example of using the DVH to prescribe a plan would be requiring that at least 99% of the PTV receives 95% of the prescribed dose, less than 1% of the volume receives 105% of

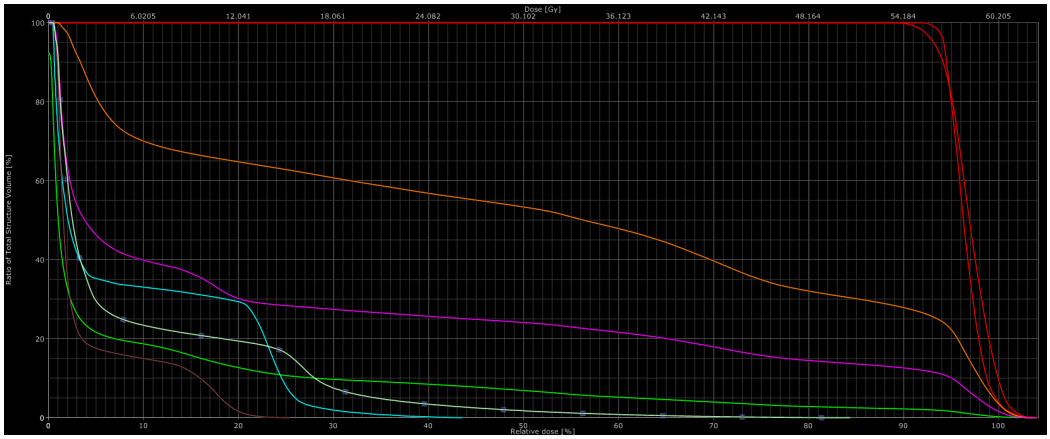


Figure 2.7: An example dose volume histogram taken from a radiotherapy plan used to treat a patient with lung cancer. Each line corresponds to a different region of interest: body (green), GTV (red), spinal cord (cyan), right lung (orange), left lung (brown), whole lung (purple) and the oesophagus (pale green). Note that the x- and y-axes cover the range $[0, 105]$ and $[0, 100]$, respectively. These axes can be read in the electronic copy of this document.

the prescribed dose and the median lies between 99% and 101%. Or another example may be that 95% of the PTV receives 60Gy while no more than half of the rectum exceeds 50Gy. An objective function is then minimised to create a plan which is then inspected for viability. Since mathematical optimality is not always the same as clinical optimality the dose constraints may be modified based on the first iteration of the plan to interactively create a suitable plan. For instance, when a PTV is near the skin a plan may build up dose and undo the skin sparing effect, requiring manual intervention to adjust the planning constraints.

Rotational IMRT

Rotational IMRT (or VMAT) sweeps the beam source through an arc while also changing the MLC shape to conform to the shape of the PTV. Both the gantry speed and the dose delivered during the arc can also be varied in order to meet dose constraints. The use of sweeping beams in VMAT allows for a high dose to be directed to the PTV while sparing surrounding tissues with a short treatment time - a typical VMAT treatment may only take two minutes. However there is more of a dose wash effect with this approach - low dose to a large region, which for example may play a role in the occurrence of radiation induced pneumonitis or in secondary cancer rates in breast cancer patients [1]. The main disadvantage to VMAT lies in the increased planning and verification burden due to the increased complexity of accounting for the changing MLCs and gantry position during irradiation.

Stereotactic Ablative Radiotherapy

SABR uses small fields with a high dose per fraction delivered from many angles to irradiate a PTV and is used to treat particularly small tumours. SABR is most commonly used for treating lung cancer, though it may also be used for treating secondary cancers in other parts of the body. The beams that are used allow for the delivery of a large dose to the PTV while minimising the damage to surrounding healthy tissue as the smaller fields result in a lower dose to these healthy tissues than standard RT. SABR requires fewer fractions (3-8) than other forms of RT and may be administered using a standard linac or by specialised equipment such as the CyberKnife robot, though it typically requires the use of 4DCT for planning.

Image Guided Radiotherapy

IGRT is a field of active research with many different forms and uses. Imaging equipment such as CBCT is used to image patients before radiotherapy. This allows realignment of the patient and a reduction in intra-fraction and inter-fraction alignment errors and consequently an improved delivery of the prescribed dose. Another example of IGRT is to monitor patient position throughout the radiotherapy delivery which allows for the stopping of treatment if a patient moves out of safe boundaries, though this also requires the acquisition of a breathing trace from the patient. An example of a commercial IGRT system that provides these functionalities is the ExacTrac X-ray patient monitoring system which allows for real time verification of patient position before and during radiotherapy.

IGRT can also refer to the use of imaging methods to automatically or semi-automatically re-segment the GTV and OAR in order to verify and possibly adjust the radiotherapy plan at the start of each fraction. This would require the development of accurate, reliable and fast image analysis algorithms to delineate regions of interest in real time. This system would then need to be coupled to a fast dose planning system. This is an ambitious goal and will only be reached in a piece wise fashion by the increased use of imaging to monitor patient treatment and the development of both, more robust and fast image processing frameworks in addition to fast dose planning systems.

2.5 Lung Cancer

In men lun cancer is the leading cause of cancer death and the the most commonly diagnosed cancer. In women it is the fourth most diagnosed and second leading cause of cancer death. Overall, lung cancer has the second highest incidence rate and the highest mortality rate in the world [58]. In 2011 43,463 people in the UK were diagnosed with lung cancer and there were 35,184 reported deaths attributed to lung cancer [31]. The causes of lung cancer include smoking (active and passive), exposure to radon gas, genetic predisposition and contact with certain

substances, for example asbestos. Diagnosis methods include biopsy, functional imaging and structural imaging methods. Once a patient has been diagnosed the treatment is dictated by the disease progression, disease type and the general health of the patient. Treatment options include surgery, chemotherapy, radiotherapy or a combination of these options.

2.5.1 Anatomy

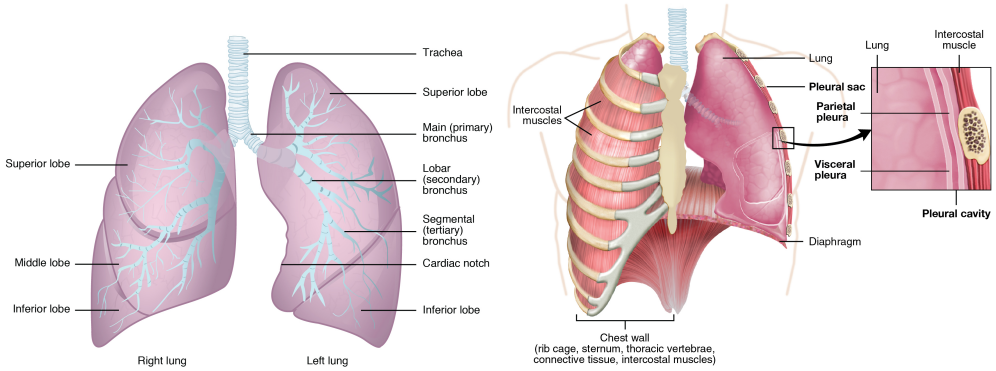


Figure 2.8: Lung anatomy [81].

The lungs are a pair of air-filled organs that reside on either side of the chest. Air passes through the nose and mouth, through the trachea (windpipe) which then splits into branches called bronchi which continue to split throughout the lung, getting smaller and smaller to facilitate efficient oxygen transfer to the blood, the primary function of the lungs. Each lung is divided into lobes, 3 on the right and two on the left. The structure and location of the lungs can be seen in Figure 2.8.

2.5.2 Diagnosis

Lung cancer diagnosis typically begins with a referral from a GP (general practioner). If a patient is exhibiting symptoms (breathlessness, a persistent cough, coughing up blood) they will be referred for a chest x-ray.

Chest X-ray

A chest x-ray is a quick and cheap method used to obtain an initial diagnosis: a single 2D projection of the chest is acquired and analysed. A tumour will be visible as a white/grey mass on the image. However the chest x-ray lacks the discriminating power to tell the difference between a tumour and other problems like a lung abscess which may look very similar on theses images. If there is a possibility lung cancer is present the patient will be referred to a specialist.

CT Scan

The next step after a potential positive result from the chest x-ray is a CT scan (Section 2.2) which allows for a much more detailed view of any potential tumours. A contrast agent may be used to improve the image quality.

PET-CT Scan

Another common step in the diagnosis of lung cancer is to carry out a PET-CT or a PET scan which allows for the determination of the presence and extent of the cancer by measuring metabolic uptake. PET (Positron Emission Tomography) makes uses of a radioactive tracer molecule (such as FDG - fluorodeoxyglucose) that is injected into the patient. This tracer is a molecule that is preferentially taken up by cancerous regions where it then decays to produce a positron which is annihilated by a nearby electron to emit a pair of gamma photons. These two photons are picked up by a ring of detectors to determine their point of origin by measurement of their time of flight. The detection of many photons allows for the reconstruction of a tomographic image.

Biopsy

The final step in a patient's diagnosis may be a biopsy. This allows for a tissue sample to be taken and assessed under a microscope to ascertain the stage and grade of the cancer present. The biopsy may be taken by bronchoscopy if the cancer is near the centre of the chest, otherwise more invasive surgical techniques may be used.

Staging

Lung cancer is divided into two types, small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). Small cell lung cancer is the less common of the two and has only two stages: limited and extensive. The rest of this section will focus on the more common non-small cell lung cancer.

The staging system for NSCLC is a scale of 1 to 4 and is based upon the extent of the cancer in the lung and on its spread to the lymph nodes or other organs. The higher the stage, the further the disease has progressed. Details of the stages are laid out in Table 2.1. After the type (small cell or non-small cell) and stage of disease has been determined an appropriate treatment plan is created for the patient.

Stage 1	Cancer is confined to the lung and has not spread to the nearby lymph nodes. 1A tumour size: <3cm 1B tumour size: 3-5cm
Stage 2	Cancer is larger and may have spread throughout the lungs or to the lymph nodes. 2A tumour size: 5-7cm OR tumour size of <5cm and the cancer is present in the nearby lymph nodes. 2B tumour size >7cm OR tumour size of 5-7cm and the cancer is present in the nearby lymph nodes OR the cancer has spread to muscle tissue but not lymph nodes OR the cancer has made its way to the bronchus OR the cancer has caused lung collapse OR there are multiple small tumours present through the lung.
Stage 3	3A cancerous tissue has spread to either to the lymph nodes or into other surrounding tissues such as the chest wall. 3B cancer has spread to another important part of the body such as the heart or a main blood vessel.
Stage 4	The cancer has spread to both lungs, to another part of the body or there is a build up of fluid containing cancerous cells around the heart and lungs.

Table 2.1: The staging system used for NSCLC. Higher numbers are indicitave of more advanced cancer.

2.5.3 Treatment Procedure

There are several ways to treat NSCLC and the main factors that determine the choice of treatment are the grade of disease, the stage of the disease and the general health of the patient. (SCLC is usually treated with chemotherapy or chemotherapy and radiotherapy together or possibly with surgery if the cancer is diagnosed at a very early stage.)

Surgery

Surgical removal of a whole lung (pneumonectomy), a lung lobe (lobectomy) or a smaller segment of the lung (segmentectomy) are all surgical options depending on the stage of the cancer, the further the disease has progressed the more radical the removal of lung tissue. In order for a patient to be suitable for surgery they must have a good general state of health. Possible complications due to surgery include inflammation of the lung and bleeding.

Chemotherapy

Chemotherapy involves the patient taking a course of drugs (typically cisplatin and etoposide or carboplatin with etoposide) that interfere with cell division processes. Cancer cells grow more quickly and divide at a higher rate than healthy cells. By interfering with the cell division process the cancer cells are killed at a much quicker rate than healthy cells, allowing for the treatment of the cancer. The side-effects vary according to the drugs used and each patient may respond to varying degrees. Common side-effects include hair loss, fatigue, mouth ulcers and nausea. Chemotherapy may be used *a*) post surgery to destroy remnant cancer *b*) alongside radiotherapy to improve treatment or *c*) palliatively to treat the symptoms when there is no curative treatment option.

Radiotherapy

External beam radiotherapy (discussed in more detail in Section 2.4) is the primary radiotherapeutic choice for treating lung cancer, though brachytherapy may also be used in cases where the cancer is accessible by bronchoscopy. Radiotherapy may be used post-surgery and alongside chemotherapy. Some of the side-effects include skin soreness, chest pain, fatigue and difficulty swallowing.

In cases where the tumour is very small, inoperable or the patient declines surgery, SABR (Section 2.4.3) may be used instead of conventional external beam radiotherapy.

2.5.4 Radiation Induced Pneumonitis

Radiation Induced Pneumonitis (a specific instance of radiation induced lung injury) is a condition that develops post radiotherapy in 13-37% of patients [92]. It is an inflammation of lung tissue which leads to impaired respiratory function and in many cases can be the cause of death. The causal mechanisms of radiation induced pneumonitis are not well understood and there has been significant research into the prediction of the occurrence of radiation induced lung injury. Much of the literature focuses on the use of dose and chemotherapy parameters related to the patient's treatment and other clinical factors such as age, tumour location and pre-existing lung conditions. The literature suggests that the V_{45} [ii], V_{40}, V_{30}, V_{20}, V_5 and Mean Lung Dose (MLD) are indicative of the risk of pneumonitis [14, 26, 33, 82, 91, 113, 121]. Particular constraints on the DVH have been shown to negatively correlate with pneumonitis and may be a preventative strategy [59] during treatment.

More recent research has shown that the effective lung dose may be a better parameter than the MLD for grade 3 pneumonitis and suggests that, for a constant MLD, high doses to small volumes carry a larger risk than low doses to large volumes [82]. In addition to these dose

ii. The V_x is the volume that receives at least x% of the dose.

parameters, chemotherapy parameters [33, 55, 56, 82, 91, 96], age [26, 33, 82], sex [33], tumour location [82, 121], smoking history [59, 121], COPD [91, 121] and gross tumour volume [121] have all been shown to be significant in the prediction of pneumonitis. It has also been shown that a given daily dose $\geq$ 2Gy, the V_{20} and a lower-lobe tumour location are predictors of *fatal* pneumonitis [82]. There is also evidence that predictors for pneumonitis may not be consistent between grade 2 and grade 3 radiation induced pneumonitis [33].

In Chapter 5 a method for the prediction of radiation induced pneumonitis by image analysis of planning CT scans is presented. This is a novel method that is capable of predicting whether a patient will develop radiation induced pneumonitis prior to treatment. There has been much research into identifying existing lung disease by image analysis [5, 67, 102, 103, 116, 119] but none into the use of image analysis techniques for the prediction and prevention of radiation induced pneumonitis. The work most closely related to that of Chapter 5 is that of Chen et al, where it was demonstrated that 66 dose and 27 non-dose parameters can achieve predictive performance of Area Under ROC of 0.76 using a neural network [23] or a support vector machine [22].

2.6 Prostate Cancer

Prostate cancer was responsible for up 14% of the total new cancer cases and 6% of male cancer deaths in 2008, it is the second most commonly diagnosed cancer in men and the sixth leading cause of cancer deaths in men worldwide [58]. In 2011, 41,736 men in the UK were diagnosed with prostate cancer and there were 10,793 reported deaths attributed to prostate cancer [32]. The causes of prostate cancer are poorly understood and there exists a variety of methods for diagnosis including prostate-specific antigen (PSA) testing, digital rectal examination (DRE), imaging and biopsy. Once a patient has been diagnosed the treatment is dictated by the severity and progression of the disease. Low risk tumours are often left untreated while a policy of "watchful waiting" is employed to monitor the condition. High risk tumours may be treated with surgery, radiotherapy, chemotherapy, cryosurgery, high intensity focussed ultrasound (HIFU), hormone therapy or some combination of these options.

2.6.1 Anatomy

The prostate is a small gland that is part of the male reproductive system. It is approximately the size of a walnut and is located below the bladder and next to the rectum, surrounding the urethra. The main function of the prostate is to secrete a fluid which makes up the majority of the volume of semen. The location and relative size of the prostate can be seen in Figure 2.9. There are four zones in the prostate: the peripheral, central, transition and anterior fibromuscular zones. Of these the peripheral zone is the largest and also the most common origin for prostate cancer.

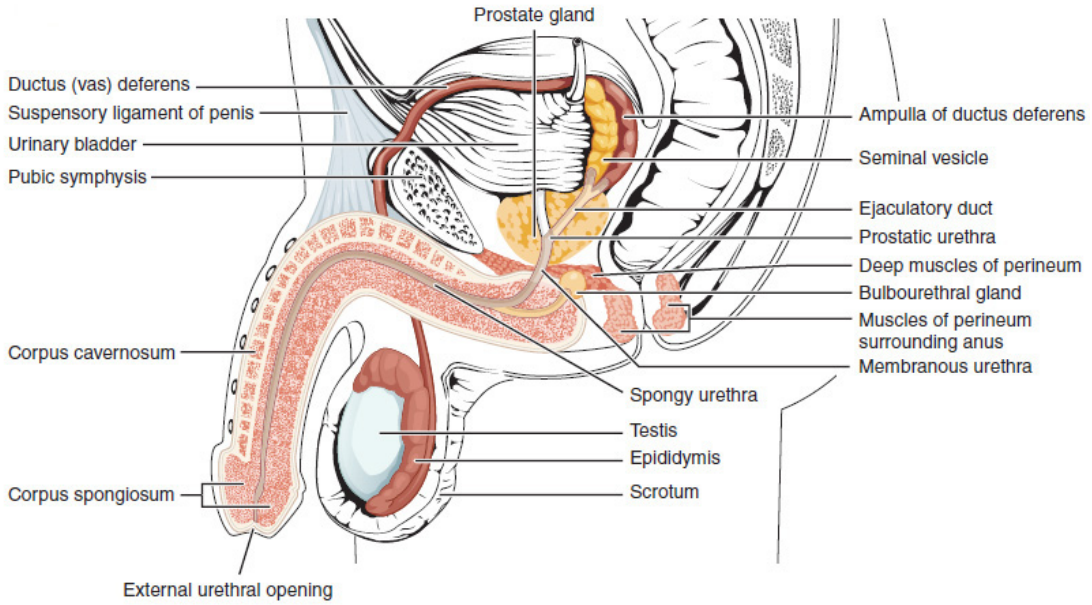


Figure 2.9: Prostate anatomy [81].

2.6.2 Diagnosis

Diagnosis of prostate cancer is often difficult as the symptoms can also be the result of benign enlargement of the prostate and this is why several complementary tests are used for diagnosis, usually beginning with PSA.

PSA

PSA is a protein produced by the prostate gland and its levels can increase in the presence of cancer tissue. It can therefore be used as a marker for prostate cancer, although PSA tests have a relatively poor sensitivity/specificity profile [54]. There is currently no PSA screening in the UK as the costs and risks are thought to outweigh the benefits [42] due to the large number of false positive results.

DRE

The next step after a PSA test is a digital[iii] rectal examination (DRE) by a GP, which examines the surface of the prostate for any abnormal changes. A tumour will often make the prostate hard and bumpy while benign prostate enlargement is more likely to feel firm and smooth. Therefore, the primary benefit of DRE is to rule out benign prostate enlargement that comes naturally with age.

iii. Here digital is used in the sense of relating to a finger.

Biopsy

If a patient is deemed to be at risk of prostate cancer following PSA testing, DRE and after taking into account age, family history and ethnicity they will be referred to a hospital for a biopsy test. Most commonly this test is done by transrectal ultrasound (TRUS) guided biopsy. The TRUS biopsy allows samples of the prostate to be taken for further histological testing and assessment (preferably at least five cores are taken from each prostate lobe). However, the ultrasound only permits the clinician to identify regions or geometric position within the prostate, it does not enable them to see the tumour. This means that several samples must be taken at various points in the prostate and up to 1 in 5 cancers may be missed by this test.

Grading

After biopsy the extracted tissues are examined under microscopy to determine the presence of cancer cells and the Gleason score of any cancer found. The Gleason grade is a measure of the aggressiveness of *cells*. The Gleason score is calculated by the summation of two Gleason grades: the most common Gleason grade within a biopsy sample and the highest Gleason grade of the rest of the cells. The scale used for each grading ranges from 1-5. Grades 1 and 2 are not cancer so scores will be in the range 6-10. A score of 6 would indicate that all the cells are likely to grow slowly whilst a score of 10 that all of the cells are likely to grow quickly. Appropriate treatment options are influenced by the Gleason score.

Imaging

MRI scans may be used after PSA testing and the grading of biopsy results. The MRI allows for a determination of the extent of the cancer spread and if it is used, the scan is acquired once surgery or radiotherapy has been prescribed. A CT scan or bone scan (nuclear medicine) may also be used to check for the spread of cancer to the patient's bones.

T Stage

The T stage is a measure of the size of the primary tumour. There is a simple scale, **TX**: the tumour cannot be measured, **T0**: the primary tumour cannot be found, **Tis**: cancer cells are growing only in the most superficial layer of tissue and **T1-T4**: where the number indicates the size and spread of the cancer into surrounding tissues; the higher the number, the greater the size and spread.

Patients finally diagnosed with prostate cancer can be divided into two main groups: those with low risk cancer and those with high risk cancer. The former group (low PSA levels, grade and T stage) are usually left untreated and carefully monitored for any change in symptoms and PSA tests. However some men may elect to undergo treatment at this stage rather than wait.

The patients with high risk cancers will undergo treatment to deal with the cancer as outlined in the next section.

2.6.3 Treatment Procedure

Treatment options for prostate cancer depend upon the extent and aggressiveness of the disease in each patient. Men with cancer that is determined to be low risk or who are considered unlikely to require treatment within their predicted lifetime are often not treated but instead monitored regularly in a period of "watchful waiting". During this time there are regular DRE and PSA tests and at the first sign of change in the cancer the patient will be considered for further treatment. Once it is established that the patient is to be treated there are a number of options available; hormone therapy; prostatectomy; chemotherapy; cryosurgery; high intensity focused ultrasound; brachytherapy; and external beam radiotherapy.

Hormone Therapy

Hormone therapy is used to shrink the tumour by inhibiting the presence or the affect of testosterone in the prostate. Testosterone influences the size of the prostate and also of prostate cancers so a reduction in testosterone will lead to a shrinking of the tumour and reduced symptoms of prostate cancer. This treatment is useful when the tumour is protruding from the prostate gland but has not yet reached any other structures such as the lymph nodes or bones. The treatment can be administered by LHRH (Luteinizing-hormone-releasing hormone) analogue injection which lowers the testosterone levels in a temporary and reversible manner. The injections are administered either every month or every 3 months and patients can be taught to inject themselves. Another option is anti-androgen tablets which either lower testosterone levels or inhibit its function. These may be taken alone or in combination with LHRH analogue injections. A third option is an orchidectomy: a one off surgical procedure to remove the parts of the testicles that produce testosterone. This is irreversible, unlike hormone injection or anti-androgen tablets. Whichever method is chosen the common side-effects include hot flushes, a loss of sexual desire, impotence and occasionally breast tenderness or enlargement. Ultimately, hormone therapy does not provide a cure for prostate cancer but is an effective method for managing the disease as it can keep the cancer in check for a number of years. It may also be used in conjunction with radiotherapy or other treatments to improve patient outcome.

Prostatectomy

Prostatectomy is the surgical removal of the prostate and the cancerous tumour. It has the advantage that the true stage and grade of the cancer can be determined post surgery. If the cancer is confined to the prostate the entire removal of the gland is curative, otherwise further treatment may be required, such as external beam radiotherapy. PSA levels drop to zero after surgery and will only increase again if the disease returns allowing for early detection of recurrent disease. The removal of the enlarged prostate should also help alleviate urinary problems caused by cancer obstructing the flow of urine from the bladder. Potential side-effects of surgery include wound discomfort, a short period of incontinence and impotence. Surgery is best suited for early stage prostate cancer and in order to be eligible for surgery patients must be fit for anaesthetic and have a Gleason score of 8 or less.

Chemotherapy

Chemotherapy is usually considered after hormone therapy failure and is often used in later stage prostate cancers with continuing hormone therapy or other treatments like radiotherapy. By itself, chemotherapy is not always curative, but is used to reduce symptoms and slow cancer growth. It involves the patient taking a course of drugs (typically docetaxel, mitoxantrone, paclitaxel or cabazitaxel) that interfere with cell division processes. Cancer cells grow more quickly and divide more often than healthy cells, so by interfering with the cell division the cancer cells are killed at a much quicker rate than healthy cells, thus allowing for the treatment of the cancer. The side-effects vary according to the drug used and each patient may respond to varying degrees. Common side-effects include anaemia, low immune resistance, bruising or bleeding, loss of appetite, nausea, hair loss, fatigue and diarrhoea. Because of these side-effects a patient must be relatively fit before treatment.

Cryosurgery

Cryosurgery is used when the prostate cancer is locally advanced and has not spread outside of the prostate gland. In the UK it is usually used in the case of recurrent prostate cancer after radiotherapy, sometimes hormone therapy is required beforehand to shrink large prostates before surgery. A cryogenic gas, such as argon, is injected through a number of needles placed into the prostate gland through the skin and guided by ultrasound. During the surgery the prostate temperature is lowered to $-140°C$ while the surrounding area is maintained at a safe temperature. The very low temperature kills any prostate and cancer cells. The advantages of cryosurgery are that it is minimally invasive, is repeatable and can be used when patients are not suitable for major surgical procedures. The main disadvantages are impotence, short term incontinence and damage to the rectum in a small number of cases.

High Intensity Focused Ultrasound

HIFU involves the use of a rectal ultrasound probe: high intensity ultrasound waves are focussed on the prostate to generate localised intense heat with minimal damage to surrounding areas. The procedure is carried out under general anaesthetic and can take up to three hours. HIFU is a relatively new technique and is currently under active research.

Brachytherapy

Radioactive (I^{125}, half life of approximately 60 days) seeds are implanted into the prostate using a series of needles guided by ultrasound to deliver a therapeutic dose to the prostate and spare surrounding regions. Brachytherapy is best suited to cases where the cancer is confined to the prostate (T1 or T2 stage), there is a Gleason score of 7 or less and the prostate is relatively small. As with many of the other treatments, urinary problems and impotence may arise post treatment and radiation protection precautions are necessary around small children and pregnant women for two months post implantation.

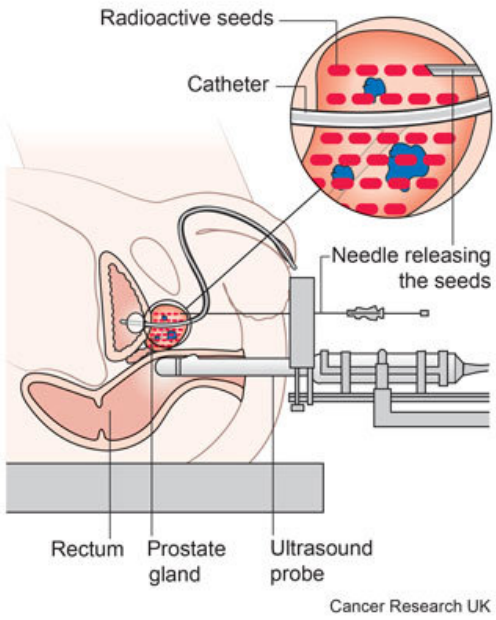


Figure 2.10: Diagram showing the typical administration of brachytherapy for a prostate cancer patient. Image from [112].

External Beam Radiotherapy

External beam radiotherapy, details of which can be found in Section 2.4, is a common treatment choice for prostate cancer. Typical treatment consists of hormone therapy followed by radiotherapy: at the Edinburgh Cancer Centre a patient undergoes three months of hormone treatment, followed by the implantation of fiducial markers and then the administration of 3 or 4 field conformal radiotherapy. IMRT may be used in some circumstances where dose constraints cannot otherwise be met. Some of the short term side-effects include fatigue, urinary and bowel problems, nausea and skin soreness. Longer term side-effects may include urinary problems, bowel problems and impotence.

State of the Art Developments

The Efinger project is researching a "multi-scale mechanical measurements" approach for the diagnosis of prostate cancer [47]. The project is seeking to develop a minimally invasive probe (transurethral, transrectal or laparoscopic) to measure the mechanical properties of the prostate in vivo. These mechanical properties are indicative of the histological properties of the prostate and can therefore be used in diagnosis.

The da Vinci surgical robot is a new technique for minimally invasive surgical procedures, including prostate cancer [74]. The system combines robotics and 3D vision systems to allow surgeons to perform much less invasive surgeries resulting in faster recoveries and fewer complications.

2.6.4 Focal Lesion Boosting

External beam radiotherapy, like the other methods outlined in Section 2.6.3, treats the whole of the prostate gland in all cases, regardless of the stage of disease. Treating the whole gland on low risk patients means that a therapeutic dose is delivered to the tumour but the healthy tissue in the prostate is also subjected to a large dose. There is potential for improving patient treatment when there is a dominant cancer focus in the prostate. That is, when the prostate cancer is located in a single dominant focus (a focal lesion) it may be possible to "boost" the focal disease. This boosting would involve administering a larger dose to the focal disease and this may allow for the overall reduction in dose to the surrounding (healthy) prostate tissue. However "focal boosting" is not in widespread use due to the challenges that must be overcome to implement a viable, safe and effective clinical protocol.

Challenges & Current Research

The challenges faced by focal therapy are primarily identifying the focal disease, developing effective treatment plans and determining the efficacy of this approach versus the conventional whole gland treatments.

Identification: There is great difficulty in manually identifying regions of focal disease in the prostate. A range of modalities are used including PET/CT, SPECT and single parameter MRI. The process for each of these modalities is very time consuming and requires a great deal of clinical expertise. There has been some preliminary research into the use of machine learning for this kind of identification. Shah et al [97] looked at localising peripheral zone prostate cancers on multi-parametric MRI with histological gold standard information and achieved an f-measure of 89% with a support vector machine. Groenendaal et al [48] considered a logistic regression model for the voxelwise prediction of peripheral prostate cancer on diffusion-weighted imaging, dynamic contrast-enhanced magnetic resonance, achieving an area under ROC of 0.89 when using a histological gold standard for evaluation. Chapter 6 outlines an automated approach for the generation of focal disease contours on T2 MRI. One advantage of the approach presented in this work is that a contour is generated in a fully automated process unlike the methods laid out above which only present a classification of the voxels and do not proceed to generating a contour of the disease. There remains work to be done in transferring this identification from MRI to the radiotherapy planning CT to allow for this information to inform treatment planning.

Treatment Plans: After the identification of the focal disease a radiotherapy plan must be designed that will suitably irradiate the disease while still maintaining the prescribed dose constraints for other organs at risk. This may not always be possible, depending on the focal disease location.

Efficacy: There are no firm clinical conclusions about the efficacy of focal prostate therapy. There are ongoing trials to determine if this should be another option for the treatment of prostate cancer. A recent review, [7], concluded that:

> "conclusions regarding optimal techniques and/or efficacy of this approach are elusive, and this approach cannot be considered standard of care. There is a need to build consensus and evidence. Ongoing prospective randomized trials are underway and will help to better define the role of differential prostate boosts based on imaging defined GTVs."

2.7 Summary

This chapter has laid out a brief introduction to x-ray CT and MR imaging, radiotherapy and prostate and lung cancer and the challenges of treating these diseases. This basic understanding of the problem is the first step in developing tools and systems to aid the clinicians treating patients with cancer. CT imaging, MR imaging and radiotherapy are all used for the treatment of many different cancers. This chapter has focussed on lung cancer and prostate cancer. In Chapter 5 a methodology for the automatic prediction of radiation induced pneumonitis is presented and in Chapter 6 a framework for the segmentation of prostate focal disease is demonstrated. It is hoped that both of these contributions may begin to address at least some of the challenges outlined earlier in this chapter.

Chapter 3

Texture and Medical Image Analysis

> "the feel, appearance, or consistency of a surface or a substance" [i]

3.1 Introduction

Texture is a property of a surface that is easily recognised but is an incredibly difficult concept to define and characterise. Almost all surfaces have *natural texture* caused by the physical variation of the profile of the surface. *Visual texture* is caused be the aforementioned physical surface texture (like variations in roughness) or by differences in reflectivity due to colour changes on a surface. It is this *visual texture* that image analysis is concerned with. When images are considered, texture is the variation of pixel intensities as a function of position. A way to visualise the texture of an image is to use the pixel location as x and y coordinates and the grey level at each location as the z coordinate to create a surface. A large amount of significant information can be extracted from the texture within an image which has led to many applications of texture analysis in the last half century. A taxonomy of common texture analysis methods is presented in Section 3.3 and short review of example applications of texture, with a particular focus on medical imaging, in Section 3.3.6.

There is no agreed upon definition of texture in the field of image analysis and computer vision. Some definitions are arrived at from a conceptual approach while others are derived from the practical application at hand. This lack of a proper definition emphasises how broad a concept texture can be and has led to the large number of different methods used to characterise texture in different circumstances. A good, though broad, definition of texture is provided by Petrou:

> "Texture is the variation of data at scales smaller than the scales of interest."[88]

This definition neatly captures the nature of texture, though it does not move one closer towards a unified theory of texture. Nevertheless this will be the working definition used in this thesis and chapter. An example that illustrates this definition: if one is attempting to identify organs on an MR scan, the texture is the variation of the grey levels within each of the organs that is then characterised in order to differentiate between the organs. That is, if one is attempting

i. Oxford English Dictionary

to segment the bladder and the prostate on a MR scan, the texture is the grey level variations within the bladder and within the prostate caused by the prostate and bladder being composed of different tissues which appear differently on an MRI.

Texture in an image can also be problematic when low level vision methods are used. For example, edge detection algorithms can run into issues when extra edges within an object are detected due to the texture of an object. Returning to the MR example above: the texture of different regions within the prostate may reduce the quality of a segmentation of the whole prostate. (Another example of this, illustrated by a block of marble, can be seen in Petrou [88].)

The remainder of this chapter will cover the role of texture in the human visualisation process, present a taxonomy of texture approaches and applications, discuss the nature of texture in three dimensions and then offer a more detailed discussion of several texture measures that are used in this work.

3.2 Relation to Human Visualisation

The interaction between human vision and texture is important because most texture analysis algorithms will be evaluated against a human baseline. Yet this may not always be the case, as shown by the work in Chapter 5 where texture is used to detect information that is undetectable by the human visual system. A further example would be the texture based classification of a disease evaluated by a histological gold standard. Texture is central to human vision as it affords clues as to the composition of object surfaces and it is used in pre-attentive recognition [60]. (Pre-attentive recognition is the ability to unconsciously recognise an object or pattern by looking at it).

Julesz postulated that the human visual system cannot discriminate between textures with different third order statistics but with identical first and second order statistics [61, 62] (defined in Section 3.6). This raises the question of how the performance of a texture algorithm is measured. An algorithm which can discriminate between textures that the pre-attentive human visual system cannot may outperform humans in some tasks, or it may find too many different regions if the aim is to emulate the performance of humans. However, in some cases where the second order statistics between pairs of textures are identical it is still possible to pre-attentively discriminate the pairs. This led to the discovery that the texture discrimination is based solely on the density of “textons” [63]. The textons identified by Julesz were “color, elongated blobs (line segments) of given width, orientation, and length, and the terminators (end-points) of these elongated blobs”.

Understanding biological systems can inspire algorithmic development like artificial neural networks or in the case of texture, multichannel filtering approaches: the demonstration that the

brain converts retinal images into filtered images of various frequencies and orientations [20] has led to the investigation of signal processing techniques for texture analysis.

In summary, because texture is an inherently visual conception texture analysis is closely related to and influenced by our understanding of the human visualisation system.

3.3 Taxonomy of Approaches

There are a variety of approaches to texture analysis reflecting the different problems being addressed but also because of the lack of a unified framework for understanding and characterising texture, as explained earlier. Because of this lack of a single framework there are a range of different approaches that have been applied to the field of texture analysis. This section presents texture analysis methods belonging to the standard classes. Statistical, geometric, model based and signal processing methods are discussed.

3.3.1 Statistical

Statistical measures attempt to implicitly capture the textural properties of a region by measuring the distribution of grey level values. Statistical measures were among the first to be developed for texture analysis and many different descriptors are available. Many of the features that can be calculated by statistical methods are correlated with each other and so a feature reduction/selection method is usually employed when these features are used. Examples of statistical methods include first order statistics, grey level co-occurrence matrices, grey level run length matrices, grey level size zone matrices (each of which will be further elaborated upon later on in this chapter) and autocorrelation features. The autocorrelation function measures the self-similarity of a region and so is a measure of the regularity and fineness/coarseness of a region.

3.3.2 Structural/Geometric

Structural (or geometric) approaches define texture as being made up of small, well-defined primitives that are placed according to a set of rules. This kind of formulation is well suited to texture synthesis as it allows for the creation of deterministic and regular stochastic textures but is not as well suited to characterising naturally occurring texture as the primitives or the rules may be difficult to predict and establish reliably. Morphological methods also fall into this class of structural methods. Morphological methods use mathematical morphology operations to determine textural characteristics. An example of a morphology based method is opening an image with progressively larger structuring elements to determine how many pixels in the image are removed at each scale, this gives a pattern spectrum which can be used to describe the texture in an image.

3.3.3 Model Based

Model based techniques seek to parameterise the texture according to a theoretical model. For example a Markov random field (MRF), an autoreggressive model or a fractal model. Taking the MRF model as an example: an MRF is defined as a field in which the value of a point depends directly on only the values of the neighbouring pixels. In order to use an MRF to characterise texture, a clique (or neighbourhood of each pixel) must be defined in addition to a function which used to calculate the probability of a pixel being a certain value given the value of its neighbours in the clique. The parameters of this function are used to characterise texture. Once these parameters have been learnt, the MRF may be used to synthesise texture that looks similar to the original. Additionally completely synthetic texture may be generated by the use of MRFs by manual selection of the model parameters.

3.3.4 Signal Processing

Texture measures based on signal processing techniques include spatial domain filters, the Fourier transform and extensions of these algorithms. Spatial domain methods include using edge detection algorithms to count the density of edges in a region and moment based features. The $(p+q)^{\text{th}}$ moments over an image, I, are defined according to

$$m_{pq} = \sum_{(x,y)\in R} x^p y^q I(x,y) \tag{3.1}$$

If R is a rectangle and the moments are calculated for every pixel in the image, then the moment based approach is equivalent to applying a series of spatial filters to the image to generate features at each pixel. More common are frequency domain approaches: Fourier transform methods perform poorly as they lack any localisation terms and hence are restricted to measuring global properties. However Gabor filters and wavelets have widespread applications in texture analysis due to their ability to characterise local frequency responses. Gabor filters will be further explained in Section 3.9.

3.3.5 Stationary/Non-Stationary Texture

A stationary texture image has only one type of texture present, while a non-stationary image has more than one type of texture. Typically an image with stationary texture is classified and a label is applied to the whole image, while images with non-stationary texture are segmented to delineate the structures or regions containing the different texture in the image. Different texture analysis methods are used for stationary and non-stationary images: FOS, GLCM and GLRLM[ii] are each global measures and are used for stationary texture images while local measures like Gabor filters and Linear Binary Patterns (LBPs) are used for non-stationary

ii. These are defined later in the chapter.

texture cases. However global methods can be transformed into local methods by applying them on local sub-regions of an image to generate local texture descriptors and localised methods can be used as global measures by averaging or computing a global statistic of local texture features.

3.3.6 Applications in Medical Image Analysis

Texture analysis has been used widely in the literature over the course of a number of years in the medical imaging field. This section provides a short overview of some recent applications in the field.

Korfiatis et al [65] demonstrated that 3D co-occurrence features are capable of classifying interstitial lung disease - lung parenchyma was separated into three classes: normal, ground glass and reticular. Linear Binary Operators have been used for content-based image retrieval in [17]. Diagnosis of liver tumour disease with an accuracy of 94% was demonstrated by [66]. Chicklore et al explored the use of texture features in functional imaging (PET) and concluded that, although more work is required, texture may play a useful role in the analysis of images from functional modalities [24]. Texture has also been shown to be useful in the characterising of structural changes during radiotherapy in head and neck cancer patients [93]. For a comprehensive review of three dimensional texture in medical imaging, the reader is directed to the excellent review by Depeursinge et al [36]. This review explores various uses of texture across a range of modalities and patient organs. More recently, Prasanna et al [89] have developed a novel technique and demonstrated its ability to distinguish between "two subtly different types of pathologies on MRI in the context of brain tumours and breast cancer".

This short collection of example applications demonstrates the potential for the clinical utility of texture based approaches and this author believes that texture will continue to be a widely researched topic in the coming decade.

3.4 Texture Analysis in Three Dimensions

The extension of texture analysis from two dimensions to three is in some ways a natural one with the ever increasing availability of volumetric imaging data, particularly in the medical imaging arena. However it is not a straightforward development, partly due to the increased visualisation and conceptual burdens, the non-trivial extension of 2D algorithms and also the increased storage and computational demands imposed by this new data. The visualisation and conceptual burden can be demonstrated by considering the visualisation of a 2D image as a surface as mentioned above: the grey level values providing the z coordinates. If this process was to be repeated in 3D we would need a 4D volume to represent the grey level at each x, y, z position. Initial attempts at 3D texture analysis were pseudo-3D methods: applying 2D

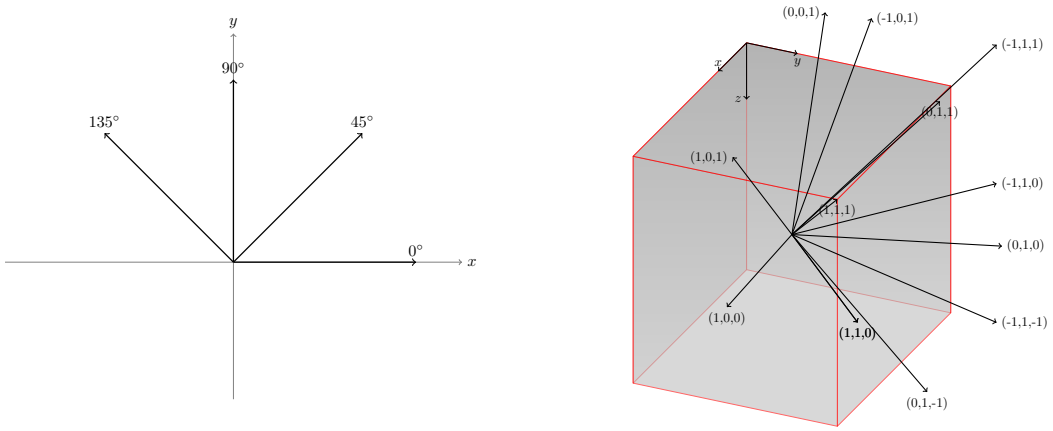


Figure 3.1: Diagrams showing the 4 directions in which texture is usually calculated in 2D (right) and the 13 directions used in 3D (left).

techniques in a slice-wise fashion or by projecting a 3D image onto a plane before calculating a 2D texture descriptor. Neither of these methods come close to fully exploiting the information that is gained by moving from a 2D to 3D interpretation in the presence of isotropic imaging data. (If $\Delta z >> \Delta x, \Delta y$, where Δx, Δy and Δz are the pixel spacings in the x, y and z directions, then it is necessary to use a slice wise approach.) Volumetric texture analysis is a relatively new field largely because isotropic volumetric imaging data has not been widely available until recently. Given the large amount of research and wide applications that have come from 2D texture, it is expected that 3D texture research will continue to expand and hopefully make a large contribution to the field of medical image analysis.

As an example of the advantage of moving from 2D to 3D texture [109]: if co-occurrence or run length features are to be calculated at each point in a image, as may be desirable in a segmentation task, it is necessary to calculate the texture on a small region around each pixel in the image. This small region must be large enough for the features to be statistically relevant, but if the region is too large this will act as an averaging/blurring effect over the calculated texture. The advantage of 3D (over 2D) texture in this situation is that the same number of pixels may be included while remaining spatially closer to the central pixel. For example a square of length 5 has only 25 sample pixels while a cube of length 5 contains 125 pixels for analysis.

However the large number of pixels available in 3D also brings challenges in capturing all of the information present in a volume. Since images are quantised to a grid the standard method for calculating texture in various directions is to examine the space in 4 directions in 2D and 13 unique directions in 3D. These are calculated by considering a 3×3 or a $3 \times 3 \times 3$ volume and can be seen in Figure 3.1. As the size of the region increases, the number of pixels not

considered also increases. In 3D this effect greater than in 2D, as can be seen in Figure 3.2. This graph clearly displays how many more pixels are discarded in 3D for larger box sizes. This suggests that for smaller volumes GLCMs and similar methods may be more appropriate whereas for larger volumes signal processing methods may be better suited as they do not suffer from this "structuring element" problem.

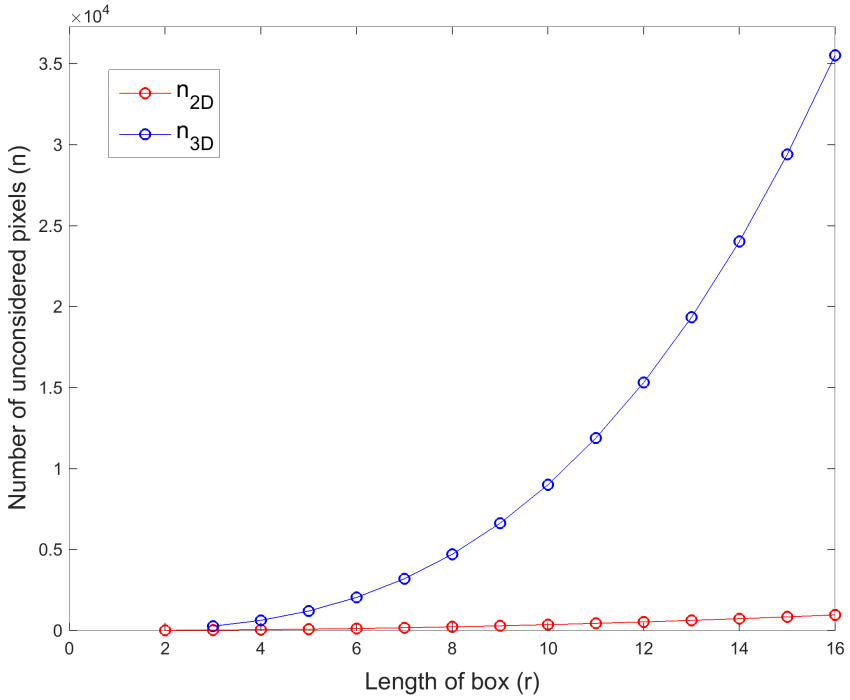


Figure 3.2: Visualisation of how the number of pixels not considered by the usual 4 directions (2D) and 13 directions (3D) varies with box size. The curves in the graph are given by $n_{2D} = d^2 - 4(d-1) - 1$ and $n_{3D} = d^3 - 13(d-1) - 1$.

Most of the texture descriptors described in this chapter were originally described in 2D and have been extended to 3D. The suitability of their extension to 3D and the "naturalness" of their three dimensionality are discussed within each of their sections.

3.5 First Order Statistics

First Order Statistics are derived from the image histogram (or grey level probability) which means that no higher order information about the spatial relation between pixels is taken into account, hence the name. This means that first order statistics are trivial to compute, but have limited discriminatory power. This can be seen in Figure 3.3: each image has identical first order statistics but because only the grey level probability is used to calculate the features the (noisy) chequerboard cannot be distinguished from the union jack.

The seven first order statistics are calculated according to the following equations:

$$\text{Mean, } \mu = \frac{1}{N}\sum_{i=1}^{N} x(i) \tag{3.2a}$$

$$\text{Variance, } \sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2 \tag{3.2b}$$

$$\text{Coarseness} = 1 - \frac{1}{1+\sigma^2} \tag{3.2c}$$

$$\text{Skew} = \frac{\frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^3}{\left(\frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2\right)^{\frac{3}{2}}} \tag{3.2d}$$

$$\text{Kurtosis} = \frac{\frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^4}{\left(\frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2\right)^{2}} \tag{3.2e}$$

$$\text{Energy} = \sum_{i=1}^{N} x_i^2 \tag{3.2f}$$

$$\text{Entropy} = -\sum_{x} P(x) log(P(x)) \tag{3.2g}$$

$$\tag{3.2h}$$

where $P(x)$ is the image histogram.

First order statistics are often included in many texture analysis studies and combined with more advanced features as they are both trivial to compute and may contain information complementary to other texture descriptors and in the case of obviously distinguishable textures the mean and entropy are often useful features. It is for these reasons that they are also included in this work.

3.6 Grey Level Co-occurrence Matrices

Grey Level Co-occurrence Matrices (GLCM) are used to measure the second order statistics present in an image by considering the relationship between pairs of pixels in a given direction at a given distance.

A rotationally invariant GLCM may be calculated by considering all pairs of pixels at a given distance. These can be more appropriate when the texture in an image is rotationally invariant. Another motivation for the use of rotationally invariant descriptors is in scenarios where the images of textured objects have been acquired from different perspectives and viewing angles. This is not the case for the medical images considered in this study. Patients were all placed

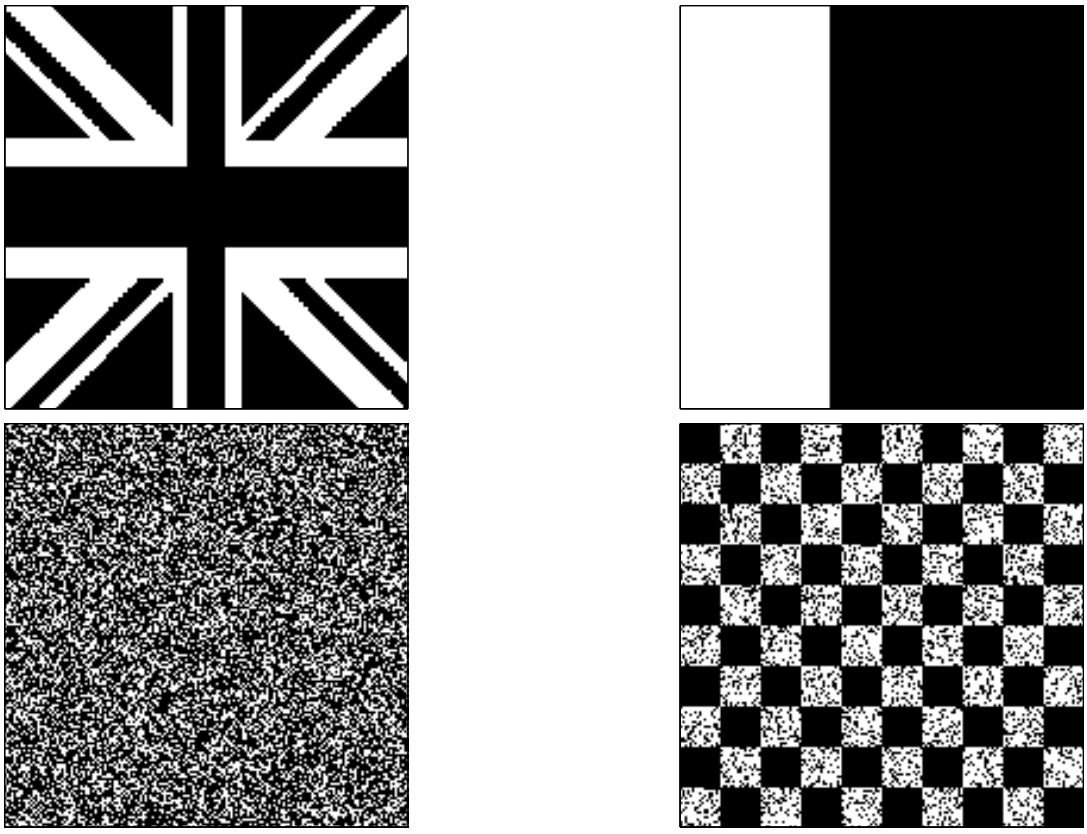


Figure 3.3: Four example images with identical first order statistics - a system based only on FOS would be unable to discriminate between each of these images despite their very obvious differences.

in a head first supine position which aligns all the patients in the same, or very similar, frame of reference before image acquisition. (This rule would apply to most CT and MR imaging as a group of scans are likely to share a common frame of reference even though MR can be acquired in any arbitrary plane. However this may need to be reconsidered for ultrasound imaging where the probe is free to move and images will not normally be aligned.) For this reason the rotationally invariant GLCM was not considered. Instead the rest of this section focuses on GLCMs that incorporate the directional information as it should allow the extraction of more information from textured regions.

In 2D a GLCM is formally defined as the number of occurrences of a pixel with grey-level, i, at a distance, d, and angle, α, from a pixel with grey-level, j, in an image, I ($\Delta x = d\cos\alpha$, $\Delta y = d\sin\alpha$):

$$GLCM_{d,\alpha}(i,j) = \sum_{p}\sum_{q}\begin{cases}1, & \text{if } I(p,q) = i \text{ and } I(p+\Delta x, q+\Delta y) = j\\ 0, & \text{otherwise}\end{cases} \tag{3.3}$$

It is possible to generate as many GLCMs as there are combinations of angles and distances

1	1	2	2	4
1	1	2	2	2
1	3	3	3	1
3	3	4	4	3

idx	1	2	3	4
1	6	1	3	0
2	1	4	2	1
3	3	2	2	2
4	0	1	2	0

Figure 3.4: Example of a 4×4 GLCM (right) calculated from the matrix on the left - the values of the grey levels are displayed in the matrix.

in an image. Since images are made up of values on a discrete grid the available distances and angles are quantised (unless interpolation is used): for a distance of one, there are 8 neighbouring pixels which results in 8 directions (0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°) from which to choose. However GLCMs are symmetric matrices as can be trivially deduced from Equation 3.3 so the 0° GLCM is equivalent to the 180° GLCM, 45° to 225°, 90° to 270° and 135° to 315°. This results in only 4 unique directions in two dimensions for a distance of one, see Figure 3.1 for an illustration of these. When the distance is increased, however, there are more neighbours to consider. These neighbours can be found by finding all the pixels that satisfy, $d - 0.5 \leq \sqrt{\Delta x^2 + \Delta y^2} < d + 0.5$, where d is the distance being considered. This leads to 6 unique directions for $d = 2$, 8 unique directions for $d = 3$, etc. In order to keep the number of GLCMs reasonable, only the four unique directions from the $d = 1$ case have been considered in this work when $d > 1$.

Equation 3.4 shows the extension of co-occurrence matrices to 3D. The GLCM is still two dimensional but is now parameterised by an additional angle, β.

$$GLCM_{d,\alpha,\beta}(i,j) = \sum_r \sum_p \sum_q \begin{cases} 1, & \text{if } I(p,q,r) = i \text{ and } I(p+\Delta x, q+\Delta y, r+\Delta z) = j \\ 0, & \text{otherwise} \end{cases} \tag{3.4}$$

The extension to three dimensions increases the number of directions that must be considered. For $d = 1$ there are 13 unique directions, as illustrated in Figure 3.1. As with the 2D case, only these 13 directions will be considered for $d > 1$, the effects of discarding pixels as d increases was discussed in Section 3.4.

GLCMs produce a very rich representation of an image and its texture but can be very unwieldy and bulky as they have $(\text{Number of Grey Levels})^2$ elements. One method for extracting features to characterise the texture is to use the ratio of elements in the GLCM which expresses the relative abundance of co-occurrences in the image. The problem with this approach is the large number of features it can generate and the inability to know *a priori* which of these features will be significant for a given task. These features must be tuned on a per task basis. The more frequently used method for feature extraction was proposed by Haralick [50], where 14

characteristics of the GLCM are computed and used as features:

$$\text{Angular Second Moment} = \sum_i \sum_j p(i,j)^2 \tag{3.5a}$$

$$\text{Contrast} = \sum_{n=0}^{N_g-1} n^2 \left\{ \sum_i \sum_j p(i,j) \right\}_{|i-j|=n} \tag{3.5b}$$

$$\text{Correlation} = \frac{\sum_i \sum_j ij - \mu_x \mu_y}{\sigma_x \sigma_y} \tag{3.5c}$$

$$\text{Sum of Squares: Variance} = \sum_i \sum_j (i-\mu)^2 p(i,j) \tag{3.5d}$$

$$\text{Inverse Difference Moment} = \sum_i \sum_j \frac{1}{(i-j)^2+1} p(i,j) \tag{3.5e}$$

$$\text{Sum Average } (f_6) = \sum_{i=2}^{2N_g} i p^{x+y}(i) \tag{3.5f}$$

$$\text{Sum Variance} = \sum_{i=2}^{2N_g} (i-f_6) p_{x+y}(i) \tag{3.5g}$$

$$\text{Sum entropy} = -\sum_{i=2}^{2N_g} p_{x+y}(i) log(p_{x+y}(i)) \tag{3.5h}$$

$$\text{Entropy}(f_9) = -\sum_i \sum_j p(i,j) log(p(i,j)) \tag{3.5i}$$

$$\text{Difference Variance} = var(p_{x-y}) \tag{3.5j}$$

$$\text{Difference Entropy} = -\sum_{i=0}^{N_g-1} p_{x-y}(i) log(p_{x-y}(i)) \tag{3.5k}$$

$$\text{Information Measure of Correlation 1} = \frac{H_{xy} - H_{xy1}}{max(H_x, H_y)} \tag{3.5l}$$

$$\text{Information Measure of Correlation 2} = (1 - exp(-2(H_{xy2} - H_{xy})))^{0.5} \tag{3.5m}$$

$$\text{Maximal Correlation Coefficient} = (\text{Second largest eigenvalue of } Q)^{0.5} \tag{3.5n}$$

where N_g = number of grey levels, $p_x(i) = \sum_j p(i,j)$, $p_y(i) = \sum_i p(i,j)$, $p_{x+y}(k) = \sum_i \sum_j p(i,j)$, $_{i+j=k}$ $p_{x-y}(k) = \sum_i \sum_j p(i,j)$, $_{|i-j|=k}$ $\mu_x = mean(p_x)$, $\mu_y = mean(p_y)$, $\sigma_x = std(p_x)$, $\sigma_y = std(p_y)$, $H_{xy} = f_9$, $H_x = entropy(p_x)$, $H_y = entropy(p_y)$ and $Q(i,j) = \sum_k \frac{p(i,k)p(j,k)}{p_x(k)p_y(k)}$. This method of computing a set of descriptive statistics is appropriate because the GLCM can be thought of as a joint probability density function, and probability density functions can be well characterised by computing a set of statistics to describe them.

3.7 Grey Level Run Length Matrices

Grey Level Run Length Matrices (GLRLMs) count the number of "runs" in an image in a particular direction, a run is a series of collinear pixels of the same value (or which lie in a range of values). This can be thought of as a higher order approach than GLCMs - instead of looking at pairs of pixels the number of runs in a image are counted for each grey level, as such the interactions between many pixels instead of between pairs of pixels are now considered. Galloway [46] defined the GLRLM:

> "The [GLRLM] matrix element (i, j) specifies the number of times that the picture contains a run of length j, in the given direction, consisting of points having grey level i (or lying in grey level range i)."

Galloway also described five statistics derived from this matrix to be used as features:

$$\text{Short Run Emphasis} = \frac{1}{n_r}\sum_{i=1}^{M}\sum_{j=1}^{N}\frac{p(i,j)}{j^2} \tag{3.6a}$$

$$\text{Long Run Emphasis} = \frac{1}{n_r}\sum_{i=1}^{M}\sum_{j=1}^{N}p(i,j)j^2 \tag{3.6b}$$

$$\text{Grey Level Nonuniformity} = \frac{1}{n_r}\sum_{i=1}^{M}\left(\sum_{j=1}^{N}p(i,j)\right)^2 \tag{3.6c}$$

$$\text{Run Length Nonuniformity} = \frac{1}{n_r}\sum_{j=1}^{N}\left(\sum_{i=1}^{M}p(i,j)\right)^2 \tag{3.6d}$$

$$\text{Grey Percentage} = \frac{n_r}{n_p}, \tag{3.6e}$$

where $p(i, j)$ is the RLM and n_r, n_p are the total number of runs and pixels. However, in these first five features there is a bias towards consideration of the runs in the image, in order to correct for this and make better use of the grey level information Chu et al [25] proposed two more features which are analogous to the first two above as can be seen by comparing their equations:

$$\text{Low Gray-Level Run Emphasis} = \frac{1}{n_r}\sum_{i=1}^{M}\sum_{j=1}^{N}\frac{p(i,j)}{i^2} \tag{3.7a}$$

$$\text{High Gray-Level Run Emphasis} = \frac{1}{n_r}\sum_{i=1}^{M}\sum_{j=1}^{N}p(i,j)i^2 \tag{3.7b}$$

Subsequently four more features were defined by Dasarathy and Holder [34] based on the joint

distribution probabilities rather than the grey level and run length probabilities separately:

$$\text{Short Run Low Gray-Level Emphasis} = \frac{1}{n_r}\sum_{i=1}^{M}\sum_{j=1}^{N}\frac{p(i,j)}{i^2 j^2} \tag{3.8a}$$

$$\text{Short Run High Gray-Level Emphasis} = \frac{1}{n_r}\sum_{i=1}^{M}\sum_{j=1}^{N}\frac{p(i,j)i^2}{j^2} \tag{3.8b}$$

$$\text{Long Run Low Gray-Level Emphasis} = \frac{1}{n_r}\sum_{i=1}^{M}\sum_{j=1}^{N}\frac{p(i,j)j^2}{i^2} \tag{3.8c}$$

$$\text{Long Run High Gray-Level Emphasis} = \frac{1}{n_r}\sum_{i=1}^{M}\sum_{j=1}^{N}p(i,j)i^2 j^2 \tag{3.8d}$$

These 11 features make up the standard set of features that are usually calculated when using GLRLMs. Typically they will be calculated in the same 4 directions in 2D and 13 directions in 3D as described in Section 3.6. Like the GLCM features, many of the GLRLM features are correlated with one another and a feature reduction/selection technique is usually required. The extension of GLRLMs from 2D to 3D is a simple one requiring the additional directions to be taken into account and all the relevant pixels to be properly counted.

3.8 Grey Level Size Zone Matrices

Grey level size zone matrices (GLSZMs) are an analogous extension to GLRLMs, instead of counting the number of a runs in a particular direction, the number zones (areas or volumes, defined by an arbitrary connectivity matrix) are counted [110]. A zone is defined as a region of connected pixels of the same grey-level.

An entry, (i, j), in a GLSZM is the number of zones with grey-level, i, and of size, j, found in an image. The number of rows in the matrix is fixed by the grey level range in the image but the number of columns will vary according to the texture in the image - the more homogeneous an image, the wider and flatter the matrix will be, while an image with many inhomogeneities will be more compact, with larger values. Once the GLSZM has been calculated, a set of features analogous to the 11 GLRLM features are calculated, for example, long run emphasis becomes large zone emphasis and Long Run High Gray-Level Emphasis becomes Large Zone High Gray-Level Emphasis. The equations for each feature are not repeated here as they are identical to the GLRLM equations but with any reference to runs replaced by zones.

The connectivity matrix that is used to define zones can be any arbitrary logical matrix and can therefore be either 2D or 3D, meaning that GLSZMs can be thought of as "natively" volumetric rather than being "extended" to 3D as with GLCMs and GLRLMs. That GLSZMs are not tied to a particular dimensionality means that this approach is also capable of avoiding the problem with not examining all the pixels in a volume as described in Section 3.4. If a full

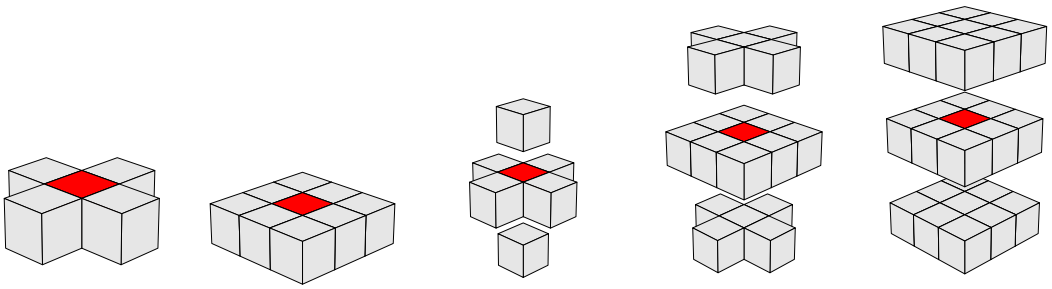


Figure 3.5: Diagrams displaying the connectivity matrices used for GLSZM calculation. From left to right 4, 8 (2D) connectivity, 6, 18, 26 (3D) connectivity.

connectivity matrix is used (the 26-connectivity in Figure 3.5) then a GLSZM will take into account all the pixels in a volume - alleviating one of the issues with statistical approaches versus filtering/transform approaches. In the work presented later in Chapter 5, 5 connectivity matrices were used to generate GLSZMs: each with connectivity of 4, 8 (2D) and 6, 18, 26 (3D), these can be seen in Figure 3.5.

3.9 Gabor Filters

The Gabor transform, first introduced by Gabor [45], is a special case of the short-time Fourier transform (STFT) and provides information about the frequency and phase content of a signal in a localised window as it changes over time. The STFT is calculated by applying a windowing function (which is only non-zero in a local region) to a time varying signal and sliding the window along the time axis. The Fourier transform is then calculated on the resulting signal which generates a two dimensional output signal that measures the frequency and phase of the original signal as it changes in time. The Gabor transform of the signal $x(t)$ at time τ and frequency f is

$$G(t,f) = \int_{-\infty}^{\infty} \exp\left(-\pi(\tau - t)^2\right) \exp\left(-j2\pi f\tau\right) x(\tau) d\tau \tag{3.9}$$

Two dimensional Gabor filters are the elliptic generalisation of the one dimensional transform and when applied to a 2D image (here we move away from considering a time-varying signal to a spatially varying signal: an image) they are capable of characterising the spectral and spatial content of an image. The human visual cortex representation was shown to closely correspond to two dimensional Gabor filters [35, 73] and this is thought to be the reason that Gabor filters are well suited to pattern recognition and texture classification problems.

Gabor filters have also been extended to 3D [9]:

$$\varphi_{f,\theta,\phi} = S \times \exp\left(-\left(\left(\frac{x'}{\sigma_x}\right)^2 + \left(\frac{y'}{\sigma_y}\right)^2 + \left(\frac{z'}{\sigma_z}\right)^2\right)\right) \times \exp\left(j2\pi(xu + yv + zw)\right) \tag{3.10}$$

where $u = Fsin\phi cos\theta$, $v = Fsin\phi sin\theta$, $w = Fcos\phi$, $[x'y'z']^T = R \times [xyz]^T$, S is a normalisation scale and $F = \sqrt{u^2 + v^2 + w^2}$ is the amplitude of the complex sinusoid wave with frequency (u, v, w). $\theta \in (0 \leq \theta < \pi)$ and $\phi \in (0 \leq \phi < \pi)$ are the orientations of the wave vector in the 3D frequency domain and $\sigma_x, \sigma_y, \sigma_z$ define the width of the Gaussian envelope in the x, y and z axes, respectively. R defines the rotation matrix for transforming the Gaussian envelope to coincide with the orientation of the sinusoid.

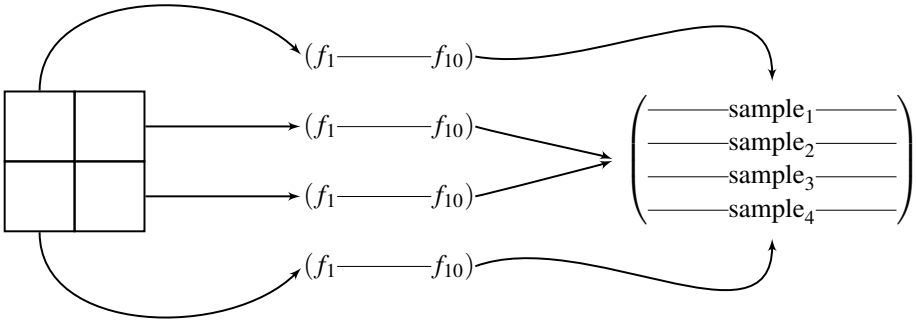


Figure 3.6: Illustration of how pixel wise features are generated using Gabor (or filters in general) filters. On the far left there is a toy image with four pixels. Ten Gabor filters are applied and the filtered value for each pixel is put into a 10×1 feature vector. These feature vectors are then concatenated to form a 4×10 feature matrix for use in a classification/segmentation/etc task.

Features are typically generated by constructing a filter bank over a range of ϕ, θ and F. The application of a Gabor filter bank to an input image generates a set of filtered images representing the response of the image to the filter. In order to construct a texture descriptor from these filtered images a choice must be made between the extraction of local or global features according to the task at hand. For the purposes of classification, one may wish to:

1. Classify an image as a whole. In this case a global feature or set of features is required: A first order statistic or several statistics of each filtered image are calculated and used as the texture descriptors - the energy is usually the statistic of choice, though less frequently the mean and variance are used. This results in a total number of features equal to the product of the number of filters and the number of statistics. In Chapter 5 this approach was chosen, using the energy as the feature as it measures the amount of information in the image at the given scale and direction determined by the filter.
2. Or to classify sub-regions of an image for segmentation/retrieval. In this case a local feature or set of features is required: Local descriptors can be generated for each pixel in the original image by concatenating the value of the pixels in each of the filtered images into a feature vector. This generates a feature vector at each location in the original image, whose length is the number of filters in the filter bank.

Gabor filters have seen widespread use in image processing (medical and otherwise) tasks

including: image registration [21, 69, 98] segmentation and classification [27, 28, 37, 57, 68, 78, 87, 99, 117], content based image retrieval [3, 6, 77, 83, 120] and motion tracking [52].

3.10 Local Binary Pattern

Local binary patterns [79] (LBP) offer another statistical measure of texture. A local binary operator is applied to each pixel in the image to convert the pixel to a binary number. The operator takes a central pixel and its neighbours, the neighbours are thresholded and set to 0 if their value is less than that of the central pixel and set to 1 if higher than the central pixel. These neighbours are then transversed in a clockwise fashion and concatenated to form a binary number. This binary number, or its decimal conversion, is assigned to the central pixel. This results in the generation of a new image made up of the binary numbers. This operator can be extended and made multi-resolution by changing the radius of the neighbours used. However when the radius (R) is increased, the number of pixels also increases and so it is necessary to limit the number of neighbours on the circle to a smaller number of equally spaced points (P). This all results in the local binary operator being defined as:

$$\mathrm{LBP}_{R,P} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p, \tag{3.11}$$

where

$$s(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.12}$$

This operator is grey scale (or illumination) independent because it measures how texture is varying in relation to the grey level value of the central pixel.

At this point the LBP operator is both sensitive to the orientation of texture and the choice of which neighbour to start with when assigning the binary bits. In order to make the operator rotationally invariant a bit wise right shift is applied to the LBP numbers (this is equivalent to circularly shifting the numbers, while maintaining their order) until we find the sequence with the lowest value.

In order to generate a set of features for use in a classification or segmentation task, a histogram of the binary patterns is calculated over a local region around each pixel in the image. The values of each bin in the histogram may be used as features or a dissimilarity measure between histograms may be used to segment an image. The histograms are calculated to extract more information surrounding the pixel than the size of the LBP neighbourhood.

Uniform patterns [80] are another extension to LBPs which involve the observation that the majority of texture patterns have a small number of transitions from 0 to 1. A pattern is defined

as being uniform if it has 2 or fewer transitions in its binary pattern. This generates a new operator:

$$\mathrm{LBP}^{riu2}_{R,P} = \begin{cases} \sum_{P-1}^{p=0} s(m_p - m_c) & \text{if } U(LBP_{P,R}) \leq 2 \\ P+1 & \text{otherwise} \end{cases} \tag{3.13}$$

where

$$U(LBP_{R,P}) = |s(g_{P-1} - g_c) - s(g_0 - g_c)| + \sum_{p=1}^{P-1} |s(g_p - g_c) - s(g_{p-1} - g_c| \tag{3.14}$$

The histogram is then calculated, as described above, for $\mathrm{LBP}^{riu2}_{R,P}$ and features are generated accordingly. As there are small numbers non-uniform patterns in most images, which could lead to under sampling of these bins in the LBP histogram, the grouping of all these patterns into a single miscellaneous bin helps to mitigate against this. The uniform pattern operator also brings with it an increased rotational invariance and an increased noise insensitivity (uniform patterns are some times called “stable” patterns).

Another feature that may be computed is a measure of the local contrast [118] which is otherwise discarded during LBP calculations. This contrast operator is defined as

$$VAR_{P,R} = \frac{1}{P} \sum_{p=0}^{P-1} (g_p - \mu)^2, \tag{3.15}$$

where,

$$\mu = \frac{1}{P} \sum_{p=0}^{P-1} g_p \tag{3.16}$$

This measure is complementary to the $\mathrm{LBP}^{riu2}_{R,P}$ operator and so these features may be combined to improve texture description and discriminatory power.

The extension of LBP to 3D is a simple one that involves redefining the neighbourhood from a circle of radius r to a sphere of radius r and sampling equally spaced points on the surface of the sphere. This modification is similar to the GLSZM and Gabor methods in that it can be thought of as a natural extension to 3D because only the neighbourhood metric must be changed. However this technique will still suffer from ignoring pixels at larger distances, primarily due to memory constraints. An efficient FFT based implementation of 3D LBP computation has also been developed [40]. LBP methods have seen extensive usage in the literature and are a valuable tool for analysing texture - they are conceptually simple but have significant discriminatory power. However the main disadvantage is that the radius and the number of points on the circle/sphere require optimisation for a given task.

3.11 Summary

Texture is a concept with a broad scope and many definitions depending on the field in question. In this chapter we have introduced the subject and argued for a narrower definition in the current context. The notion of texture in volumetric data has been explored and an introduction to the applications of various texture analysis methods has been presented alongside the theory behind a number of techniques that are later used in Chapters 5 and 6. In summary texture analysis is a powerful tool with many wide ranging applications. Texture is an important tool in the toolbox for those interested in the field of medical image analysis and should continue to provide further avenues for research and development of novel methods and techniques to make better use of the increasing amount of imaging data that is acquired in the modern hopsital.

Chapter 4

Machine Learning and Medical Image Analysis

> "Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed." - Arthur Samuel (1959)
>
> "A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E." [75]

4.1 Introduction

Machine learning is a broad and developing field that is a confluence of developments in statistical theories and computer science. The statistical problem was conceived as "what information can be extracted from a set of data when using particular modelling assumptions", whilst the computer science question was concerned with building machines that can solve problems and with the determination of which problems were tractable [76]. Machine learning seeks to answer both of these questions and also the implementation of systems and algorithms that allow a computer to represent data sets and build a generalised model capable of predicting an outcome from previously unseen samples. There are many different applications of machine learning in use today including voice recognition, computer vision, news aggregation, self driving cars, recommendation systems and medical diagnosis. This is a rapidly growing field that will undoubtedly change greatly in the coming decades, especially given the ever increasingly widespread data acquisition and storage facilities made available by recent advancements in and the relative ubiquity of computing technology.

The rest of this chapter will further explicate the field of machine learning with a description of applications, focussing on the field of medical image analysis. Section 4.2 presents a taxonomy of learning techniques. Various supervised learning methods are presented in Section 4.3 and finally the chapter is completed with some discussions of feature preprocessing, performance evaluation and training strategies.

4.2 Taxonomy of Approaches

Machine learning algorithms can be divided into clear categories based on the learning style they use. Categorisation by the underlying algorithm involved is another complementary division of machine learning algorithms, though it can be more ambiguous as some techniques may fit into more than one category. The division outlined below is robust, if rather broad.

4.2.1 Supervised Learning

Supervised learning methods operate on a set of samples for which both features and known class labels are available. This means that the learning algorithm must find a way to represent a relation between the features and the class labels while maintaining a level of generalisation so that the model is capable of also predicting on previously unseen samples. Supervised learning methods can be used for classification tasks where a discrete label is to be predicted or for regression tasks where a continuous variable is the desired output. Examples of supervised learning in action include handwriting and optical character recognition [70], email spam detectors [11] and tissue classification [115]. More details of some supervised learning methods that have been used in Chapters 5 and 6 are explained later in Section 4.3.

4.2.2 Unsupervised Learning

In contrast to supervised learning, unsupervised learning does not require any correct labels for the learning step. This is advantageous as the acquisition of features may be cheap while the acquisition of correct labels is expensive. For example, it can be cheap and easy to calculate texture features on a CT scan, but it is a time consuming task for a clinician to outline all the organs and any disease present on an image. Another example of unsupervised learning can be seen in the use of clustering algorithms to serve similar news articles together within news aggregators. Some unsupervised methods, like the K-means algorithm [71] require, the number of classes to be provided. This must be determined by prior domain knowledge or by trial and error. Other unsupervised methods are capable of inferring the number of classes or clusters to use, an example of one of these algorithms is DBSCAN (Density-based spatial clustering of applications with noise) [39]: another clustering algorithm which views clusters as areas of high density separated by areas of low density. Other types of unsupervised learning methods include Gaussian mixture models, manifold learning methods and matrix factorization approaches[51]. PCA, which will be described later in terms of feature preprocessing, can also be considered an unsupervised learning technique.

4.2.3 Semi-Supervised Learning

Semi-supervised learning methods have been developed to take advantage of situations where there are is a mixture of labelled and unlabelled data available. The structure of unlabelled data can allow a better and more generalised decision boundary to be formed (and over fitting avoided) than when only labelled data is used. An example of a semi-supervised algorithm is the transductive support vector machine (t-SVM) [8] which aims to draw a decision boundary that maximises the margin over all the unlabelled data by way of an additional loss function for the unlabelled data, more details of the SVM algorithm are presented in Section 4.3.2.

4.2.4 Reinforcement Learning

Reinforcement learning [105] is a technique used particularly in the field of robotics where a programme is designed to learn behaviour as it is exposed to penalties and rewards in the surrounding environment. As such there are no class labels or continuous variables to predict, rather the aim is to achieve certain goals. Google's self-driving car utilises an A* search algorithm and reinforcement learning in the planning stages of a drive.

4.3 Supervised Classification Models

The remainder of this chapter focuses on supervised learning methods as they are the most appropriate for the class of problems that are tackled in this work (see Chapters 5 and 6), where ground truth class labels are available. Focus will only be given to classification models and not regression. Issues involving model complexity and bias-variance trade-offs will be discussed in Section 4.6. The main advantage of supervised learning methods is the ability for an algorithm to correct itself by punishing wrong predictions. The disadvantages of supervised learning methods lie in the collection of labels: a learned model can only be as good as the reliability of the class labels which may be costly to acquire in many scenarios. For example, in the medical field, it can be time consuming for a clinician to outline regions on a CT scan (which provides class labels for each pixel/voxel), additionally there is no guarantee that regions are outlined correctly - this will introduce incorrect class labels, introducing noise at the training, cross validation and testing steps of a classification regime.

4.3.1 k-Nearest Neighbours

k-nearest neighbours (kNN) is an instance based approach that holds all the samples in an optimised data structure. Predictions on new samples are carried out by finding the nearest k neighbours and assigning a class label based on a majority vote of the neighbours. kNN is a non-generalising model because it simply holds all the training data in memory for classification. k must be chosen beforehand: varying values of k will affect the performance of the algorithm

and so this may be tuned by cross validation. (An odd k is obviously desirable to prevent tied votes.) As k increases the classifier is less susceptible to noise in the data but will result in a smoother decision boundary. Any distance metric can also be used, though the Euclidean distance is most common. For a binary classification task (class labels $y = [-1, 1]$), once the training examples have been sorted according to their distance from the sample to be classified, the predicted classification label, $\hat{c}$is given by:

$$\hat{c} = \sum_{i=1}^{k} y_i \tag{4.1}$$

This formula can be easily extended to a multi-class problem.

The non-parametric nature of kNN allows it to easily represent very complicated decision boundaries, which is one of the reasons why it remains a successful algorithm, despite its simplicity.

Variations of kNN include using a radius (rNN) instead of a fixed number of points when finding the neighbours of a sample. This allows for more votes to be taken in high sample density areas and prevents neighbours that are a long distance away from making contributions to the classification for samples in low density areas. Another variation is to weight the votes according to their distance from the sample to be classified, where the predicted classification label is given by:

$$\hat{c} = sign(\sum_{i=1}^{k} d_i \times y_i) \tag{4.2}$$

kNN performance suffers as the number of features increases because the number of dimensions in which the distance metric is calculated necessarily increases. This is an example of the so called “curse of dimensionality”: as the number of dimensions increases, the distance metric between points decreases: consider the Euclidean distance for two points separated by unit distance in each dimension is given by $(1+1)^{\frac{1}{2}} = 1.414$ in 2D, by $(5)^{(\frac{1}{5})} = 1.38$ in 5D and by $(1000)^{(\frac{1}{1000})} = 1.01$ in 1000D. This results in all points in higher dimensional space appearing to be “near” each other - rendering the kNN approach nearly useless. However, kNN has been used successfully in many arenas including gene expression analysis [84], face recognition [106] and breast cancer detection [2].

4.3.2 Support Vector Machine

The support vector machine [29] is a model based algorithm. Unlike kNN there is a distinct model training step and this model is then later used for classification. SVMs have proven popular in the last 20 years and as a result they are well understood and there are several mature open source implementations available. SVMs are formulated for the binary classification task, though they have been extended to multi-class problems by using one-vs-one or one-vs-all strategies and there are also extensions of the SVM to regression problems. In addition to the training step, a cross validation step is usually used with SVMs to tune one or more hyper-parameters of the model.

The SVM is a development of the support vector classifier (SVC) which itself is a generalisation of the maximal margin classifier: the rest of this section will explain how a SVM functions, beginning with the maximal margin classifier.

A hyperplane is a flat subspace, that is not required to pass through the origin, of one less dimension than the space in which it is present. That is, a hyperplane in 2D is a line and in 3D a 2D plane. The equation for a hyperplane in p dimension space is:

$$\beta_0 + \sum_{i=1}^{p} \beta_i X_i = 0 \tag{4.3}$$

If a point $X = (X_0, ..., X_p)^T$ does not satisfy this equation, then the function on the left hand side will have either a positive or a negative value, that is the hyperplane partitions a p dimensional space into two segments. A hyperplane can then be used as a classifier by partitioning a space into two segments: an observation, x, can be tested by determining

$$\hat{c} = sign(\beta_0 + \sum_{i=1}^{p} \beta_i x_i) \tag{4.4}$$

and the magnitude of the right hand expression will give a measure of the confidence of a label. The closer an observation is to the hyperplane, the less confidence about the class label.

However for a perfectly separable data set there are an infinite number of hyperplanes which separate the classes. The obvious choice of hyperplanes is the one that is furthest away from each of the training samples i.e. the hyperplane which maximises the margin according to the optimisation problem:

$$\text{maximise(M)} \tag{4.5a}$$

$$\text{subject to} \sum_{i=1}^{p} \beta_i X_i = 1, \tag{4.5b}$$

$$y_i(\beta_0 + \beta_1 x_{i1} + ... + \beta_1 x_{i1}) \geq M \; \forall \; i = 1, \ldots, n \tag{4.5c}$$

Equations 4.5b and 4.5c constrain the hyperplane such that each point is on the correct side

and at least a distance, M, away from it. Not all the training points determine the position of the hyperplane, only the samples that are closest to the hyperplane determine the margin, any other training sample is free to move as long as it does not cross the margin or the hyperplane. These training samples that determine the hyperplane's position are said to *support* the hyperplane.

The maximum margin classifier breaks down when no separating hyperplane exists - in order to treat this case the concept of a soft margin must be added. A soft margin allows for a number of training points to be within the margin or even on the wrong side of the hyperplane in order to better classify *most* of the training data and to make the model more robust to variations in individual training instances - this is the essence of the SVC. There is a parameter, C, in the SVC that controls the number of and the severity of violations of the margin. This parameter is usually tuned using cross validation. A high value of C will allow many violations and hence be a softer fit to the data than a lower value of C would permit. In a fashion similar to the maximal margin classifier, only the samples that lie on the margin or within the margin affect the hyperplane position and therefore the classification. For the case of the soft margin, the optimisation problem becomes:

$$\text{maximise(M)} \tag{4.6a}$$

$$\text{subject to} \sum_{i=1}^{p} \beta_i X_i = 1, \tag{4.6b}$$

$$y_i(\beta_0 + \beta_1 x_{i1} + ... + \beta_1 x_{i1}) \geq M(1-\varepsilon_i) \; \forall \; i = 1, \ldots, n \tag{4.6c}$$

$$\varepsilon_i \geq 0, \sum_{i=1}^{n} \varepsilon_i \leq C \tag{4.6d}$$

The SVC only supports linear classification boundaries and one way to overcome this limitation is to calculate new features from the original ones and use these for classification. For example using the square of each feature in addition to the original features. This allows for the creation of non-linear decision boundaries in the original feature space with the trade-off of a very quick increase in computational complexity as the overall number of features rises as higher order polynomials are used or other interactions between features are taken into account. An SVM, however, uses a kernel function to allow for a non-linear class boundary without a large increase in computational complexity. This is achieved by observing that the SVC can be calculated in terms of the inner products of feature vectors, rather than the vectors themselves, which allows for the replacement of the inner product with a generalisation of the inner product: a kernel. By using a polynomial kernel an SVC is fitted in higher dimensional space than in the original feature space; the combination of an SVC and a kernel is an SVM.

Common kernels are polynomial kernels of the form $K(x_i, x_j) = (1 + \sum_{j=1}^{p} x_{ij} x_{i'j})^d$ and radial kernels of the form $K(x_i, x_j) = exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2)$. Each of these kernels have parameters

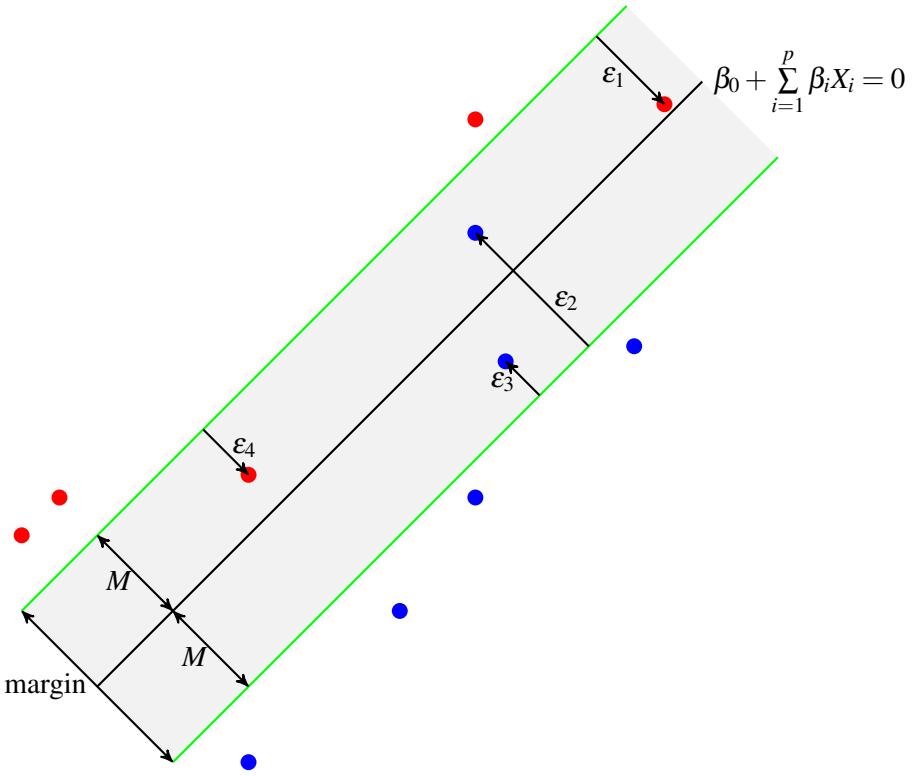


Figure 4.1: Support vector classifier hyperplane separation. This figure shows, diagrammatically, the optimisation problem in Equation 4.6.

(d and γ) that must be tuned, typically by cross validation.

The main advantage of the SVM is that it allows for the fitting of a classification boundary in a higher dimensional feature space than that represented by the training data, indeed the kernel trick enables the calculation in some feature spaces which would otherwise be intractable due to the computational requirements if attempted without the kernel parametrisation.

SVMs were initially thought to be a completely new class of machine learning methods, however, subsequent research has demonstrated a close connection between SVMs and logistic regression methods due to the similarities of the loss functions that are minimised by the models. As mentioned above, SVMs have been extended to multi-class problems, regression and semi-supervised learning problems. The SVM continues to be a popular machine learning technique and should remain so given that it is well understood, performs well on a range of data sets and because there a number of popular libraries implementing the algorithm. Some examples of its use are for liver disease discrimination [68], hyperspectral image classification [108], breast cancer detection [19], facial recognition [49] and the use of SVMs as components of deep learning frameworks [107].

4.3.3 Naive Bayes

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} \tag{4.7}$$

A Naive Bayes Classifier (NBC) is based on Bayes' theorem (Eq. 4.7) and the assumption that all features are independent of each other. This is almost always a poor assumption (hence "naive"), but the NBC can still demonstrate good performance even when this assumption is unjustified. An NBC is also easy to implement and computationally inexpensive, making it attractive for a variety of applications [11].

The NBC is formulated as follows. Given C_j are class labels and x_i are the features:

$$P(C_j|x_1,\ldots,x_n) = \frac{P(C_j)P(x_1,\ldots,x_n|y)}{P(x_1,\ldots,x_n)} \tag{4.8}$$

The assumption of independence gives:

$$\forall i : P(x_i|C_j,x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n) = P(x_i|C_j) \tag{4.9}$$

That is, the probability of a feature taking a value given a class label is not dependent on the value of the other features.

Equation 4.8 then becomes:

$$P(C_j|x_1,\ldots,x_n) = \frac{P(C_j)\prod_{i=1}^{n} P(x_i|C_j)}{P(x_1,\ldots,x_n)} \tag{4.10}$$

Since the denominator is constant for a given input:

$$P(C_j|x_1,\ldots,x_n) \propto P(C_j)\prod_{i=1}^{n} P(x_i|C_j) \tag{4.11}$$

Predicted class labels, $\hat{c}$, are then assigned according to:

$$\hat{c} = \underset{j\in\{1,\ldots,J\}}{\operatorname{argmax}} P(C_j)\prod_{i=1}^{n} P(x_i|C_j) \tag{4.12}$$

Different NBCs can be constructed by varying the distribution used to model $P(x_i|C_j)$.

4.3.4 Decision Trees

Decision trees are a model based supervised learning method that partitions the feature space by binary recursion: the feature space is split in two over and over again at the training stage. This generates a tree which is traversed in order to test a sample. Decision trees are capable of regression, multi-class classification, can make use of numerical and categorical features and are able to handle missing values. The rest of this section focuses on classification decision trees and binary classification problems, though the modifications to regression/multi-class are trivial.

A decision tree is constructed by splitting the training set into two groups according to a single feature. Each feature is used to split the classes into regions and the split that minimises the impurity is chosen. This creates two new "leaf nodes" which are also split in the same fashion as the first node and this process is repeated until one of the stopping criteria is reached. The impurity is a measure of the number of each class correctly partitioned by the decision made at each node, one measure could be the classification error but this performs poorly and so either the "Gini impurity" or the "cross entropy" are used:

$$\text{Gini } H(X_m) = \sum_k p_{mk}(1 - p_{mk}) \tag{4.13a}$$

$$\text{Cross-Entropy } H(X_m) = \sum_k p_{mk} log(p_{mk}) \tag{4.13b}$$

where $p_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k)$ for a node m, representing a region in the feature space R with N_m samples.

The stopping criteria are a set of hyper-parameters that must be chosen prior to training:

- maximum depth of nodes permitted in the tree: high values will lead to over fitting while low values may prevent the tree from obtaining sufficient complexity,
- minimum samples per split: the minimum number of samples required to split a node, high values will prevent small numbers of samples from skewing the model,
- minimum samples per leaf: the minimum number of samples required to form a new leaf, again high values should prevent over fitting,
- maximum leaf nodes: this is usually used instead of maximum depth and controls the overall size of the tree rather than the depth.

Decision trees have many advantages; they are easy to understand, conceptually easy to explain and it is believed that they may more closely reflect human decision making processes. In addition it is easy to visualise a decision tree as a flow chart which exposes the importance of each feature at each stage and so a decision tree is a "white box" model - it is transparent (unlike, for example, artificial neural networks). Decision trees are also capable of handling missing values. If a test case is missing a value, both branches of a node are traversed and the predicted probabilities are used to choose the more likely label. Categorical and numerical data

are both first class citizens in decision trees - qualitative features do not have to be converted to a numerical scale to be used as a decision tree can split using categorical data in the same fashion as numerical data.

However decision trees also have significant disadvantages - they tend to over fit problems easily and do not generalise well and it is for this reason that they can be very unstable - small changes in a training set can generate a very different looking decision tree. These problems can be mitigated against and overcome by the use of bagging (Section 4.3.4), Random Forests (Section 4.3.5) or Boosting (Section 4.3.6). Even with these advances, decision trees have some disadvantages. For example, finding the optimal decision tree is a NP-complete and so a greedy algorithm (binary recursion) must be used to build the tree. Some concepts such as XOR can be difficult for a decision tree to learn as it is not well represented by a tree. A decision tree is susceptible to bias if the training set is heavily biased towards one class - this can be seen by inspecting the impurity measures above. Despite these disadvantages, decision trees are useful learners, especially in their use as "weak learners" as the basis for more complicated learning algorithms like AdaBoost.

Bagging

Tree bagging [15] averages the predictions of many decision trees in order to reduce the variance of the trained model, which should allow for more stability and more generalisation. This averaging would ideally be done on a number of different training sets, but data is expensive to collect, so a single training set is instead sub-sampled and each sub-sample is used to train a decision tree. Thousands of trees may be constructed and bagged. Each tree will have a very high variance and a low bias, but the averaging over the trees lowers the overall variance again. When doing classification the bagged trees are averaged by taking the majority vote of the tree or by averaging over the probabilistic prediction of each tree, in order to form a single predictor. One benefit of bagging is that the number of trees is not a crucial parameter as the model will not over fit with large numbers of trees. Very large numbers of trees may be used: the only constraint on the number of trees is that enough trees have been used to ensure that the error rate has stabilised. The main disadvantage of bagging is that it obscures the importance of each feature that is so easily available when using a single tree - it is no longer possible to display a single, interpretable, tree. One way around this problem is to measure the feature importance by finding the total decrease in impurity over all the splits caused by each feature, averaged over all the bagged trees.

4.3.5 Random Forests

A disadvantage of bagging trees is that they will be highly correlated, as the same features are used for training each tree. This is because the sampling is only performed on training examples, not on the features. If there is one particularly dominant feature, then this feature will almost always be the first used to split the first node and from there the generated trees will all look similar - the particular values used for splitting may change, but the features used to split remain similar across trees. Averaging over a large number of similar (correlated) trees does not decrease the variance as much as averaging over dissimilar (decorrelated) trees.

In order to overcome this problem, a random forest [16] of trees is created by a similar approach to bagging, but instead of choosing to use all the features for training each tree, only a subset of features is used at a time. A number of features are chosen randomly and used to train each tree. This removes any strong predictors from a number of the trees allowing other features to be considered for the first split which may radically alter the structure of the tree. The same processes of casting a majority vote from the trees or averaging the probabilistic prediction of each tree can be used to determine the predicted label from the random forest. If there are a total of n features available, then each decision tree in the forest is usually trained with $\sqrt{n}$ features, though any number of features ($< n$) can be chosen.

The main advantage of random forests over decision trees is the increased performance by the reduction in variance, with a small increase in the bias (compared with an individual tree). However just like the bagging method, the random forest obscures the importance of features, although the feature importance method described above can again be used to gain an indication of the most significant features. Additionally, over fitting is still a danger with the random forest if the individual trees are too complex.

4.3.6 Boosting

Boosting is another method for improving the performance of decision trees, though it can be used with almost any weak learner to improve classification performance. For example, a series of kNN classifiers could be used as a set of weak learners. For the sake of discussion the rest of this section will assume that decision trees are the weak learners used.

A weak learner is a classifier that achieves a misclassification rate less than 50% but does not otherwise provide satisfactory performance. Instead of creating a forest of independent trees like the random forest classifier, it is possible to use the predictions of previous trees to help train subsequent trees. The first practical algorithm that was able to achieve this boosting was the “AdaBoost” algorithm (Adaptive Boosting) [43, 44]. AdaBoost works by training a weak learner on the training set and up-weighting incorrectly classified samples while down-weighting correctly classified samples. This leads to previous misclassified (harder) samples being given greater focus in subsequent training rounds. AdaBoost trains trees sequentially

while the random forest trains trees in parallel, this typically allows for the use of smaller trees as AdaBoost enables trees to cooperate. Once the AdaBoost model is built, predictions are made in the same manner as the random forest. There are two parameters, in addition to any parameters for the weak learners, that must be tuned when using AdaBoost: the number of learners to use and the learning rate. The learning rate is a value ≤ 1 that reduces the contribution of the classifier at each learning step when building the model.

One of the main advantages of AdaBoost comes from the fact that it is simpler to design a weak learner, a weak learner only needs to be right most of the time rather than achieve some higher performance demands. Therefore, it is possible to design a weak learner and use AdaBoost with it rather than putting effort and time into developing a strong learner. One of the problems with AdaBoost is that it also obscures relative feature importance, though if many decision stumps (trees that only make one decision) are used it is possible to determine which features are important. Remarkably, a large number of these very weak learners are capable of performing well on a number of data sets.

4.4 Feature Preprocessing

Many complications arise when dealing with real world data sets. Typically there may be an abundance of features, these features may be highly correlated, there may be many noisy features present in a data set and some features may be numerical and some categorical. Some features may have missing values and some features may be several orders of magnitude larger than others. Each of these situations can be problematic depending on the machine learning algorithm used: for example, an over abundance of features can lead to the "curse of dimensionality". However a range of techniques have been developed to solve these problems.

Categorical data can be used by some learning algorithms (decision trees) but not others (SVMs). In order to make use of what can be valuable information contained in categorical variables it is possible to convert categorical data to a numerical scale. For instance, whether a person is a smoker or not can be converted to a binary value, which allows a model like the SVM to make use of this information. Yet this becomes more complicated if the categorical feature cannot be represented as a binary. To return to the smoking example, if it is desirable to capture whether a person has smoked in the past it is possible to create a new feature that takes the values

$$f = \begin{cases} 0 & \text{never smoked} \\ 1 & \text{former smoker} \\ 2 & \text{current smoker} \end{cases}$$

but this is a rather arbitrary approach. How much separation should there be in the feature space? Should current smoker have a value of 5 instead of 2? Despite these problems it is still

beneficial to convert categorical data to a numerical representation rather than discarding the information. Another approach is to create a binary variable for each option (never smoked, former smoker and current smoker) as this removes the question about how the categories should be scaled, but this can generate a very large number of features if there are many different categories for each variable.

Data acquisition can result in missing values for some features. The most common solution to this problem is to calculate the mean of the feature over the remaining samples and use this in place of the missing data, hoping that the model is still able to correctly classify this sample using the other remaining features that were correctly collected.

Different features can often vary widely in their absolute values: one feature may have a range of $-0.1 \leq f_1 \leq 2.3$ while another may lie in the range $0 \leq f_2 \leq 10000$. This can be problematic for algorithms like kNN or SVMs that use the distance between samples since a change in f_2 would dominate almost any change in f_1. To overcome this problem features can be scaled or normalised. Scaling features involves converting all the features so that they lie in the same range, typically $0 \leq f \leq 1$ or $-1 \leq f \leq 1$ while normalised features are calculated by subtracting the mean and dividing by the variance of the feature vector to ensure that all the features have zero mean and unit variance. This means that all features are now equally weighted in the feature space and any relatively large changes in one feature will not inappropriately dominate the distance metric.

4.4.1 Feature Selection and Reduction

Problems with missing values and different scaled features are relatively easy to overcome, as explained above. A much more difficult problem arises when the number of features is large compared to the number of samples, especially if some of these features are redundant or noisy. Redundant features can suppress other, less abundant, features in the data and compromise performance. Noisy features can lead to a higher risk of over fitting as they may cause a model to be built upon chance correlations between the features and the labels - the chance of these correlations occurring obviously rises with the dimensionality of the feature space. Adding more features to a classifier, unless they are correlated with the class labels, will degrade performance. This is why feature selection or reduction methods are a key part of any machine learning pipeline. There are two solutions to this problem: feature selection and feature reduction. The former chooses a subset of the features to be used in the model while the latter combines the features into a smaller number of new features.

A simple feature selection technique is to perform univariate classification with each of the features, then take the top performers and use them in a multivariate model. The main issue with this approach is that it completely discards any interactions between features: a pair of features may be capable of separating two classes while neither of the individual features can.

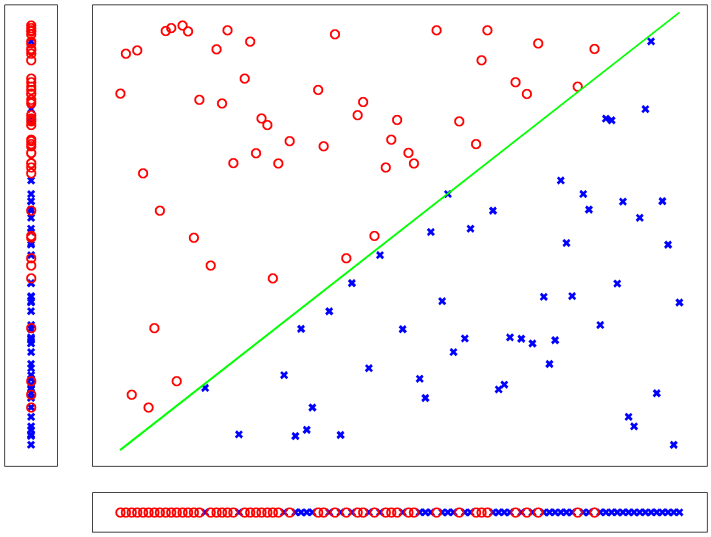


Figure 4.2: This figure demonstrates how a pair of variables may be unable to separate two classes when they are considered in a univariate fashion. The plots along the bottom and left hand side show the features considered individually, note how there is a large overlap in the classes (represented by the different markers). However when the two variables are considered together, as seen in the main plot, a clear decision boundary can be overlaid (the green line).

Exhaustively searching all the combinations of features will take an infeasible amount of time on real data and so other methods for feature selection have been developed, such as adaptive selection methods like sequential floating forward feature selection [101], mutual information based measures like minimum redundancy maximum relevance [86] and the least absolute shrinkage and selection operator [111].

Feature reduction methods, on the other hand seek to create a new set of features from the original features. The most common feature reduction method is principal component analysis (PCA) which linearly combines the features into a new set of orthogonal features of decreasing variance. These new features minimise the redundancy, measured by the covariance of features, and maximise the true signal in the data as measured by the variance. PCA can be efficiently computed by use of the SVD operator. In order to generate the reduced feature set the covariance matrix of the feature matrix $\mathbf{X}$ is first calculated

$$\Sigma = \frac{1}{n}\mathbf{X}\mathbf{X}^T \tag{4.14}$$

where $\mathbf{X}$ is the feature matrix with n features and m samples. The eigenvectors, $\mathbf{U}$, of Σ are then

calculated and the new feature matrix, $\mathbf{Z}$, is generated according to

$$\mathbf{Z} = \mathbf{XU} \tag{4.15}$$

The first x desired features are then retained by truncating number of columns in the new feature matrix, x can be chosen so a number of features are retained or so that a portion of the original variance is retained by the new data set. PCA is a non-parametric method - there are no parameters to tune, the method is applicable to any feature matrix and it is a good choice for removing correlation between features. There are two main problems with PCA *a*) it is grounded on the assumption that features with high variance are good predictors and so the new features, while capturing the variance of the data, may not provide good classification performance and *b*) PCA assumes a linear relationship between features, though this can be superseded by using kernel PCA which makes use of a kernel trick, however this requires *a priori* knowledge of the feature space in order to choose a suitable kernel. Other feature reduction methods exist, including independent component analysis, Sammon mapping and non-negative matrix factorisation [104].

4.5 Classification Performance Evaluation

There are a variety of methods to measure the performance of a learning algorithm[i]. The most basic and obvious method is the accuracy:

$$\text{accuracy} = \frac{\text{number of incorrect classifications}}{\text{number of test samples}} \tag{4.16}$$

however this measure very quickly breaks down when the number of each class is unbalanced. Take, for example, a problem with 1000 test samples, 990 of which are “dogs” and 10 are “cats”. If our classifier assigns every test case the label “dog” then the accuracy is 0.99 - indicating that the classifier seems to be performing well. Yet it will never get the “cat” samples correct. In order to properly measure the performance other metrics have been devised that are

i. This section will focus on binary classification tasks.

better able to handle such unbalanced cases.

$$\text{Sensitivity (Recall)} = \frac{TP}{TP+FN} \tag{4.17a}$$

$$\text{Specificity} = \frac{TN}{TN+FP} \tag{4.17b}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{4.17c}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{4.17d}$$

$$\text{F1 Score} = 2\frac{precision \times recall}{precision + recall} \tag{4.17e}$$

where $TP =$ the number of true positives, $TN =$ true negatives, $FP =$ false positives and $FN =$ false negatives. (A true positive, for example, is when the true label and the predicted label are both positive.) These summary statistics allow for the assessment of a classifier on a data set even if it is unbalanced, it is best to make use of several of these measures when assessing the performance of a classifier as edge cases like the example above can still be masked even with these statistics. The number of TP, TN, FP and FNs are nearly always a good indicator of the performance - these can be summarised in the form of a confusion matrix, an example of which can be seen in Table 4.3.

Predicted Label

True		**1**	**0**
	1	50	50
	0	50	1000

TP = 50	FN = 50
FP = 50	TN = 1000

Table 4.3: An example confusion matrix. Two views of the same confusion matrix for a toy classification case. This matrix shows the need for careful examination of classification performance when looking at unbalanced data. The accuracy of this classifier is 91% even though it only has a precision and recall of 50%.

4.5.1 ROC and PRC

Instead of using the the stationary sensitivity and specificity of a classifier a, "Receiver Operator Curve" (ROC) can be constructed that gives a better indication of the model performance. A ROC is a plot of the false positive rate ($1 -$ Specificity) against the true positive rate (Sensitivity). In order to generate the data for the plot the classifier must be capable of returning a probabilistic prediction (or a score of some sort). This probabilistic prediction is then thresholded at different levels producing different sets of class labels for which the sensitivity and specificity are calculated in order to generate a series of points to plot in the ROC space. The ROC then allows for the examination of the trade-off between sensitivity and specificity. An example of several ROCs can be seen in Figure 4.4. The dashed line from $(0,0)$ to $(1,1)$ represents the performance by guessing the class randomly and better

performing classifiers are represented by curves in the top left of the plot as they are capable of maintaining a high sensitivity and specificity simultaneously. An additional measure of a classifiers performance naturally drops out of the ROC: the area under the curve. The larger the area under the curve, the better the classifier with a value of 1 being a perfect learner and a value of 0.5 being equivalent to random guessing.

A similar approach may be taken with the precision and recall scores achieved by a classifier, but now curves in the top right section of the plot are high performing classifiers. The area under precision recall curves is another good measure of the performance of a classifier.

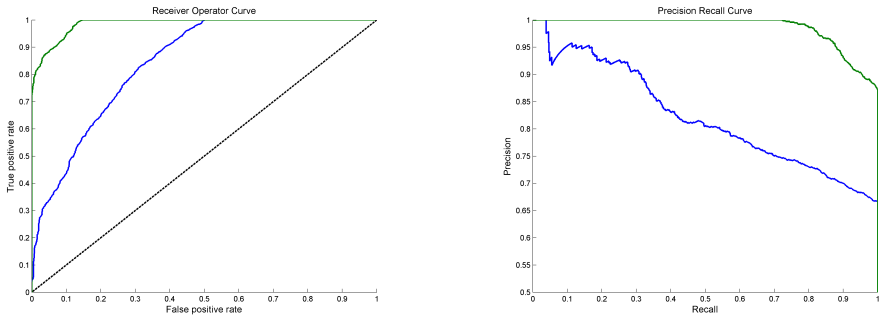


Figure 4.4: Left: Two examples of Receiver Operator Curves, the green curve represents a superior classifier than the blue curve. Right: Two examples of Precision Recall Curves corresponding to the same classifiers as the Receiver Operator Curves in the left hand plot.

4.5.2 Bias-Variance Trade-Off

Model complexity is a very important property when trying to build a classifier. Naïvely one would expect that reducing the training error will always generate a better model, yet Figure 4.5 demonstrates that a reduction in the training error may lead to an increase in the error on an unseen test set. This is described as over fitting the model or having a model with high variance. The model is well fitted to the training data but is not well generalised to properly predict new data.

Many classifiers have parameters that can tune the model complexity - for kNN it is k, for SVMs it is C and for trees it is the parameters that control the size and depth of the tree. These parameters must be chosen to optimise the bias-variance trade-off. It has already been shown that a “high variance - low bias” scenario will lead to poor results but a “low variance - high bias” model will result in a model that is under fit: it will not be sufficiently complex and will perform poorly on test data. In order to find the optimal variance/bias trade-off another set of data is usually used: the cross validation set. In a situation where there is an abundance of data the validation set approach can be used, otherwise k-fold validation is another viable method. These techniques are described in Section 4.6.

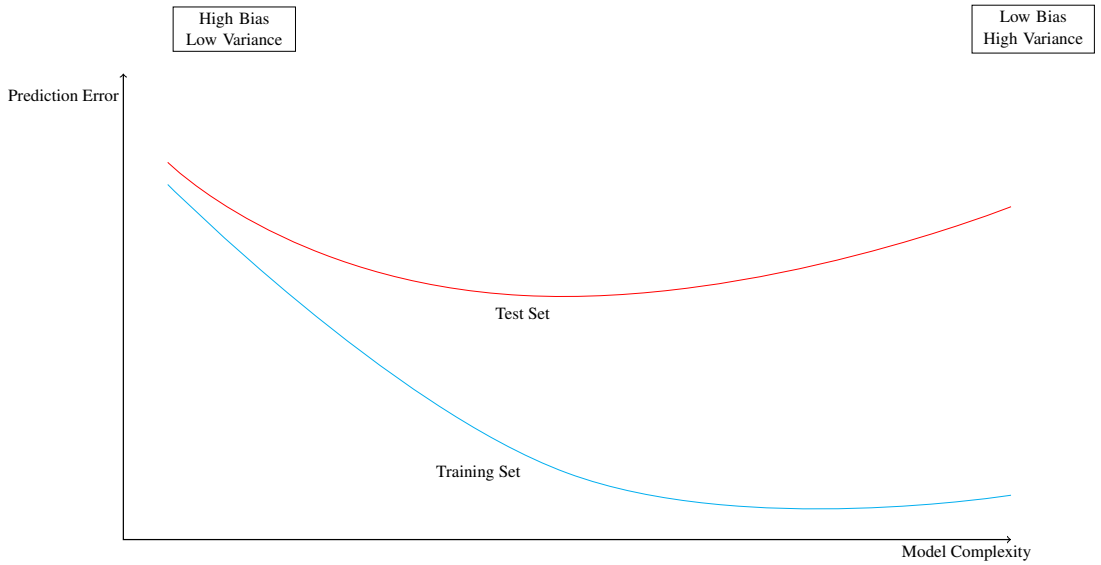


Figure 4.5: This graph displays the trade off between bias variance and the pitfalls of ever increasing model complexity. At the far left hand side the model is not complex enough to capture the variance in the data - prediction error is poor for both sets. On the far right hand side the model performs very well on the training set, but poorly on the test set as it has over fit the training set. There is a minimum in the test set error where the model is sophisticated enough to perform well on the training and test sets - this is the optimal balance between bias and variance of a model.

4.6 Training and Cross Validation Strategies

In order to train a model and to verify its performance, there are a range of different methods that can be employed, the suitability of each depending upon the size and nature of the data set. Some of the most common methods and some related concerns are addressed in this section. Only binary classification tasks will be considered, though many of the methods can be used for multi-class problems.

4.6.1 Balanced Training

Some learning algorithms such as decision trees and SVMs are susceptible to bias from uneven numbers of samples of each class in the data set. For example, if there are a large number of positive samples, then the classifier can be overwhelmed and always predict a positive label. This susceptibility to a unbalanced training set can be overcome by training with a balanced subset of the data which is created by using all the samples from the less common class label and an equal number of randomly selected samples from the other, more common class label. This proves to be an effective method for training although it can discard a significant number of training samples. Care must also be taken when measuring the performance when dealing with unbalanced data sets - this is explained in more detail in Section 4.5.

4.6.2 Validation Set

Perhaps the simplest manner in which a model can be trained and tested is by the use of a validation set by which the data set is split into two: a training set and a test set. The learning model is then trained on the training set, and the test set is used to measure the performance of the model on unseen data. A typical split would be 70% training, 30% testing. If the learning model requires a cross validation set, then the data set can be split according to a 70% training, 20% cross validation and 10% testing split. The main problem with this method is that performance can vary widely according to which samples are used for training and which are used for testing. It also fails to efficiently make use of the available data - each sample is used only once for training or for testing.

4.6.3 Leave One Out

Leave one out cross validation is a step up from the validation set approach: instead of splitting the data into a training set and a test set, only a single case is removed to form the test set, the model is then trained with the training set and used to predict a label for this test case. This test case is then returned to the data set and the next sample is set aside to be used as the test case, this process is repeated until each of the samples have been used as the test case. This allows every sample to be used for testing and ensures that as much of the data as possible is used for training. An obvious disadvantage of this method is that many models are trained which can take a long time on a large data set. As such, this method is ideal when there are a small number of cases: computation time is not increased dramatically and the small amount of data is fully exploited.

Leave *x* Out

This method is identical to the leave one out approach above but x cases are set aside to form the test set. For n training samples, this creates $\frac{n!}{x!(n-x)!}$ (overlapping) pairs of test sets.

Leave One Label Out

In order to perform leave one label out validation, an additional label - not the class labels - must be assigned to each sample. This could be date or time information or perhaps gender. For example, taking the “year” as the extra label, all samples for year 1 would be assigned to the test set and training is carried out on the rest of the samples before testing on the test set and repeating for each of the other labels. This method could be useful to track how a model performs on data from different years, or how a model performs differently on patients with different stages of disease.

4.6.4 k-Fold Cross Validation

k-fold cross validation is typically used for larger data sets when model training may take a significant period of time making leave one out impractical. The data set is split into k, roughly, equal size groups. For each of these groups a model is trained with the remaining $k-1$ groups and testing is performed on the group left out. This generates k measures of the performance, which are then averaged to provide the k-fold performance. Usually $k = 5$ or $k = 10$ to limit the computational burden. Obviously if $k = (\text{number of samples})$ then this is equivalent to leave one out validation. Stratified k-fold cross validation is an extension of this method that ensures that each of the k groups used for training are balanced and must be used with some learning algorithms.

4.7 Summary

Machine learning is a fast growing area of research due to the large amount of commercial and academic interest in its theory and applications. It is used extensively in the wider medical field and particularly in medical imaging. Applications of machine learning and artificial intelligence will only continue to grow and develop. It is this author's belief that they will become an integral part of patient care over the coming years and decades. In particular, the use of machine learning to aid the interpretation of medical images should reduce the burden and errors caused by intra- and inter-clinician variability when assessing images.

Chapter 5

Predicting the Occurrence of Radiation Induced Pneumonitis

5.1 Introduction

As outlined in Section 2.5.4, treatment of lung cancer using radiotherapy can cause radiation induced pneumonitis in 13-37% of patients. This is a debilitating condition that can be very serious in already ill patients. There is currently no reliable way to predict whether a particular patient will develop radiation induced pneumonitis. This chapter uses a cohort of 57 patients to investigate the feasibility of using features derived from a patients planning CT scan to predict the occurrence of radiation induced pneumonitis. The rest of this chapter details the data used (5.2), outlines the method (5.3), describes the features used (5.4) and then presents the results (5.6), followed by a discussion (5.7).

5.2 Study Data

The data used in this retrospective study consists of radiotherapy planning CT scans and associated clinical parameters from 83 patients treated at the Edinburgh Cancer Centre, Western General Hospital between January 2009 and September 2010. Of the 83 patients, 21 were excluded due to uncertain patient outcome (although all 21 cases were thought to be asymptomatic, the clinician was either unable to determine the outcome or the patient died before a diagnosis could be made) and a further 5 were excluded due to data processing complications leaving a total of 57 patients, 14 of which developed radiation induced pneumonitis and 43 remained asymptomatic after radiotherapy. The outcome of each patient was assessed in the standard post-radiotherapy follow up by an expert clinician familar with each case.

Each of the 57 patients were treated with external beam radiotherapy: 48 on a Varian Clinac 600C/D, 2 on a Varian Clinac iX and 7 on a Varian Clinac 21EX. All radiotherapy plans were calculated using the Varian Pencil Beam Convolution algorithm (v8.1.2). 32 of the patients were treated with a 55Gy in 20 fractionation schedule and 15 with a 60Gy in 30 fractionation schedule, the remaining fractionation schedules, alongside a summary of various radiotherapy

Number of Patients	Fractionation Schedule
32	55/20
15	60/30
6	54/36
1	52/26*
1	66/33
1	54/36
1	55/30

	Min.	Max.	Mean
V_{20}/%	4.0	36.5	23.7
V_{10}/%	5.0	70.0	42.5
V_5/%	6.0	80.0	51.2
MLD/Gy	2.3	20.7	14.3
PTV/cm^3	87.4	1323.3	463.1

Table 5.1: Left: Summary of the fractionation schemes used in the treatment of patients in this study. *60/30 was planned but stopped early due to traumatic spinal injury. Right: Summary of the radiotherapy dose delivered and the size of the PTV. V_x is the volume of lung receiving at least x Gy.

parameters can be seen in Table 5.1. 3 patients received sequential chemotherapy (2, 3 or 4 cycles of carboplatin/gemcitabine), 16 patients received concurrent chemotherapy (2, 3 or 4 cycles of cisplatin/vinoreline) and 37 of the patients received no chemotherapy. The radiotherapy planning CT scans were acquired using a 3mm CT slice thickness, (IGE HiSpeed Fx/i, GE Medical Systems, Milwaukee, WI, USA) resulting in a resolution of approximately 1mm in the axial plane with a 2048 HU range.

5.3 Methodology Overview

In order to automatically predict whether a patient is likely to develop radiation induced pneumonitis during lung cancer treatment a set of texture features are calculated from the radiotherapy planning CT scan and combined with 13 clinical features into a feature vector for each patient. PCA is then applied to reduce the dimensionality of the feature space and a classification model is trained and tested in a leave one out fashion to determine the viability of this approach.

The full procedure for predicting the outcome for a single patient is:

1. Lung parenchyma is extracted from the planning CT scan,
2. Blood vessels and airways are removed by a simple thresholding technique,
3. The remaining lung volume is then resampled to create an isometric array to compensate for the 3mm slice thickness,
4. Volume is then scaled to a lower number of grey levels,
5. A number of texture features are calculated from the volume,
6. A feature matrix is created by the concatenation of texture and clinical feature vectors,
7. Classification is performed in a leave one out fashion with an optional PCA step.

It should be noted that the PCA step was carried out outside the leave one out loop for the SVM classifier. That is, the PCA was applied to the whole feature matrix once before splitting

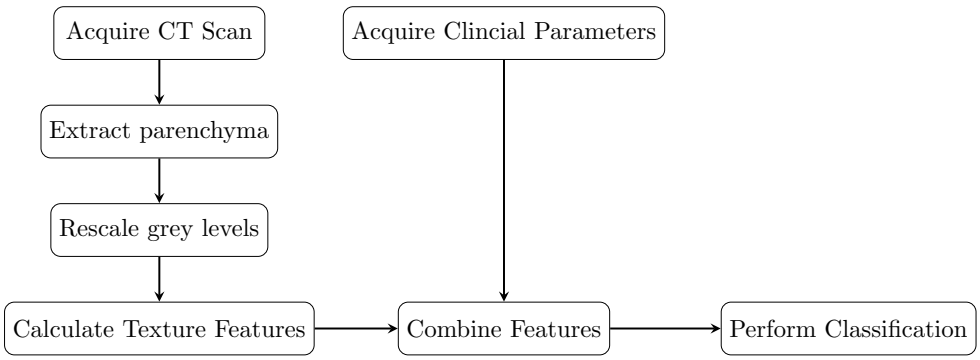


Figure 5.1: Flow chart outlining the main steps in the pipeline for predicting pneumonitis from lung CT scans.

to do the leave one out classification. This should have a minimal effect on the results as PCA is an unsupervised method - it is not exposed to the class labels and so there should not be any inappropriate information "leakage". (The PCA was implemented outside the loop to avoid slowing down already extensive execution times.) However, the when the Decision Tree classifier was used, the PCA step was implemented inside the leave one out loop - that is, PCA was applied at each step of the leave one out process. (This was possible due to the much smaller execution times for the Decision Trees compared to the SVM method). The comparable results between the SVM and the Decison Tree methods provide evidence to back up the assertion that the PCA outside the loop does not have a large effect on the results.

5.4 Feature Extraction

5.4.1 Clinical Features

There are a number of clinical and dosimetric features that have been shown to be significant when predicting radiation induced pneumonitis (see Section 2.5.4). 13 such features were chosen from a much larger set of parameters collected in the course of the normal clinical work flow for a patient with lung cancer. This had the advantage of not adding an additional burden to the existing clinical work flow. The 13 features were chosen because of their completeness across the 83 subjects, that is the vast majority of the subjects had values recorded for each feature. Additionally, each of the features were easy to convert to a numerical/boolean value in order to be used as inputs to a classification model (Section 4.4).

The 13 features used were age, smoker/non-smoker, if the patient suffers from asthma or similar conditions, T-stage, N-stage, if the patient was treated with chemotherapy, the radiotherapy dose and number of fractions, the V_{20}, V_{10}, V_5, MLD and the size of the planning treatment

volume (PTV). Table 5.2 shows these features and their division into pre-treatment features and treatment dependent parameters.

Pre-Treatment		**Treatment Dependent**	
Age	Integer	Treated with Chemotherapy	$[0,1,2]$
Smoker/non-smoker	$[0,0.5,1]$	Radiotherapy dose	Double
COPD	Boolean	Radiotherapy fractions	Integer
T-stage	Integer	V_{20}	Double
N-stage	Integer	V_{10}	Double
		V_5	Double
		Mean Lung Dose	Double
		Size of PTV	Double

Table 5.2: A summery of the clinical parameters (and their formats) used as features in the prediction of radiation induced pneumonitis.

All of the treatment dependent features are numerical, except for the chemotherapy based feature which was assigned to a ternary scale: 0 for no chemotherapy, 1 for a patient treated with sequential chemotherapy and 2 when a patient is treated with concurrent chemotherapy. Age, T-stage and N-stage are also numerical features but smoking history and COPD were converted from categorical features to a numeric scale. COPD was converted to a binary feature and smoking history was converted to a ternary scale: 0 for never smoked, 0.5 for ex-smoker and 1 for current smokers (which is equivalent to $[0,1,2]$ or $[-1,0,1]$.

5.4.2 Texture Features

In order to extract meaningful texture features, the lung parenchyma was extracted from the CT scans. The lungs on each of the CT scans were automatically segmented using the Varian Eclipse™ software which is able to extract the outside contour of the lungs. The GTV contour defined by the clinicians was then used to remove the tumour from the image volume for the same reason that the bone and blood vessels were removed: the tumour itself does not develop pneumonitis. In principle this step could also be done automatically as the tumour volume usually falls into a significantly different grey level range than the parenchyma, however the clinical volume was used here for simplicity and to focus on the outcome prediction task.

Following the extraction of the lungs, a threshold was applied to the image to remove the majority of blood vessels, airways and bone, which do not develop pneumonitis and may interfere with the texture calculation and add noise to the features in the form of unrelated information. This threshold was determined by manual inspection of the lung volume histograms from a sample of the patient population (both symptomatic and asymptomatic) and was tuned to ensure that the maximum amount of parenchyma was retained while removing the majority of other tissues.

When using 3D texture analysis, best practice requires isomorphic sampling resolution of the

3D volume. However, the slice thickness of this CT data is 3mm, with a 1mm resolution in the x- and y-directions. Thus a resampling is required to produce an isometric volume: in order to keep as much information as possible, the data was upsampled in the z-direction to produce a pseudo-isomorphic resolution of 1mm. This upsampling was carried out by repeating each of the slices in the z-direction. An interpolation method could have been used here instead, but it is believed that this would not have made an appreciable improvement to the feature matrices. For example, the interpolation would result in a "blurring" effect in the GLCMs. The upsampling is sufficient as the main aim is to ensure the information present in the z-direction is not unfairly downweighted due to the increased pixel spacing.

Finally grey level resampling was used to convert the images from 2048 grey levels to a more manageable 8, 16, 32 or 64 grey level range, this step is required for the statistical texture measurements used (Sections 3.5-3.8). If this step was omitted, a very sparse GLCM, for example, would be generated from the 2048 grey level image.

FOS (n=7: mean, variance, coarseness, skew, kurtosis, energy and entropy), GLCM (n=1820: 14 Haralick features in 13 directions over 10 different pixel distances), GLRLM (n=143: 11 features in 13 directions), GLSZM (n=55: 11 features for 5 different connectivity matrices) and Gabor (n=144: a filter bank constructed using $F = [25 : 25 : 100]$, $[\theta = -\pi/3 : \pi/6 : \pi/2]$ and $[\phi = -\pi/3 : \pi/6 : \pi/2]$ was applied to an image and the energy in the filtered result was used as a feature) features were calculated for each of the rescaled lung volumes, all of these methods are described in more detail in Chapter 3, all of the calculations were computed using custom implementations in MATLAB. This procedure resulted in a total of 2162 texture features, in addition to the 13 clinical features, being available for use in the classification pipeline.

Typical execution time[i] is in the region of 15-20 minutes per case for all of the texture excluding the Gabor features and of the order of an hour per case when including the Gabor features. These times are achieved by use of a vectorised GLCM method and an FFT based Gabor implementation. This is too slow for a real time application, but for the proposed use case where texture and classification would be run offline (perhaps overnight) on a clinical system then these execution times are acceptable - the texture stage is the longest of any with the preprocessing taking less than a minute and the label prediction stage taking no more than a few seconds. Further optimisation of the texture calculations is likely possible by translating the functions to a native implementation or by parallelising portions of the code, for example, the application of multiple Gabor filters is an "embarrassingly parallel" operation. The classification model (training) would only need to be precomputed once and so the time taken for this computation is largely irrelevant in terms of a real world application.

---

i. All calculation times recorded on an Intel (R) Xeon (R) 5150 Dual Core CPU @ 2.66Ghz.

5.4.3 Feature Reduction

Given the highly correlated nature of many of the features (GLCMs calculated in different directions at different distances, for example) it is desirable to reduce the feature space significantly. PCA was used (Section 4.4.1) to reduce the redundancy of features by mapping to a new feature space. The number of eigenvectors (new features) retained was either set to a fixed number or determined by measuring the amount of variance explained by the new features, that is, the number of eigenvectors that were retained was the smallest number of eigenvectors that account for $x\%$ of the original variance where, for example, $x = [90, 95, 99]$.

5.5 Preliminary Analysis

This section consists of a short, preliminary, analysis of the clinical and imaging data to demonstrate that there is relevant information within the CT data and within the 13 clinical features that is indicative of the radiation induced pneumonitis outcome and hence warrants further investigation of the viability of the approach outlined above.

Imaging Data

A simple comparison of the histograms of the HU values (from the CT data) from each patient group (symptomatic and asymptomatic) was carried out. The results from this can be seen in Figure 5.2. The figure shows a subtle, but distinct difference between the two groups and after fitting a probability distribution to the data, a two-tailed Kolmogorov–Smirnov (KS) test found (at the $p = 0.01$ level) that the two population samples were drawn from different distributions. This suggests that there is a distinct difference in the HU distributions in the two different patient groups and therefore there is information in the CT data that may be exploited to predict the occurrence of radiation induced pneumonitis.

Further tests were carried out on a per-patient basis in an attempt to use only the probability distributions to predict the occurrence of radiation induced pneumonitis. KS tests were used to compare all the cases to each other and to a “global” population distribution. However, no viable method was found using such a relatively simple approach. Texture analysis methods are much more capable of extracting information from the CT images than a simple histogram/probability distribution approach.

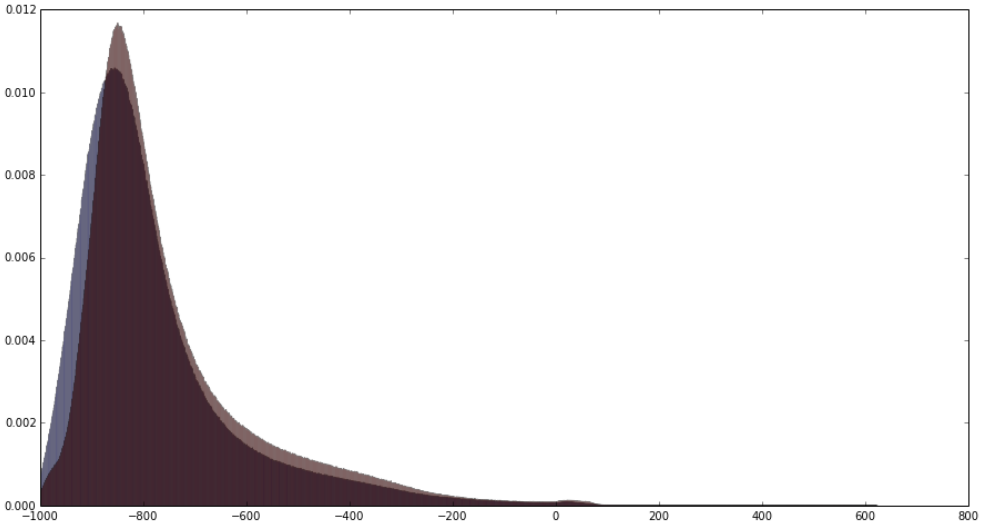


Figure 5.2: Comparison of the (normalised) histograms of the Hounsfield units for symptomatic and asymptomatic populations. The red histogram was generated from all the HU values from all 14 symptomatic cases. The blue histogram is a random sampling of all the HU values from the asymptomatic cases, sampled to ensure an equal amount of data in each group. The x-axis is HU.

Clinical Data

The 13 clinical features were analysed to determine if any were capable of separating the classes. Figure 5.3 shows a summary plot of all the features. From these plots it is clear to see that none of the features are individually capable of separating the classes. However, most notably, the age at the time of diagnosis displays the most obvious difference between the two patient groups.

After this analysis, leave one out classification was performed on the 13 clinical parameters. An SVM classifier performed with 56.6% Sensitivity, 65.7% Specificity and an Area Under ROC of 0.58. This is poorer performance than reported by Chen et al [22, 23]. This is expected given that there are far fewer features used in this study (93 compared to 13).

These results are obviously not good enough to predict pneumonitis reliably. This level of performance is likely due to the small number of features used and perhaps due to the relative lack (compared with Chen et al [22, 23]., for example) of disease state information encoded in the features. A possible line of further research could be to include such information from the patient records. It would be possible to add more clinical features directly related to the disease such as tumour position or more histological information. However, one advantage of the limited number of features used is the ease of collection - they are readily collected in the standard lung cancer treatment pathway without the need for extra tests. As will be seen later, combining these features with the texture information derived from the imaging data improves the predictive performance.

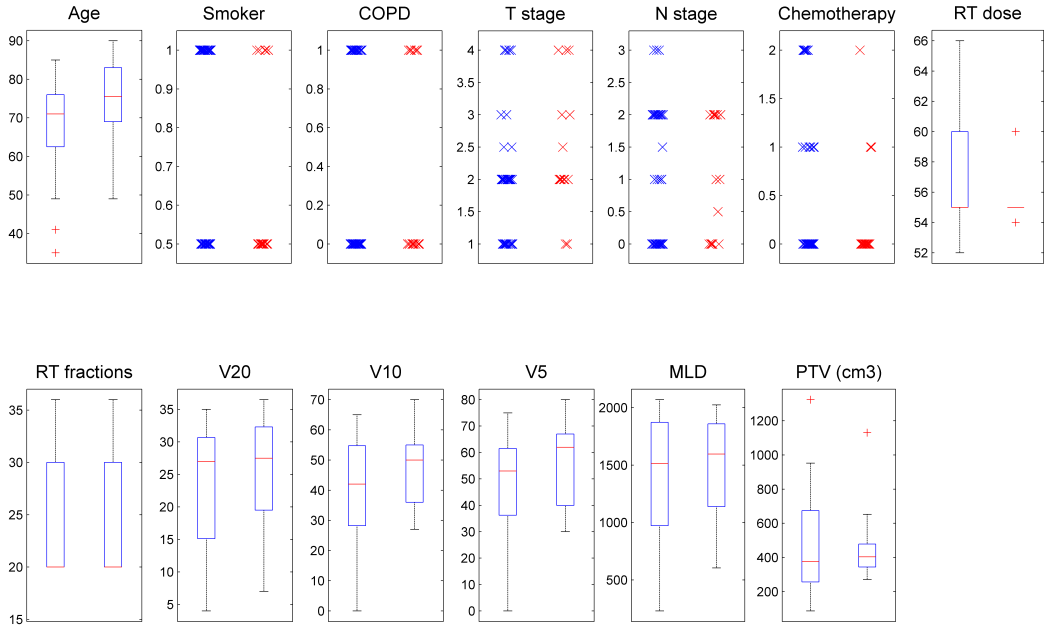


Figure 5.3: Plots showing the variation in clinical parameters between the two classes (symptomatic and asymptomatic). Box plots have been used for the non-categorical features and scatter plots have been used for the categorical feature as box plots were found to collapse and provide misleading results. In each plot the box plot on the left is the asymptomatic cases (blue in the scatter plots) and the symptomatic cases are on the right (red in the scatter plots). The plots show that no single feature is capable of separating the two classes. Note that a small amount of noise has been added to the class labels on the scatter plots to visualise the density of overlapping points.

5.6 Results

This section presents the results from a variety of methods that were employed to test the hypothesis that it is possible to use image analysis of radiotherapy planning CT scans to predict the occurrence of radiation induced pneumonitis. Several different approaches were undertaken, beginning with the whole lung volume. Subsequent analysis was then performed on sub-regions of the lung: regions near the GTV and regions defined by the planned dose. The results in this section demonstrate that all three of these methods are capable of predicting the occurrence of radiation induced pneumonitis.

Because of the size of the data set ($n = 57$, $n_{sym} = 14$, $n_{asym} = 43$) leave one out classification was used. If a simple split into a training and test set was employed, there would be very few symptomatic cases to test with (4 cases, using a typical 70/30 train/test split). However, by using a leave one out approach it is possible to maximise the data - every case can be used as a test case and every other case can be used for training (or cross validation). However, the usual disadvantages with this approach apply: a new model must be trained for each test case.

A number of classifiers were tested in this work, including a Naive Bayes Classifier (NBC), Support Vector Machines (SVM), Decision Trees and k-Nearest Neighbours (kNN). The NBC and kNN classifiers were not capable of reliably predicting the occurrence of radiation induced pneumonitis on the whole lung data set and so these two classifiers were not used on the subsequent permutations of the data. With a larger corpus of data, these classifiers may be of use and should be reconsidered for any real world application of the work presented in this chapter. Additionally, once it was demonstrated that the Decision Tree performance was, at least on par with the SVM performance for the whole lung volume data set, the SVM classifier was dropped from subsequent analysis. This was due to the much faster run time of the Decision Tree algorithm versus the prohibitively slow SVM - the Decision Tree method provided a (approximately) 100 times speed up. It should be noted that the SVM was slow, primarily, because of the cross validation grid search to optimise its internal parameters for each step in the leave one out process. Pre-calculating these values from the data was not an option as that would have been "cheating" by using the same data for training and cross-validation.

Other variables in the classification scheme are the number of grey levels the image was rescaled to before texture calculation (8, 16, 32 or 64) and whether a balanced training set was used. That is, whether all the available cases (56) were used at each step of the leave one out training, or if a balanced number of cases were used for training. For example, if the test case that was removed from the leave one out pool was asymptomatic, then a balanced training set would consist of all 14 symptomatic cases and 14 randomly sampled asymptomatic cases. If the test case was a symptomatic case, then 13 cases from each class would form a balanced training set. The main advantage of a balanced approach is to prevent a more prevalent class from dominating the training step and impairing the ability of the classifier to train a model that is capable of representing both classes - a particular problem with Decision Trees. See Section 4.6.1 for more details.

5.6.1 Whole Lung

The obvious first region to calculate texture features from is the whole of the lung volume. The main advantages of this method are that it requires no further preprocessing and that it will make use of as much information within the lung tissues as possible.

Initial results using SVMs are presented, followed by subsequent results using Decision Trees.

SVM

An SVM classifier with a Gaussian kernel was used in a leave one out fashion, with a cross validation carried out for each fold in order to tune the hyperparameters C (the soft-margin parameter) and γ (the exponent of the Gaussian). As an unfortunate side-effect of this process it is much more difficult to ascertain the relative feature importance as the performance of the features is now obscured by the training of multiple SVMs - one for each case.

In order to determine the best performing set of features, the performance of an SVM was evaluated for all combinations of grey levels, of variance retained after PCA, whether a balanced training set was used and of different combinations of the texture features used. This gives a large hyperparameter space over which to assess the predictive ability of this approach. A full list of the different combination of features is laid out in Table 5.3.

Grey Levels	Features Used	Balanced Training	PCA Variance Retained/%
8	All texture	Yes	99
16	All texture with clinical	No	95
32	FOS		No PCA
64	GLCM		
	GLRLM		
	GLSZM		
	Gabor		

Table 5.3: Summary of the different hyperparameters for the SVM classification scheme.

The maximum classification performance achieved was: Area Under ROC = 0.873, Sensitivity = 92%, Specificity = 72%, and Accuracy = 87%. This was achieved by training an SVM with a balanced training set of all the texture features combined with the clinical features, having used PCA to retain 95% of the variance explained by the features. This result and the nine next best combinations (sorted by Area Under ROC) are presented in Table 5.4. The table shows that the balanced approach is likely to yield better results and so the results in Tables 5.5 and 5.6 only consider SVMs trained with balanced feature matrices.

Classifier	Grey Levels	Features Used	Variance retained/%	Sensitivity	Specificity	Accuracy	Area Under ROC
SVM-bal	8	All texture with clinical	95	0.923	0.721	0.772	0.873
SVM-bal	8	All texture with clinical	No PCA	0.929	0.767	0.807	0.867
SVM-bal	8	GLCM	99	0.857	0.767	0.790	0.823
SVM-bal	16	All texture with clinical	99	0.786	0.837	0.825	0.804
SVM-bal	64	Gabor Filter	99	0.857	0.651	0.702	0.804
SVM	8	All texture	99	0.857	0.814	0.825	0.794
SVM	16	All texture	No PCA	0.786	0.721	0.737	0.779
SVM	16	All texture	99	0.786	0.837	0.825	0.770
SVM	64	All texture	99	0.786	0.837	0.825	0.736
SVM	8	All texture	No PCA	0.786	0.767	0.772	0.732

Table 5.4: Top performing SVMs for the whole lung data set.

Table 5.5 displays the average performance of the SVMs trained with balanced data sets. The values have been averaged over the number of grey levels and different values of PCA used (the first and the last column of Table 5.3). This averaging allows a measure of the stability of the texture features and their relation to the occurrence of radiation induced pneumonitis. The average AUROC (0.699) is lower than the top performing classifier, but still demonstrates a promising level of performance.

Table 5.5 shows that there is a increase in performance when the texture features are combined with the clinical features compared to using either set independently. A t-test was used to determine whether the increase in performance, when using both feature sets together, was a statistically significant improvement over using the clinical features alone. The performance improvement when comparing "Texture with clinical" vs "Texture alone", however, was not found to be statistically significant. These results can be seen in Table 5.6. These results indicate that the texture features are capable of adding new information in addition to the information provided by the clinical features.

	Specificity	Sensitivity	Accuracy	Area Under ROC
All texture features with clinical	0.6899	0.7321	0.7003	0.6990
All texture features	0.6705	0.6964	0.6769	0.6521
Clinical features	0.6570	0.5655	0.6345	0.5815

Table 5.5: Average performance of the SVMs trained with a balanced training set. The table displays an improvement in average performance when the texture features are combined with the clinical features.

"Texture with Clinical" vs "Clinical"	Specificity	Sensitivity	Accuracy	Area Under ROC
Significant Increase	No	Yes	Yes	Yes
p-value	0.2067	0.0101	0.0084	0.0136

"Texture with Clinical" vs "Texture"	Specificity	Sensitivity	Accuracy	Area Under ROC
Significant Increase	No	No	No	No
p-value	0.2839	0.2304	0.2127	0.1144

Table 5.6: p-values from a t-test comparing the performance of SVMs that were trained with either the texture features combined with the clinical features or trained with one of the individual feature sets. The results presented are for the SVMs trained with a balanced training set. The p-values demonstrate a significant increase in performance.

Decision Trees

A Decision Tree classifier was used instead of the SVM classifier above in an attempt to speed up the classification run time. Furthermore, using a different classifier may validate the use of the chosen features - if multiple learners are capable of achieving similar levels of performance, then it further suggests that the underlying features have predictive power.

The Decision Tree classifier was run in the same manner as the SVMs - a leave one out approach was used to maximise amount of training and test data at each stage and a balanced data set was used for each training step. It should also be noted that cross validation step was not carried out on a per-case basis. Instead a preliminary grid search over sensible Decision Tree parameters was performed. It is often simpler to interpret the parameters of a Decision Tree than those used in an SVM and so a range of parameters were chosen in an attempt to prevent over-fitting. For example, it is possible to reason about how shallow a tree should be to avoid over fitting.

Table 5.7 shows the top performing Decision Tree classifiers (compare with the SVM results in Table 5.4). The top performing classifier showed a Sensitivity of 93%, a Specificity of 79% and an Area Under ROC of 90%. This demonstrates performance of the Decision Tree classifier was at least comparable to that of the SVM based approach.

Figure 5.4 shows the AUROC performance of the trained models (using only the GLCM features) as the number of components selected during PCA and the number of grey levels used for texture calculation varied. The Figure demonstrates that the performance is relatively stable across these two parameters. The model parameters used in training of the Decision Trees shown in the Figure were a max depth of 2, a minimum samples per split of 4 and a minimum samples per leaf of 7. These parameters produce a shallow tree with (given the size of the data set) fairly strict requirements on the number of samples at each leaf node. This should mean that the tree was not over fitted - this suggests that the method should be generalisable on a larger data set.

Discussion

The results presented here (using texture calculated on the whole lung volume) demonstrate that it is possible to predict the occurrence of radiation induced pneumonitis using texture features with clinical features or with texture features alone. Additionally, both the SVM and the Decison Tree classifiers perform well on this data set. Both these facts suggest that there is an underlying difference in the lung tissue of a patient that remains asymptomatic versus a patient that develops radiation induced pneumonitis and that this difference can be characterised using texture features.

Given that performance of the Decision Trees are comparable to the results from the SVMs, the Decision Tree classifier will be preferred over the SVMs for subsequent results. This is primarily due to the much shorter training time requried for the Decision trees.

Grey Levels	Features	PCA Components	Area Under ROC
64	GLCM	7	0.9020
64	GLCM	9	0.8945
64	GLCM	0.95	0.8870
8	GLSZM	9	0.8704
64	GLCM	8	0.8696
8	GLCM	8	0.8530
32	GLCM	8	0.8430
64	GLSZM	8	0.8239
8	GLCM	9	0.8231
64	All texture	9	0.8189

Table 5.7: Top performing Decision Tree Classifiers. The PCA column corresponds to the number of new features retained after PCA if the value is an integer. If the value is less than one, then it corresponds to the variance retained after PCA.

The following sections present the results from applying the same approach to different subsections of the lung volume: the region surrounding the GTV (5.6.2) and the V_{20} (5.6.3).

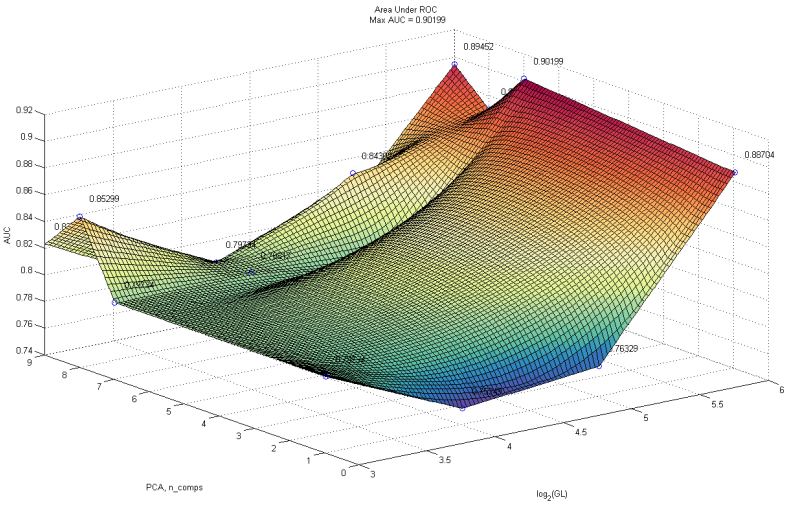


Figure 5.4: Illustration of the stability of a Decision Tree classifier as both the number of grey levels used for texture calculation and the number of PCA components retained varies. This stability is suggestive of a significant underlying difference in the lung tissue of the symptomatic and asymptomatic populations.

5.6.2 Region Surrounding the GTV

After the GTV, the surrounding lung is the area that receives the largest dose during radiotherapy. Indeed, some of the surrounding tissue will be within the defined PTV and will receive a therapeutic dose. It may be the case that the lung tissue surrounding the GTV plays a more significant role in the pneumonitis outcome than the rest of the lung volume.

To investigate if features derived from this sub-region of the lung volume were capable of predicting pneumonitis, the parenchyma in a 20mm strip around the GTV was extracted. Texture was calculated on this smaller volume in the same manner as the whole lung (with the volume padded with the HU values equivalent to air to create a regular shaped array). The texture features were then fed into the same classification pipeline as before.

An ancillary advantage of this approach is a quicker run time due to the analysed volume being significantly smaller than the whole lung volume. However, one significant disadvantage is that the GTV must be segmented for this approach to work. This means that this method could not be used on diagnostic CT scans as they are not usually segmented in the normal clinical work flow - the GTV is only segmented on the planning CT scan to enable a radiotherapy plan to be created. In effect, this means that the prediction of pneumonitis outcome could only be made once the decision to use radiotherapy has been taken.

Decision Trees

As described above, only the Decision Tree classifier was used on this data set. The results can be seen in Table 5.8. These results show that while the performance is comparable to that of the whole lung results, there is a small drop in performance. Further work is required to determine if this method performs consistently worse than the whole lung approach or if it is an insignificant variation. In either case, the texture calculated from this sub region of the lung volume is able to discriminate between the symptomatic and asymptomatic subjects.

These results further reinforce the notion that there is any underlying difference in the lung tissue in the two populations and that this difference can be characterised by texture features derived from CT data.

Grey Levels	Features	PCA Components	Area Under ROC
16	GLSZM	9	0.8738
8	GLRLM	8	0.8480
16	GLRLM	0.95	0.8040
16	GLRLM	9	0.7940
8	GLCM	0.95	0.7899
8	GLCM	9	0.7890
32	GLRLM	9	0.7874
64	GLRLM	9	0.7874
64	GLCM	8	0.7807
16	GLRLM	8	0.7799

Table 5.8: Top performing Decision Tree Classifiers for the region near the GTV data set. Texture features were calculated on a 20mm strip of parenchyma extracted from the whole lung volume. The PCA column corresponds to the number of new features retained after PCA if the value is an integer. If the value is less than one, then it corresponds to the variance retained after PCA. Compare with the whole lung results in Table 5.7.

5.6.3 Dose Map Information

An alternate approach to the region near the GTV method from the previous section is to leverage the dose information from the radiotherapy plan to extract the lung volume to be used for feature extraction. The V_{20} is believed to be an important factor in the pneumonitis outcome so it may be reasonable to assume that texture features extracted from this region may have predictive power. To this end, the dose map was extracted from the DICOM data and used to calculate the region that received 20% of the dose. The lung parenchyma from this region was then extracted from the CT data and the same texture features as before were calculated. The features were then fed into the same classification scheme. The Decision Tree classifier was again used for the same reasons as outlined above.

Decision Trees

As per the previous two regions used, only the Decision Tree classifier was used on this data set. The results can be seen in Table 5.9. The results in the table are poorer than the previous two methods - the top performing classifier only manages an Area Under ROC of 0.72. This is still a respectable level of performance, but is not as high as the performance demonstrated by the whole lung and the region near the GTV approaches.

As with the results in the previous section, further work on a larger data set would be required to determine if this method is consistently less reliable than the previous two. Despite the drop in performance it was still possible to use texture features calculated from the V_{20} to predict whether a subject would develop radiation induced pneumonitis after radiotherapy.

Grey Levels	Features	PCA Components	Area Under ROC
16	GLCM	7	0.7234
16	GLCM	0.95	0.7209
8	All Texture	0.95	0.7168
32	All Texture	0.95	0.7101
8	All with Clinical	7	0.6993
16	All Texture	7	0.6985
16	GLCM	0.95	0.6885
64	GLCM	7	0.6752
16	All Texture	0.95	0.6586
64	All Texture	7	0.6561

Table 5.9: Top performing Decision Tree Classifiers for the V_{20} data set. Texture features were calculated on the parenchyma extracted from the V_{20}. The PCA column corresponds to the number of new features retained after PCA if the value is an integer. If the value is less than one, then it corresponds to the variance retained after PCA. Compare with the whole lung results in Table 5.7.

5.7 Discussion

This chapter has shown that it is possible to use texture features (alone or combined with a small number of clinical features) to predict the occurrence of radiation induced pneumonitis from CT planning scans. This is a novel approach as related research has typically focussed on using a large number of clinical features or attempting to identify already existing lung disease on CT images.

Three regions of the lung tissue were used to calculate texture features: the whole lung; the 20mm adjacent to the GTV and the V_{20}. Texture features derived from each of these regions were capable of predicting the occurrence of radiation induced pneumonitis. This seems to indicate that the region chosen for analysis is not important (or plays a minor role). This may

indicate some level of uniformity of the importance of each lung region in the prediction of pneumonitis. That is, all the lung parenchyma may share the same characteristics that are picked up by the texture calculation and so the area of the lung exposed to radiation may not be relevant - if any of the lung is exposed in such a patient, they may develop pneumonitis.

The performance of multiple learners suggests an inherent stability in the results - multiple learners achieving similar levels of performance should reduce the likelihood of over-fitting and is another reason for further investigation with a much larger data set.

These results are promising and display a proof of concept approach to this problem. However, a much larger data set is required for further testing and optimisation of the approach. The simplicity of the training models used here should allow for the method to be easily scaled up - there should be no time or memory constraints at the classification stage if a much larger data set is used. Any future work with a larger data set should begin with the same classification scheme used here and move on to a more complicated learner if necessary. For instance, no ensemble methods (such as Random Forests or AdaBoost) have been used due to the limited size of the data set available, but these may be suitable if several hundred cases were used.

Several modifications to the methods presented were also attempted, some of them are listed here:

- An adaptive thresholding method was created to remove non-parenchyma regions in the CT data and substitute the mean of the local region. This approach made no noticeable difference and was laid aside in favour of simpler removal only threshold.
- The lung was divided into physiological sections (lobes) to calculate texture on different regions. Averaging over these regions or using a single region did not yield useful results. It is unclear why this approach failed when the dose based method and the region near the GTV method both worked.
- Each case was registered onto a common frame of reference before texture calculation. This was an attempt at a basic normalisation process. This method did not yield any worthwhile results - probably due to the registration deforming the information before the texture features were extracted.
- Sequential Floating Forward Selection [90] was implemented and used with the SVM classifier. However, this resulted in impractical run times (of several days) and was abandoned.

The original CT data was rescaled from 2048 HU to prevent the formation of very sparse GLCMs (and GLRLMS, etc) which can result in poor texture descriptors. The CT data was scaled to 8, 16, 32 and 64 grey levels to test if the number of grey levels used impacted the performance of the classifier. Since this rescaling is carried out after the extracting of the parenchyma, the range of grey levels of the remaining data is already somewhat reduced from 2048 HU to approximately 1000 HU (as can be seen in Figure 5.2). This range can still lead to

sparse GLCMs, and so the rescaling was used. The rescaling used was a linear mapping:

$$CT_{new} = \text{round}\left((CT - CT_{min}) \times \frac{GL}{CT_{max} - CT_{min}}\right) \tag{5.1}$$

where GL is the number of grey levels in the rescaled image. The performance seems to be uniform across the grey levels used. This seems to be an additional indication of the stability of the approach outlined in this chapter. It may be interesting to further investigate the impact of increasing the grey levels used beyond 64 or to combine features derived from different grey level scalings (however these features are likely to be highly correlated). It seems remarkable that the CT data can still predict pneumonitis outcome when reduced to only 8 grey levels.

The best performing feature set with the SVM classifier trained with whole lung data was made up of all the texture features and the clinical features combined (Table 5.4). Intuitively, this makes sense as this is the data set containing the most texture and clinical information. However, the GLSZM and the GLCM features also perform very well (Tables 5.7, 5.8 and 5.9) across the different regions. It is difficult to speculate upon why each of these descriptors performs well as both the PCA and the leave one out scheme makes it difficult to interrogate the relative feature importance. However it is possible to note that the FOS features are missing - they perform poorly which suggests there is some higher level interaction between pixels that is indicative of the pneumonitis outcome. The differences in the underlying tissues cannot be characterised solely by the simple first order methods. Again, it should be noted that the same sets of features repeatedly performing well indicates that the stability of the underlying relationship between the texture in the image and the pneumonitis outcome.

5.7.1 Avenues for Further Investigation

Avenues for future work in this problem space may include:

- Testing the method on diagnostic CT scans. The disadvantage of using planning CT scans is that they are acquired after the decision to use radiotherapy has been made. However diagnostic CT scans are acquired much earlier in the treatment process and the pneumonitis prediction would be more valuable at this stage. The planning CT data was used in this study as it was the only suitable data set available at the time. The approach should work equally well on the diagnostic scans, but this requires validation.
- The main issue with using PCA or any other feature reduction method is that - when coupled with the ‘leave one out’ classification scheme used in this work - the relative importance of each feature is obscured. This makes it much more difficult to tune the overall system and choose better features for classification.
- Further investigation into the relevance of the different texture features should be undertaken. This would potentially allow for a determination of the underlying tissue changes in the two populations and may even suggest clues as to the underlying biological

differences.

- If follow up CT data were available, it may be possible to further test if there is the possibility for intra-patient prediction. That is, if texture features are generated from sub-images of the lung (similar to the approach used in Chapter 6), it may then be possible to test if these features could be used to identify local regions of the lung that are at higher risk. This approach could then generate multiple class labels per patient, rather than focussing on a global approach where a patient is given a single label. These class labels could then be used during the radiotherapy planning stage as additional planning constraints to attempt to reduce the risk of triggering pneumonitis.
- Testing of this approach using EPID imaging (acquired during radiotherapy) could lead to inter-fraction determination of the risk of pneumonitis. That is, if texture calculated on the regions of the lung from EPID images could be correlated with pneumonitis outcome, it may be possible to identify patients early (during treatment).

5.7.2 Clinical Work Flow

The method presented here could easily fit into the typical clinical work flow used in the treatment of lung cancer. As an example:

1. A CT scan of a patient is acquired for the purposes of radiotherapy planning and stored in the hospital's PACS system.
2. The texture features are calculated by the PACS system. This process takes approximately 20 minutes (or less with native code) and may be run over night or during another period of time when the PACS system is under a light load.
3. The texture features are combined with the clinical factors present in the PACS system and passed to a classifier (trained once off line on historic data).
4. The classifier reports a binary value (or a probability that can be converted to a risk score).
5. The result of the classifier is stored in the PACS system, ready for a clinician to make use of in the planning of the patient's treatment.

5.7.3 Summary

In summary, a proof of concept automated approach to predicting radiation induced pneumonitis has been presented. One distinct advantage of this approach is the minimal effort required for data collection - the clinical information and the CT images are collected in the normal course of treatment for lung cancer and the methods outlined above could easily fit into a standard clinical work flow. The results outlined above are promising and warrant future research with a much larger cohort of patients.

Chapter 6

Automatic Segmentation of Prostate Focal Lesion

6.1 Introduction

As outlined in Section 2.6.4, a dominant unifocal lesion greater than 0.5cm^3 may be present in up to 70% of newly diagnosed prostate cancer patients. There is significant potential for improving treatment by treating this dominant focus or boosting the dose to the area while lowering the dose delivered to the surrounding healthy tissue. There are many challenges associated with this approach, including identifying the focal disease in a manner that enables a radiotherapy planning volume to be delineated, planning a suitable radiotherapy plan, executing on the plan and verifying the new method is curative and improves patient outcome.

This chapter focusses on the first step: attempting to automatically identify the focal disease on T2 MR scans that were acquired for diagnosis. This would be a key first step as it could greatly assist clinicians in the difficult and time consuming task of identifying the focal disease within the prostate. Subsequent steps would involve the registration of the generated contour to a CT scan, and then the use of this contour to generate a radiotherapy plan that meets the dose restraints required to "boost" the disease.

The rest of this chapter will detail the data used, explain the methods and examine the results of the approach. An overview of the model can be found in Section 6.3 and the final results are analysed in Section 6.6.

6.2 Data

The data set used for this retrospective study came from 14 patients with confirmed prostate cancer treated at the Edinburgh Cancer Centre, Western General Hospital between August 2012 and April 2013. T2 MR scans were acquired with a Symphony 1.5T (Siemens, Erlangen, Germany) for the purpose of diagnosis. Each patient was treated with androgen deprivation therapy over the course of three months before returning for gold marker seed implantation (for

Feature	Details	Total
FOS	standard first order features	7
GLCM	14 Haralick features calculated in 13 directions	182
GLRLM	11 standard run length features calculated in 13 directions	143
GLSZM	11 standard features calculated for 5 connectivities	55
LBP	a u2 mapping with 8 neighbours was used on a slice wise basis, with the final feature vector computed as the mean of the vectors	59

Table 6.1: Details of the features calculated for classifying the focal lesion.

alignment) and a planning CT scan in preparation for external beam radiotherapy. The prostate, bladder and rectum were contoured by an experienced consultant oncologist on both the MR and the CT scans and the focal disease was contoured on the MR scan only. It is not possible to identify the focal disease region on the CT images due to the poor soft tissue contrast. Additionally, the whole prostate and the focal lesion shrinks due to the hormone treatment administered between the MR scan and the CT scan, further confounding the identification of the focal lesion. The T2-weighted MR data and the outlined contours form the data set for this investigation. In the absence of histological data, the ground truth labels are the segmentations provided by the expert clinician.

6.3 Model Overview

In order to automatically segment the focal lesion a set of features were extracted using the methods described in Chapter 3. The approach taken was to classify each pixel in the region of interest in order to generate a mask of class labels that can then be converted back to a contour. The region of interest (ROI) was defined as the joint prostate-focal lesion volume (for some cases the focal lesion lay outside of the prostate volume).

This approach requires the calculation of a feature vector at each pixel within the ROI. This was carried out by extracting a 3D box ($5 \times 5 \times 5$) around each pixel and deriving 446 texture features for each subimage. See Table 6.1. This box size was chosen as a starting point in an attempt to balance the trade off between the amount of information contained in a subimage (increases with box size) and the locality of this information (decreases with box size).

These features were used to train a series of AdaBoost models (with Decision Trees as the base classifier) using a 5-fold cross validation scheme (Section 4.6.4) to determine the optimal set of hyperparameters. This optimisation step was carried out as a grid search over a range of the total number of estimators in the model and the maximum depth of each tree. The results of this optimisation step can be seen in Section 6.4.3.

An AdaBoost model with the optimal hyperparameters was then trained and used to classify each case in a leave one case out fashion. These classification labels were then mapped back

onto the MR space before being cleaned with the use of morphological operations. The resulting binary masks were then converted to a contour and compared with the original clinical segmentation.

In summary:

1. The prostate and focal lesion are outlined by an expert radiologist to establish the ground truth labels.
2. A set of features ($n = 446$) is calculated on a subimage extracted around each pixel within the prostate (and focal lesion).
3. Feature matrices and class labels are created using the 446 texture feature vectors.
4. A cross validation step is carried out to find the optimal hyperparameters.
5. A leave one out classification process is carried out:
   - The current case is withheld for later testing, while the other 13 cases are concatenated together to form a single training matrix.
   - An AdaBoost model is trained.
   - The trained model is then used to predict class labels on the test case.
   - This is repeated for each case.
6. Predicted labels are subsequently mapped back into the MR volume frame of reference to generate a binary mask of focal disease that can be compared with the original clinical contour.
7. The predicted masks are cleaned up by slice wise morphological operations.
8. Dice coefficients are calculated to assess the agreement between the clinical and predicted masks.
9. The predicted mask is converted to a contour for display.

6.4 Classification

6.4.1 Feature Extraction

A $(5 \times 5 \times 5)$ subimage was extracted around each pixel from the prostate and focal lesion. Texture features (see Table 6.1) were then calculated to generate a feature vector at every point. These feature vectors were labelled “prostate” or “focal lesion”.

In a clinical scenario these calculations could easily be performed off line by a PACS system, post-MR acquisition and before the image is inspected by a clinician, perhaps overnight to avoid unnecessary load on the system. If these features are calculated ahead of time on the whole of the MR scan, then a near real time segmentation becomes possible once a clinician has outlined the prostate: all the points within the prostate are classified and a predicted focal lesion is displayed on the screen.

All algorithms used for texture calculation were implemented in MATLAB by the author or another member of the author's group.

6.4.2 AdaBoost Classifier

An AdaBoost meta-classifier (Section 4.3.6), using Decision Trees (Section 4.3.4) as the base classifiers, was chosen for this application. AdaBoost trains a series of base classifiers, focussing on previously misclassified training samples, to achieve better performance on subsequent iterations. The AdaBoost and Decision Tree parameters used in this work are shown in Table 6.2, more details of the role of each parameter are provided in Sections 4.3.4 and 4.3.6.

AdaBoost

Number of Base Learners	100
Learning Rate	1

Decision Trees

Criterion	Entropy
Maximum Depth	50
Minimum Samples: Leaf	20
Minimum Samples: Split	50

Table 6.2: Table showing the parameters used in the AdaBoost model used for the focal lesion classification. The decision tree parameters are those used for each of the base learners.

The number of learners and maximum tree depth were determined by a grid search with a 5-fold cross validation strategy as described in the next section. PCA was included in this grid-search, but when PCA was used the results were uniformly and significantly poorer and are not included here. PCA was not included in the classification pipeline.

The Scikit-learn [85] implementation of AdaBoost was used in this work.

6.4.3 Model Development Using Cross Validation

In order to test the classification model and to determine if the texture features were capable of discriminating between prostate and focal disease tissue 5-fold stratified cross validation was employed and used to search for the best combination of model parameters, this cross validation strategy is outlined in more detail in Section 4.6.4. Due to memory constraints a traditional stratified cross-validation strategy was not used, rather, each case was individually "balanced" before the cross validation steps. The balanced feature matrices for all 14 cases were grouped together into one consolidated data set for this test. A random subset of the data was used for balanced training and testing was carried out on the remaining data. This was then repeated for a total of 5 "folds", with replacement. At each fold a series of classification performance metrics (Sensitivity, Specificity and Area Under ROC) were calculated to assess the performance of the classification system.

The advantage of this cross validation step is that it allows quick and efficient prototyping and testing of classification models and their stability over a series of trials. The results of the grid search can be seen in Table 6.3.

Number of Estimators	Maximum Depth	Sensitivity	Specificity	Area Under ROC
10	1	0.6052	0.5799	0.6302
10	10	0.5422	0.6755	0.6501
10	50	0.5150	0.7014	0.6506
10	100	0.5150	0.6999	0.6496
100	1	0.6014	0.6412	0.6604
100	10	0.5181	0.7152	0.6612
100	50	0.5259	0.7497	**0.69105**
100	100	0.5278	0.7472	0.6884
500	1	0.5916	0.6566	0.6652
500	10	0.5265	0.7478	0.6895
500	50	-	-	-
500	100	-	-	-
1000	1	0.5831	0.6607	0.6619
1000	10	-	-	-
1000	50	-	-	-
1000	100	-	-	-

Table 6.3: Performance metrics from the grid search over the number of estimators and the maximum depth parameters in the AdaBoost model. Some classifiers (blank entries) were terminated early due to exceedingly long run times. Each aborted run was allowed to continue for 48 hours before termination. See Appendix B for full results.

Once the optimal maximum depth and number of estimators were found, the cross validation strategy was repeated with 10 folds. The performance metrics can be seen in Figure 6.1 and Table 6.4.

The scores in Table 6.4 show reasonable performance from the classifier. Note that there are equal numbers of each class due to the the balancing carried out prior to the cross validation strategy. This has the side effects of making the classification metrics simpler to interpret and of losing a significant portion of the data available for testing - this is because the focal lesion is much smaller than the whole prostate volume[i] and the balancing of the data set (necessary due to memory constraints) discards a large number of the prostate test examples.

The model is better capable of correctly identifying the prostate tissue than the focal disease tissue. That is, the specificity is consistently higher than the sensitivity - the model tends to favour classifying points as “healthy”. The use of the stratified cross validation removes the possibility that this is due to a training bias produced by a majority of samples being from the

i. On average, the focal lesion makes up 14.6% of the total prostate volume. Notably, this figure is 69% for case 4, 44% for case 5 and 29% for case 9. If these three cases are excluded, the mean ratio falls to 5.7%

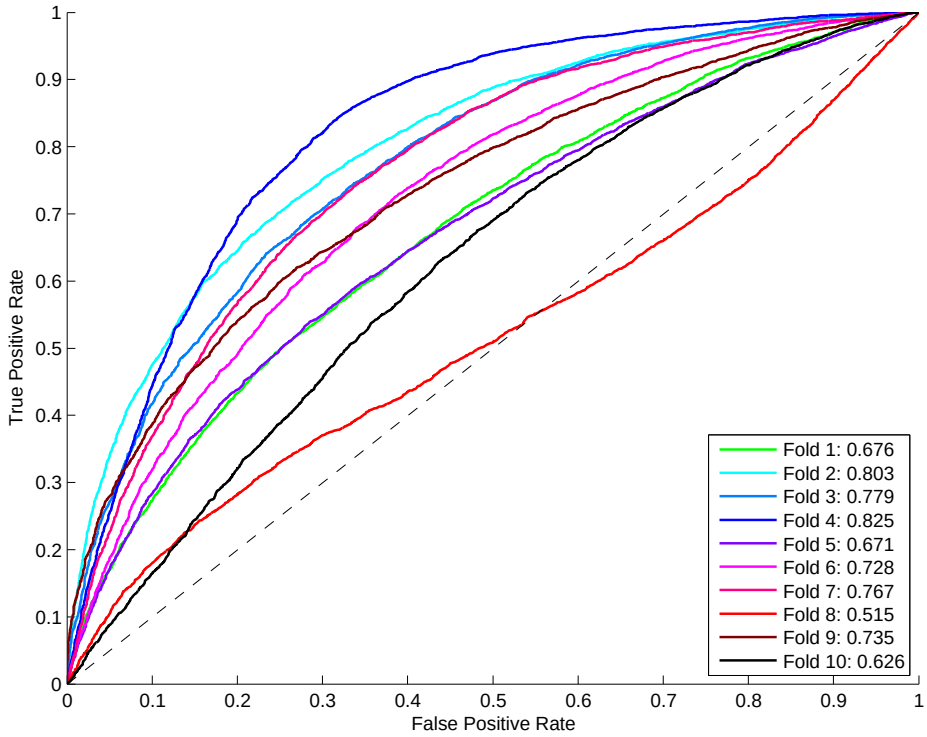


Figure 6.1: Receiver operator curves for each fold of the 10-fold stratified cross validation test using features calculated on all 14 patients as the consolidated feature matrix. See Table 6.4 for detailed performance metrics from each fold. The mean area under ROC $= 0.71$ demonstrates promising performance with this classification problem.

prostate. This difference is inherent in the model. It is reasonable to suppose that the classifier is better capable of representing the prostate tissue as it will be more regular and consistent than the focal disease tissue: by definition a tumour is an out of control growth of tissue. Allowance must also be made for class label noise - the ground truth labels provided by the expert radiologist are not absolute - histological information registered back to the MR volume would provide more certain identification of the underlying tissue type (class label) at each point in the MR image.

The graphs and the table of results demonstrate that the performance of the classification model is fairly stable with a small number of poor performing folds (fold 8, AUROC $= 0.5149$). As can be seen later in this Chapter, there are 3 cases (10, 11 and 12 - see Figure 6.2) that perform poorly, one possible explanation for the poor performance on fold 8 is that more of the samples used in this fold were drawn from these cases. The performance of these cases are discussed in more detail later in the chapter. The small standard deviations for the Specificity and AUROC also demonstrate the stability of the model. This is a good indication of the veracity of the methodology so far - the hypothesis that the two tissues can be discriminated by the texture features seems plausible at this stage. Additionally, the much larger standard deviation (0.1775) and range (0.6485) for the Specificity show that the model is much better at

Fold	TP	TN	FP	FN	Sens	Spec	Area Under ROC	Precision	Recall	Area Under PRC
1	3281	4469	1745	2933	0.5280	0.7192	0.6758	0.6528	0.5280	0.6714
2	4591	4417	1797	1623	0.7388	0.7108	0.8027	0.7187	0.7388	0.7957
3	3896	4811	1403	2318	0.6270	0.7742	0.7789	0.7352	0.6270	0.7666
4	5436	3971	2243	778	0.8748	0.6390	0.8254	0.7079	0.8748	0.7868
5	2525	5135	1078	3688	0.4064	0.8265	0.6714	0.7008	0.4064	0.6669
6	3278	4829	1384	2935	0.5276	0.7772	0.7285	0.7031	0.5276	0.7054
7	3856	4757	1456	2357	0.6206	0.7657	0.7672	0.7259	0.6206	0.7395
8	1406	5341	872	4807	0.2263	0.8597	0.5149	0.6172	0.2263	0.5475
9	3081	5158	1055	3132	0.4959	0.8302	0.7353	0.7449	0.4959	0.7488
10	3121	4160	2053	3092	0.5023	0.6696	0.6256	0.6032	0.5023	0.5909
Mean	-	-	-	-	0.5548	0.7572	0.7126	0.6910	0.5548	0.7019
Min.	-	-	-	-	0.2263	0.6390	0.5149	0.6032	0.2263	0.5475
Max.	-	-	-	-	0.8748	0.8597	0.8254	0.7449	0.8748	0.7957
Std. Dev.	-	-	-	-	0.1775	0.0720	0.0935	0.0494	0.1775	0.0831

Table 6.4: Detailed classification performance, using a range of metrics, for each fold of the 10-fold stratified cross validation test using features calculated on all 16 patients as the initial feature matrix. Data corresponds to the curves in Figure 6.1. The mean values of each metric demonstrates promising performance while the standard deviations display the stability of the model.

correctly identifying healthy tissue than diseased tissue.

Within each cross validation fold the training and test sets are made up of samples drawn from all the 14 cases available. Although these results are encouraging, further investigation into mapping the classification labels back to the original frame of reference of the MR scan for comparison with the original contours is required for any clinical application.

6.4.4 Leave One Case Out Testing

The next step after initial model validation using cross validation is to train and test in a manner which allows for the classification labels to be mapped back to their corresponding location on the MR images. As there were only 14 cases available to this study a leave one case out approach was chosen. This approach maximises the number of test instances available without reducing the amount of training data and allows us to generate a set of contours for every case.

Balanced feature matrices were created for each case, as described above. Then a single case was removed from the pool of feature matrices and designated the test case, the full, unbalanced, feature matrix for this test case is then loaded in order to generate predicted class labels for the whole of the ROI. The remaining 13 cases were then combined into a single feature matrix and used to train the AdaBoost classifier. Predicted labels were generated for the test case and the case returned to the pool, the AdaBoost model discarded and the process repeated for all the remaining cases. This results in predicted labels being generated for every sample in the data set without compromising the training or test steps. At this stage, a limited assessment of the results can be made as no direct comparison to the clinical contours can be made until

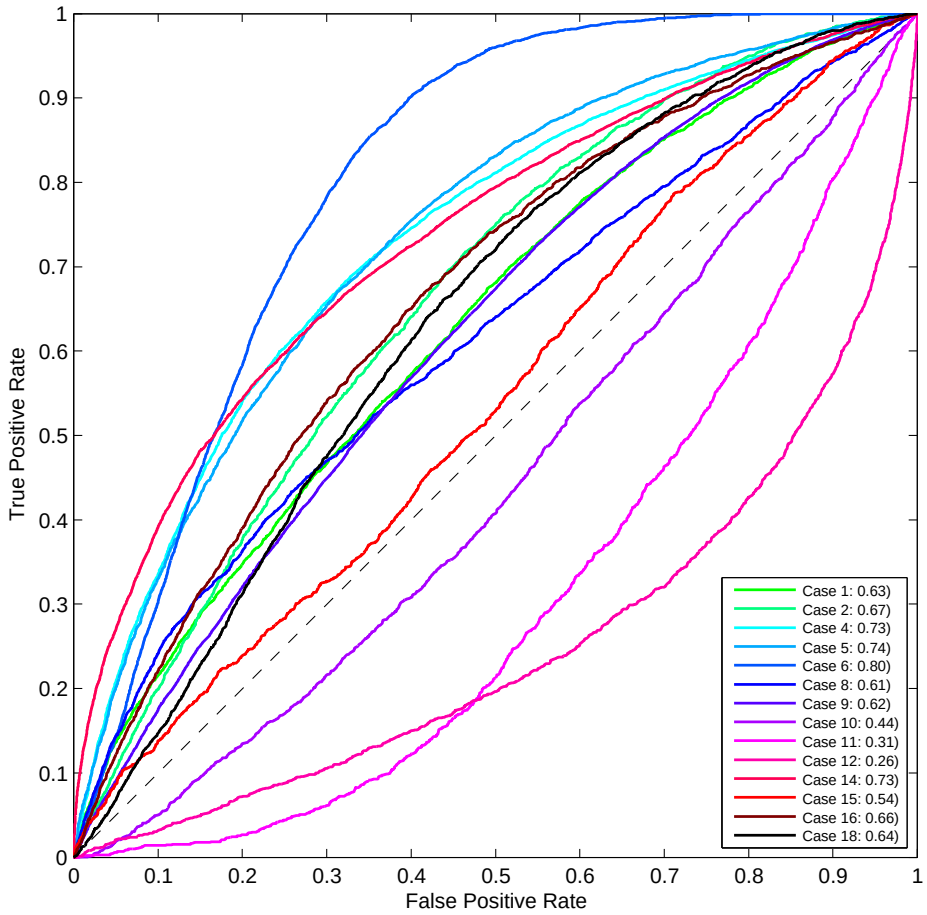


Figure 6.2: Receiver operator curves for each step of the leave one case out process. The legend provides the area under the curve for each instance. See Table 6.5 for detailed performance metrics from each case.

the labels are converted to a mask and contour.

The classification results can be seen in Figure 6.2 and Table 6.5. The mean area under ROC (AUROC) is 0.6, but the performance of the classifier varies greatly among the cases with a minimum of 0.26 and a maximum of 0.80. This variation among the cases should be examined further and shall be discussed in Section 6.7.

The four best performing cases are cases 4, 5, 6 and 14. Each of these cases have an AUROC $>$ 0.73 and can be seen towards the top left of ROC in Figure 6.2. It should be noted that Case 4 has a very large amount of disease - there are 17,177 diseased pixels and 7,777 healthy pixels. In other words, 69% of the prostate volume is cancer tissue. This can no longer be considered "focal disease" and this particular case is unlikely to be suitable for a boosted treatment plan to spare the healthy prostate. However, it is encouraging that the classifier still performs well on this case and it should be noted that it may still be possible to boost the dose to this region if a confident determination of the disease can be made. This would not help spare the healthy

Case	TP	TN	FP	FN	Sensitivity	Specificity	Area Under ROC	Precision	Recall	Are Under PRC
1	1038	77275	18006	2058	0.3353	0.8110	0.6314	0.0545	0.3353	0.0552
2	1603	12746	9479	779	0.6730	0.5735	0.6664	0.1446	0.6730	0.1563
4	10819	5653	2144	6358	0.6299	0.7250	0.7333	0.8346	0.6299	0.8440
5	4401	6120	2609	2349	0.6520	0.7011	0.7378	0.6278	0.6520	0.6638
6	4175	21849	17462	309	0.9311	0.5558	0.8033	0.1930	0.9311	0.2376
8	866	119383	21835	1881	0.3153	0.8454	0.6120	0.0381	0.3153	0.0332
9	7742	31047	12281	10381	0.4272	0.7166	0.6211	0.3867	0.4272	0.3804
10	44	81879	5147	1822	0.0236	0.9409	0.4432	0.0085	0.0236	0.0174
11	35	9582	2376	1312	0.0260	0.8013	0.3128	0.0145	0.0260	0.0674
12	44	95733	10498	1421	0.0300	0.9012	0.2552	0.0042	0.0300	0.0083
14	2958	56831	10791	3055	0.4919	0.8404	0.7343	0.2151	0.4919	0.2698
15	631	8162	5766	779	0.4475	0.5860	0.5395	0.0986	0.4475	0.1102
16	1343	50025	25301	968	0.5811	0.6641	0.6650	0.0504	0.5811	0.0523
18	1102	37905	15885	1241	0.4703	0.7047	0.6383	0.0649	0.4703	0.0588
Mean	-	-	-	-	0.4310	0.7405	0.6000	0.1954	0.4310	0.2111
Min.	-	-	-	-	0.0236	0.5558	0.2552	0.0042	0.0236	0.0083
Max.	-	-	-	-	0.9312	0.9409	0.8033	0.8346	0.9311	0.8440
Std. Dev.	-	-	-	-	0.2687	0.1212	0.1611	0.2529	0.2687	0.2565

Table 6.5: Detailed classification performance, using a range of metrics, for each case in the leave one case out process. The data corresponds to the curves shown in Figure 6.2.

tissue in this case, but the higher dose may itself be of use. The size of the focal disease is further considered in Section 6.7.

But cases 10, 11, 12 perform much more poorly with AUROC of 0.44, 0.31 and 0.26 respectively. These three cases can been seen in Figure 6.2 - they are the three lines that fall below the dashed line that represents randomly guessing the class labels. Examination of Table 6.5 shows that the model classified nearly all the points as healthy, resulting in a very low sensitivity and an artificially high specificity due to the much larger number of healthy pixels. It is not clear why these three cases have performed poorly - the ratio of healthy tissue to prostate tissue does not correlate with the performance. So it is not due to unusually large or small tumour volumes. Future research could investigate this further - it may be that with more subjects the performance on these cases improves or there may be a link between this poor performance and information about the disease state or other demographic data. If such a link were found, it may be possible to train multiple models for different population groups in order to identify the focal lesion reliably.

6.4.5 Feature Importance Investigation

When using an AdaBoost meta-classifier the relative importance of each feature can be determined if the base estimator supports it. The decision trees used in the current work support reporting the relative importance and it is therefore possible to determine the overall importance across the ensemble of decision trees. In the AdaBoost implementation used here the feature importances are defined as the summed importances across all the base estimators.

The feature importance for each step of the leave one out classification can be seen in Fig-

ure 6.3. The level of importance is stable across the cases - all of the lines are in agreement, this suggests that the classification model is not over fitting on each stage of the leave one out scheme, but is consistently finding the same features to be good predictors of focal disease.

The step down at the end of the graph is one set of GLSZM features (connectivity of 26) and the LBP features - these are the least important features. The low importance of the LBP features may be due to the parameters used to create the LBP operator used, or it may be because these features do not take into account the inter-slice information in the same manner as the other features. That is the LBP features are the only ones that are not natively 3D, but are the mean of a set of slice-wise features. It is less clear why there is a sudden drop off in importance for the GLSZM features with a connectivity of 26. A connectivity of 26 means all possible pixels are included in the structuring element. Intuitively, one would expect this to lead to better results and it is currently unclear why this is not the case.

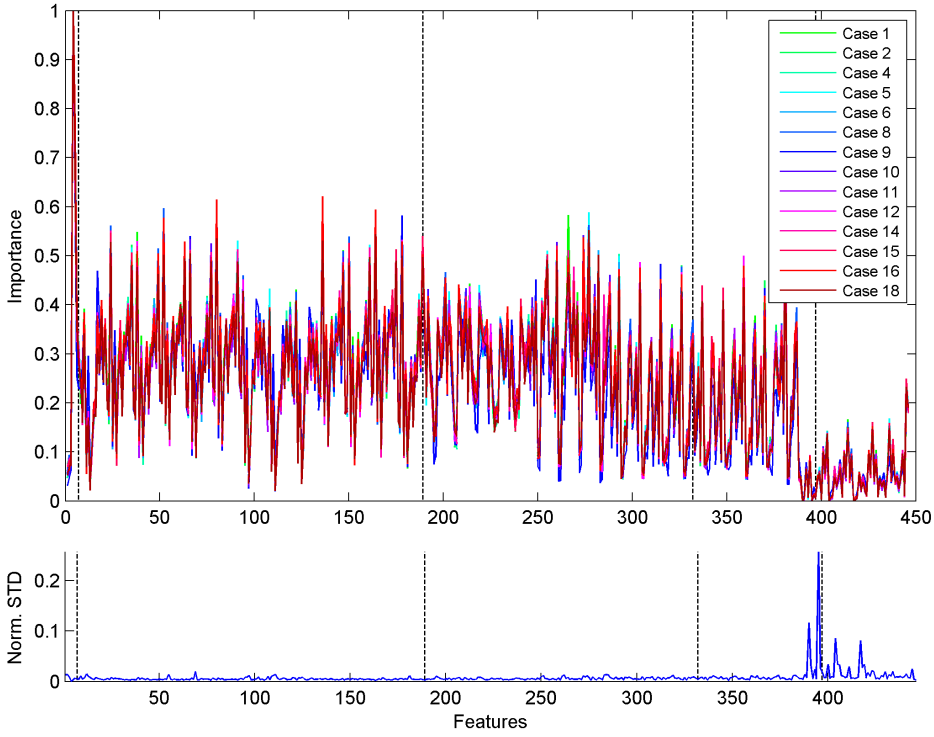


Figure 6.3: Top: Relative importance of the features in the AdaBoost model. Each line corresponds to the test case used. Note that the importances are relatively stable across the cases. (Each feature importance vector was normalised by dividing by its peak value.) Bottom: The standard deviation divided by the sum of the normalised importances for each case. This is a measure of the relative variation across the cases. The black dashed lines demarcate the different types of features - the first region are FOS, followed by GLCM, GLRLM, GLSZM and LBP.

A series of repeating patterns can easily be seen in Figure 6.3 - this is not unexpected as a number of the features are correlated. Take, for example, the GLCM features - there are 14 features calculated over 13 different directions and examination of the plot shows that the tallest peaks in this section of the feature vector are separated by a distance of 14, that is, the same feature repeatedly shows a high level of importance in multiple directions. A similar pattern is also visible for the GLRLM features and the GLSZM features being separated by 11.

If the mean (across the cases) of the normalised feature importance is sorted, the top two features are the skew and the kurtosis (both first order statistics). Of the top 10 features, 7 are the GLCM Correlation in different directions and the remaining GLCM Correlation features are present in the 75 best performing features. The GLCM features dominate, with 16 of top 20 features and 50 of the top 100 being GLCM features. The rest of the top 100 performing features are made up of 39 from the GLRLMs, 8 from the GLSZMs and the first order energy.

The high importance of the two first order features validates their inclusion in the feature set and suggests that it is always worth including them given the triviality of their calculation. Furthermore, the GLCM features clearly play a major role in the classification performance with the GLCM Correlation being consistently important across a range of directions.

Two insights from this data suggest that the texture in the prostate and the texture in the focal lesion are fairly anisotropic: *a*) the high importance of the skew and kurtosis, and *b*) the consistency of the performance of several features when calculated over different directions. If the texture in the two regions had a distinct directionality, then it would be likely that a particular direction would dominate the feature importance list. Perhaps it is more correct to conclude that the texture features that differ between the two regions are isotropic.

6.4.6 Effect of subimage size

The size of the subimages (defined by a cube around each pixel in the prostate ROI) may play a role in the effectiveness of the texture features calculated from the volume. For large boxes, there is much more information at the cost of a smearing effect due to the information being taken from further away from the central pixel. Yet, a smaller box results in a more localised measure at the cost of being unable to encapsulate the same amount of information. Therefore we expect a trade-off with the size of the subimage.

The $5 \times 5 \times 5$ subimage size was chosen as a starting point as there are 125 pixels in the box, each pixel still touches the central pixel (though the corner pixels are only connected by a point). The same texture features were calculated for box sizes of $(6 \times 6 \times 6)$, $(7 \times 7 \times 7)$, $(8 \times 8 \times 8)$, $(9 \times 9 \times 9)$ and an AdaBoost model (with the same parameters used above) applied to these features following the same leave one out scheme outlined earlier.

Figure 6.4 shows various performance metrics (averaged over the 14 cases) as the box size varies. All three measures do not vary greatly with the change in box size. Though there is a

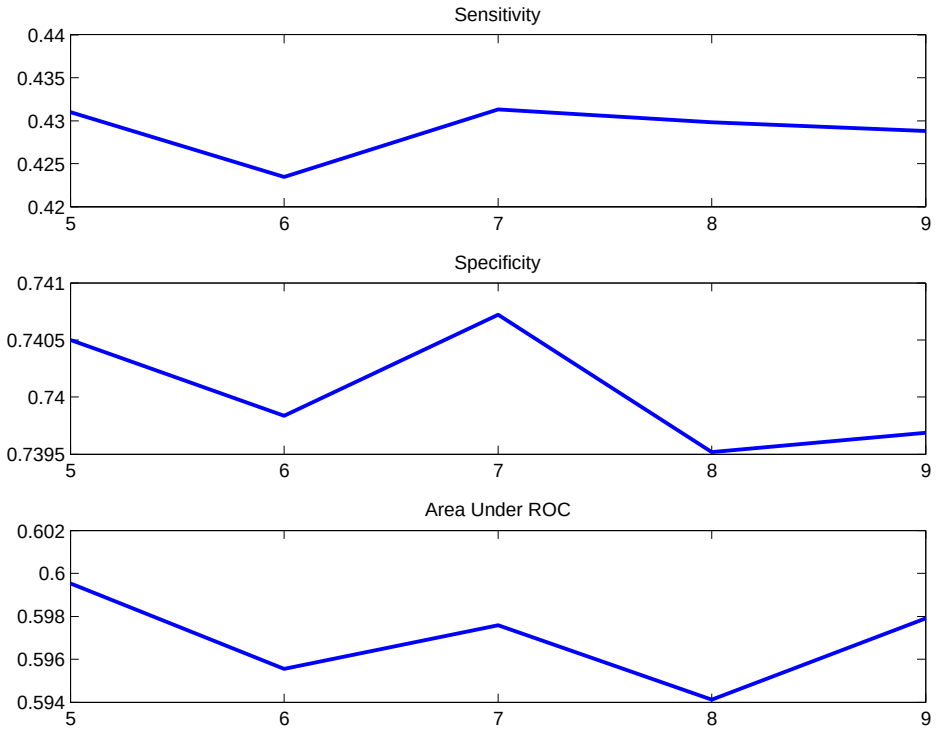


Figure 6.4: Variation in classification performance with different subimage sizes used for texture feature calculations.

drop in the AUROC when moving from subimages with length 5 to 9.

It is also interesting to note that the AUROC is lower for the even numbered subimage sizes. That is the AUROC for $n = 6$ and $n = 8$ is worse than the performance of $n = 7$ and $n = 9$, despite the overall trend of performance decreasing as the box size increases. This is likely due to a simple indexing (without interpolation) being used to create the subimages: the subimages are created by extracting the pixels around the current pixel of interest. When the box size is odd the pixel of interest is located in the centre of the subimage. When the subimage size is even the subimage is off-centre and means that the texture features are not a true representation of the space around the pixel of interest. Interpolation may be used to fix this issue - but this may also cause artefacts which may negatively impact classification performance. The simplest fix for this potential issue is to only used odd numbered box sizes.

Given the results of this analysis of the effect of the subimage size, the $5 \times 5 \times 5$ subimage size will be used throughout the rest of this chapter.

6.5 Predicted Labels Post-Processing

Obtaining predicted class labels for each pixel in isolation is useless - the labels must be transferred back onto the original MR scan for visualisation, evaluation and (eventually) clinical use. This transfer is a simple matter of reconstructing a logical mask from the predicted class labels and their corresponding pixel locations that were stored during the feature matrix generation. Once this mask has been generated it can be overlaid on the MR scan for inspection. However, post processing of these masks is required as they are not suitable for clinical use due to holes, small islands and other artefacts in the masks, such as large numbers of disjoint regions that can cause an unrealistically complex contour to be drawn around the predicted focal lesion.

The masks are post processed by slice wise operations outlined and explained below (positive pixels have been predicted to be focal disease):

- The largest connected region on the slice is found and if it contains fewer than 20 pixels the whole slice is set to zero, otherwise the following operations are applied,
- The slice is morphologically opened (with a square structuring element with a side of length 3) to further remove artefacts,
- The slice finally has a morphological closing operation applied, using a disk structuring element of radius 5, to close any holes in the mask.

Because the classification is carried out on the prostate volume as a whole, a small number of pixels on slices just above and below the true focal disease volume may be falsely labelled as a positive result. An efficient way of dealing with this problem is to examine each slice and apply a threshold to the size largest connected region. The threshold was manually tuned by inspecting slices that contained no true focal disease for predicted focal disease and set to 20 pixels, that is if a slice's largest connected region is less than 20 pixels all predicted focal disease is removed and we move on to the next slice. If the largest connected region is greater than 20 pixels, then a series of morphological operations, described below, are applied to the slice. This threshold could be tuned on a large data set before any real world application.

Morphological Operations

A brief outline of the morphological operations used in this chapter can be found in Appendix A.

In order to process the focal disease masks opening and closing operations are applied to the slices after the initial thresholding step:

- Opening: an opening using a 3×3 ones matrix is applied to the slice. This operation removes sharp edges from the mask and any small islands of isolated regions.
- Closing: a closing is then applied using a disk of radius 5 as the structuring element. This operation fills in the gaps in the masks to ensure the final contour has no holes in it.

6.6 Results

Section 6.6.1 presents the results after the post processing of the predicted class labels. Dice coefficients are calculated to assess the overlap of the clinical mask and the predicted mask in 3D and on a slice by slice basis for each case. These results are presented for the preprocessing pipeline explained in Section 6.5. Subsequent analysis of the generated masks before the morphological processing revealed that the opening step caused too many of the pixels to be removed and so the analysis was repeated with the opening step removed. These results are presented in Section 6.6.2.

6.6.1 Contour Evaluation

The Dice coefficients can be seen in Tables 6.6 and 6.7 and corresponding graphs in Figure 6.5. It should be noted that only slices where there the clinical focal disease contour is present are displayed. The cleaning step (removing all focal disease pixels on a slice if there are fewer than 20 pixels present) was successful in ensuring there was no incorrectly focal disease on any slice that the clinical contour is not present.

The tables and graphs show that the variability seen in the bare classification results have carried through to the contours. The three cases (10, 11 and 12) with very poor classification performance have all resulted in a Dice coefficient of 0 - there is no overlap between the correct contour and the predicted contour. The algorithm has catastrophically failed on these cases by not identifying any of the tissue as diseased.

Figure 6.6 shows the clinical contour and the predicted contour for case 2 (overall Dice $= 0.62$). The algorithm displays a limited capability of predicting clinical disease on this case. Note that the predicted contour is wholly within the clinical contour - this shows a high specificity and low sensitivity, just like the bare classification results.

Figure 6.7 shows case 6 (overall Dice $= 0.97$), this is the case on which the algorithm performs best with a near perfect overlap. Note that the algorithm also displays a higher specificity than sensitivity on this case.

Three dimensional visualisations of some of the results can be seen in Figures 6.8 and 6.9. (The steps in the surfaces are due to the contours being delineated on slices of the image which creates a discrete step change in the 3D contours - these have been smoothed slightly by the interpolation used for plotting). Figure 6.8 shows cases 2 and 6. From case 2 and 6 it is apparent that when the algorithm performs well, the misclassifications are located on the edge of the focal disease, as one would expect.

The relationship between the AUROC and the Dice coefficients can be see in Figure 6.10. The plot suggests that there is a linear relationship between the AUROC and the Dice coefficient once a minimum classification performance has been achieved. And below this threshold there is no correlation between the AUROC and Dice.

Case 1:

Slice	6	7	8	9	10	**3D**
Dice	0	0	0	0	0.230	0.012

Case 2:

Slice	14	15	16	17	18	19	**3D**
Dice	0.406	0.593	0.650	0.364	0.867	0.646	0.620

Case 4:

Slice	15	16	17	18	19	20	21	22	**3D**
Dice	0.723	0.625	0.382	0.419	0.349	0.793	0.284	0.404	0.480

Case 5:

Slice	22	23	24	25	26	27	**3D**
Dice	0.369	0.519	0.760	0.612	0.778	0.404	0.582

Case 6:

Slice	12	13	14	15	16	**3D**
Dice	0.966	0.979	0.989	0.985	0.946	0.977

Case 8:

Slice	11	12	13	14	15	**3D**
Dice	0.169	0	0	0	0	0.0329

Case 9:

Slice	7	8	9	10	11	12	13	14	**3D**
Dice	0.156	0	0.006	0	0.138	0.101	0.010	0.044	0.053

Table 6.6: Dice coefficients computed for comparison of the clinical segmentation and the automatic segmentation. Dice coefficents were calculated in 2D (on each slice) and in 3D (whole patient comparison) for each of the patients. Table is continued in Table 6.7.

Case 10:

Slice	11	12	13	14	**3D**
Dice	0	0	0	0	0

Case 11:

Slice	25	26	27	**3D**
Dice	0	0	0	0

Case 12:

Slice	7	8	**3D**
Dice	0	0	0

Case 14:

Slice	7	8	9	10	11	12	13	**3D**
Dice	0	0.110	0.423	0.264	0.089	0.053	0	0.161

Case 15:

Slice	17	18	19	**3D**
Dice	0.772	0.087	0	0.255

Case 16:

Slice	11	12	13	14	**3D**
Dice	0	0.217	0.571	0.613	0.296

Case 18:

Slice	18	19	20	21	22	23	24	**3D**
Dice	0	0.187	0.366	0.038	0.059	0	0.419	0.153

Table 6.7: Dice coefficients computed for comparison of the clinical segmentation and the automatic segmentation. Dice coefficents were calculated in 2D (on each slice) and in 3D (whole patient comparison) for each of the patients. (Continued from Table 6.6).

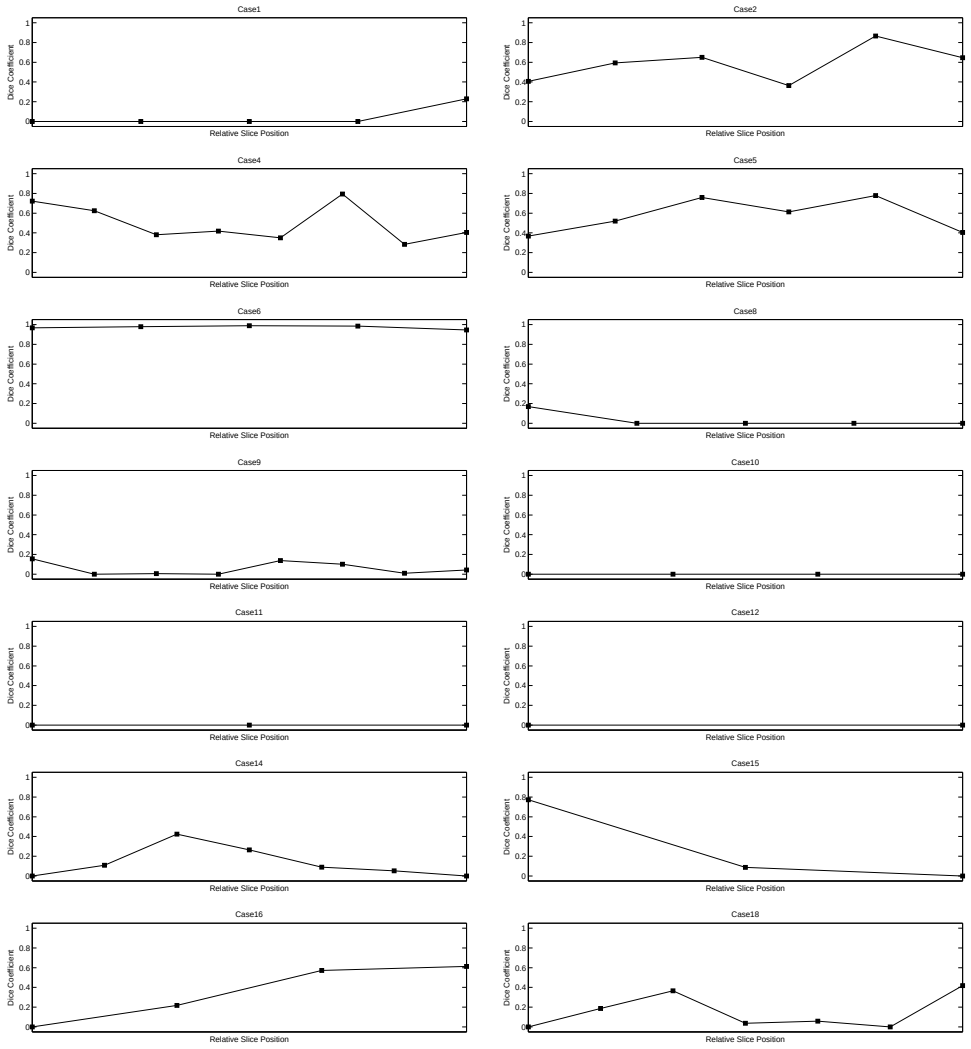


Figure 6.5: Dice coefficients computed for comparison of the clinical segmentation and the automatic segmentation. Dice coefficents are calculated on each slice for all of the patients.

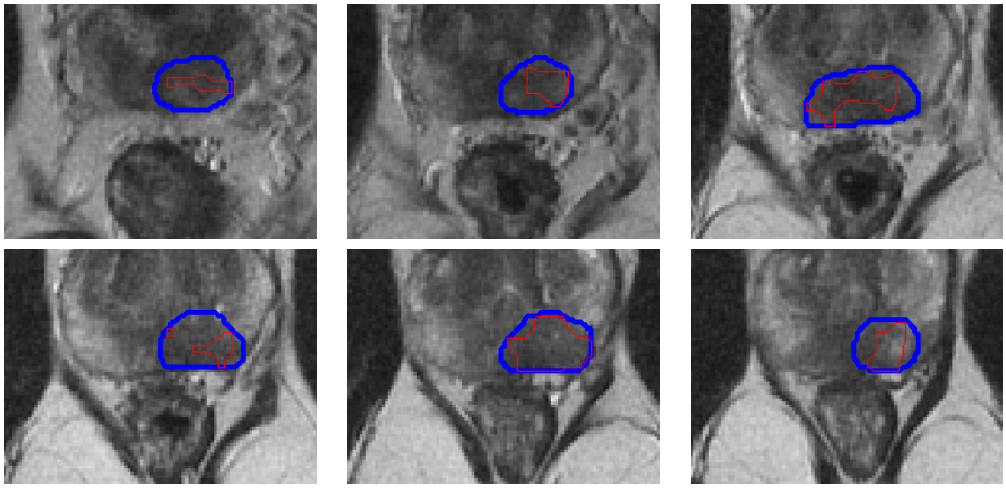


Figure 6.6: Comparison of the clinical (blue) and predicted (red) contours for case 2. Slice numbers increase clockwise from the top left. Note that the blue line is drawn twice as thick as the red line to aid interpretation.

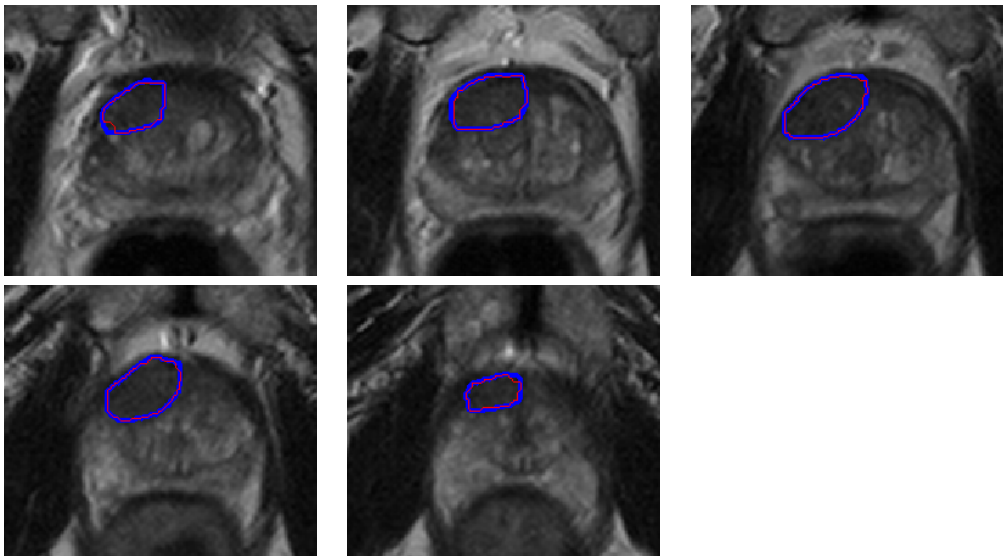
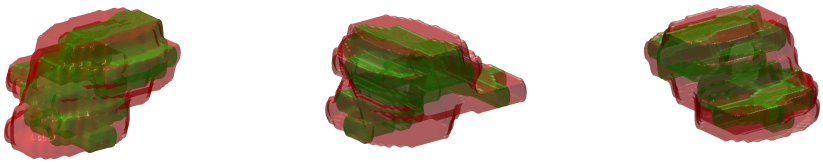


Figure 6.7: Comparison of the clinical (blue) and predicted (red) contours for case 6. Slice numbers increase clockwise from the top left. Note that the blue line is drawn twice as thick as the red line to aid interpretation.

Case 5:



Case 6:

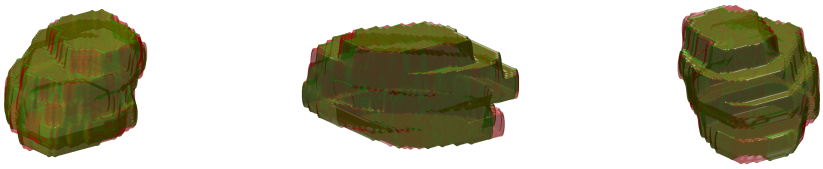
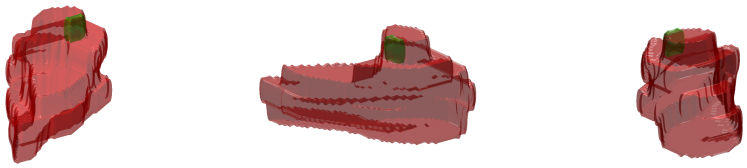
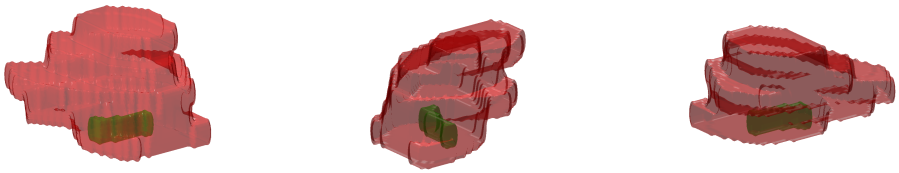


Figure 6.8: A 3D visualisation of the clinical focal lesion mask (red) and the predicted mask (green) for two of the best cases. Figure 6.9 shows the same visualisation for three of the poor performing cases.

Case 1:



Case 8:



Case 9:

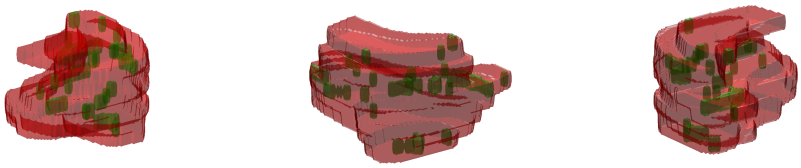


Figure 6.9: A 3D visualisation of the clinical focal lesion mask (red) and the predicted mask (green) for three of the poorer cases. Figure 6.8 shows the same visualisation for two of the better performing cases.

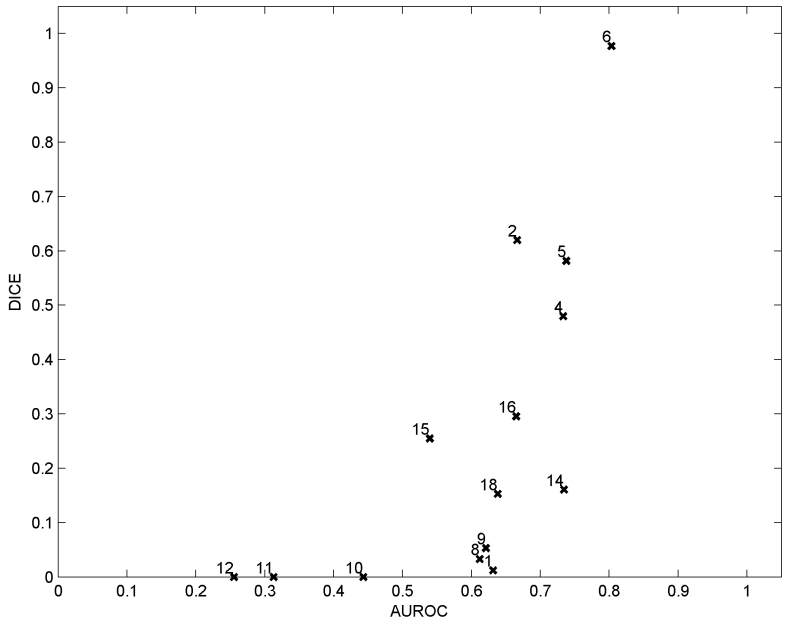


Figure 6.10: A plot showing how the Dice coefficients are related to the Area Under Receiver Operator Curve for each case. The plot shows a steep rise in the Dice coefficient once the AUROC is greater than 0.5.

6.6.2 Contour Evaluation - No Opening Step

In the process of analysing the results presented in the previous section, it was realised that the morphological opening step was discarding a large number of the pixels predicted to be focal disease. An example of this can be seen in Figure 6.11 - the top left image shows the predicted disease in white and the top right image shows the result after the morphological processing described in Section 6.5.

The morphological opening was removed from the processing pipeline and the result can be seen in the bottom left image in Figure 6.11. The clinical mask is shown in the bottom left for comparison. It is immediately apparent that the performance of the algorithm is much better having removed the opening. The rest of this section will present the Dice coefficient results like the previous section, but without the morphological opening step.

The Dice coefficient results for the approach without the opening step are shown in Figure 6.12 and in Tables 6.8 and 6.9. The increase in performance across the cases is immediately apparent. The Dice coefficients rise for 11 of the 14 cases. Notably, and as expected, the performance does not change for cases 10, 11 and 12: a very poor classification performance cannot be rescued with morphological processing.

These much higher Dice coefficients provide much more evidence that this overall approach may be viable in a clinical work flow - there is a high level of agreement between the predicted focal masks and those outlined by an expert clinician.

Figure 6.14 shows the new relationship between the AUROC and Dice coefficients. Comparing this with Figure 6.10 highlights the increase in performance with this new approach and shows that there is still a minimum required AUROC in order to achieve a reasonable Dice score. However the linear relationship between AUROC and Dice that was apparent in the previous plot has been changed significantly: it now suggests that once an AUROC around 0.65 or greater has been achieved, there is no further jump in the Dice performance. Further results would be required to validate this conclusion.

Case 1:

Slice	6	7	8	9	10	**3D**
Dice	0	0.9457	0.9526	0.8941	0.8726	0.8010

Case 2:

Slice	14	15	16	17	18	19	**3D**
Dice	0.9431	0.9828	0.9635	0.9759	0.9911	0.9924	0.9714

Case 4:

Slice	15	16	17	18	19	20	21	22	**3D**
Dice	0.9964	0.9953	0.9967	0.9923	0.9914	0.9947	0.9919	0.9918	0.9937

Case 5:

Slice	22	23	24	25	26	27	**3D**
Dice	0.9901	0.9829	0.9760	0.9901	0.9680	0.9787	0.9825

Case 6:

Slice	12	13	14	15	16	**3D**
Dice	0.9767	0.9839	0.9974	0.9927	0.9950	0.9894

Case 8:

Slice	11	12	13	14	15	**3D**
Dice	0.8913	0.8833	0.9730	0	0	0.7603

Case 9:

Slice	7	8	9	10	11	12	13	14	**3D**
Dice	0.9880	0.9837	0.9817	0.9787	0.9881	0.9727	0.9684	0.9788	0.9804

Table 6.8: Dice coefficients computed for comparison of the clinical segmentation and the automatic segmentation. Dice coefficents are calculated in 2D (on each slice) and in 3D (whole patient comparison) for each of the patients. The morphological opening step has been skipped. (Table is continued in Table 6.9.)

Case 10:

Slice	11	12	13	14	**3D**
Dice	0	0	0	0	0

Case 11:

Slice	25	26	27	**3D**
Dice	0	0	0	0

Case 12:

Slice	7	8	**3D**
Dice	0	0	0

Case 14:

Slice	7	8	9	10	11	12	13	**3D**
Dice	0.8982	0.9883	0.9853	0.9829	0.9783	0.9514	0.9442	0.9660

Case 15:

Slice	17	18	19	**3D**
Dice	0.9353	0.9545	0	0.6821

Case 16:

Slice	11	12	13	14	**3D**
Dice	0.9610	0.9905	0.9739	0.9912	0.9768

Case 18:

Slice	18	19	20	21	22	23	24	**3D**
Dice	0.9049	0.9059	0.9753	0.9658	0.9613	0.9336	0.9369	0.9420

Table 6.9: Dice coefficients computed for comparison of the clinical segmentation and the automatic segmentation. Dice coefficents are calculated in 2D (on each slice) and in 3D (whole patient comparison) for each of the patients. The morphological opening step has been skipped. (Continued from Table 6.8).

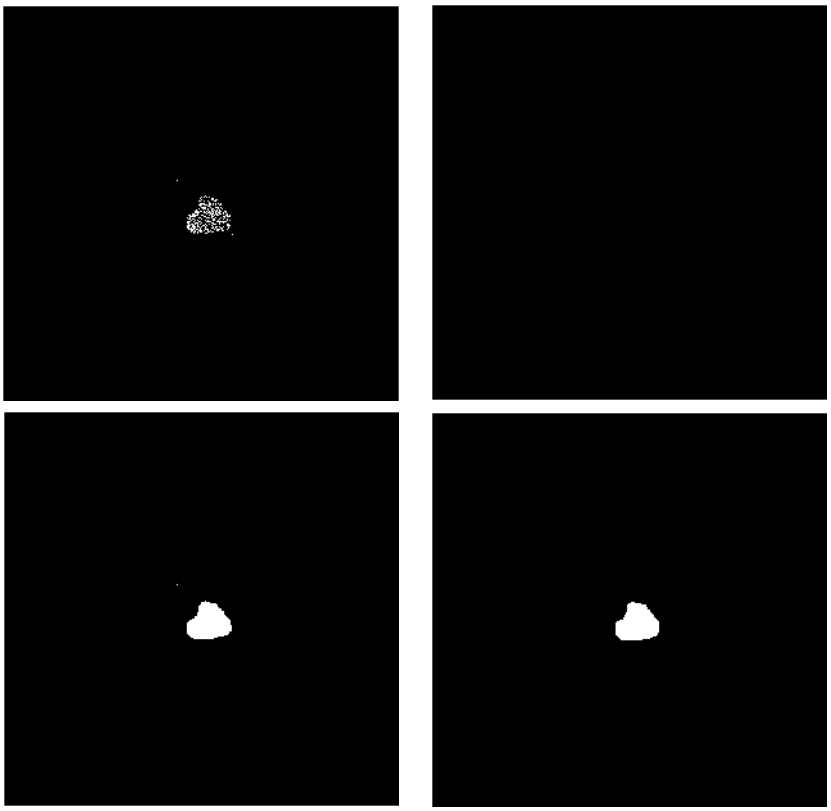


Figure 6.11: Top left: the predicted class labels for a single slice from case 9. Top right: the labels after morphological processing. Bottom left: the same labels, but with the opening step removed from the morphological processing. Bottom right: the reference mask, delineated by an expert clinician for comparison. The performance improvement is immediately apparent.

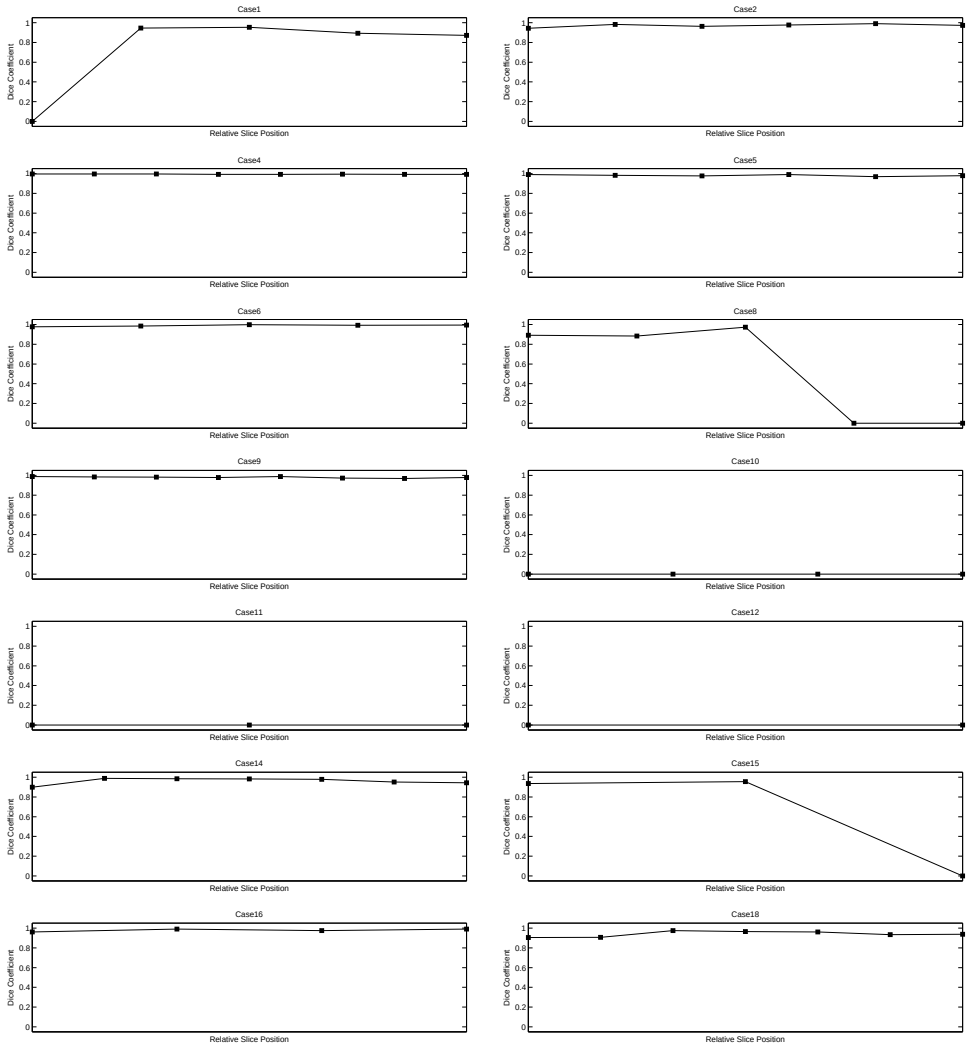
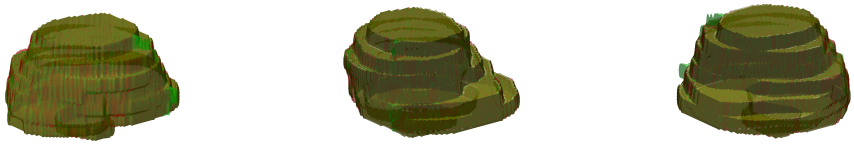


Figure 6.12: Dice coefficients computed for comparison of the clinical segmentation and the automatic segmentation. The morphological opening described in Section 6.5 has been omitted here. Dice coefficents are calculated on each slice for all of the patients.

Case 4:



Case 6:

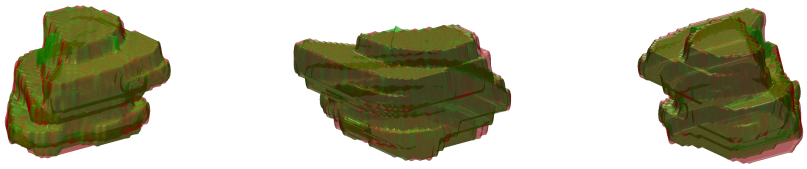


Figure 6.13: A 3D visualisation of the clinical focal lesion mask (red) and the predicted mask (green) for two of the cases with the omission of the morphological opening step. Note that the improved performance is accompanied by some artefacts - small islands of predicted focal disease. These artefacts could be easily removed with a further opening step, or by removing all regions smaller than the largest region.

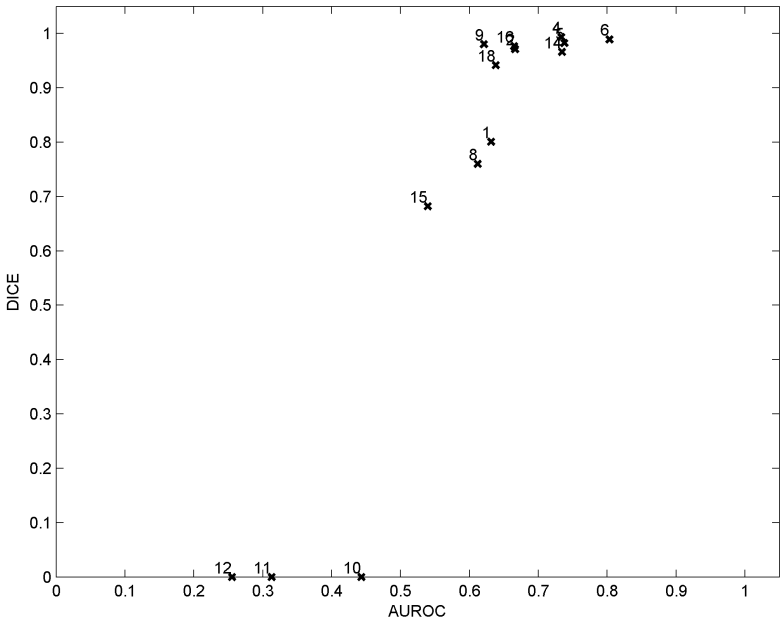


Figure 6.14: A plot showing how the Dice coefficients are related to the Area Under Receiver Operator Curve for each case when the morphological opening step is omitted.

6.6.3 Contour Evaluation - 3D

It is possible to perform the morphological operations carried out in the previous sections in 3D rather than 2D. Performing these operations in 3D is more in line with the overall approach taken in this Chapter as all the texture has already been calculated in 3D. The change to 3D is a simple one - the disc structuring element used before is replaced with a ball structuring element. Given that the slice separation ($\Delta z = 0.54mm$) of the images is not identical with the pixel spacing ($\Delta x = \Delta y = 0.4688mm$), an anisotropic ball element was used to compensate for the larger distances between pixels in the z-direction.

The Dice coefficient results for the 3D method are shown in Figure 6.15 and in Tables 6.10 and 6.11. The results are broadly comparable with those in Section 6.6.2 - overall the 3D method is slightly poorer than the 2D approach (with the exception of case 8 where the 3D method very slightly outperforms the 2D method - an increase of 0.0015). The three failed cases (10, 11 and 12) remain problematic with 3D processing due to the previously discussed effect of low classification performance on these cases. The mean absolute deviation between the two sets of 3D Dice coefficients is 0.0506, while the mean Dices are 0.7110 for the 2D case and 0.6606 for the 3D case.

The author suspects that the results are poorer for the 3D method due to a smoothing effect caused by changing from a 2D to a 3D structuring element that naturally results in a larger structuring element. This results in a smoother predicted contour/volume focal lesion.

Case 1:

Slice	6	7	8	9	10	**3D**
Dice	0	0.9111	0.8898	0.8541	0.8179	0.7770

Case 2:

Slice	14	15	16	17	18	19	**3D**
Dice	0.7418	0.7547	0.8460	0.8178	0.8380	0.7937	0.8035

Case 4:

Slice	15	16	17	18	19	20	21	22	**3D**
Dice	0.8507	0.9196	0.9503	0.9518	0.9492	0.9423	0.9372	0.9052	0.9391

Case 5:

Slice	22	23	24	25	26	27	**3D**
Dice	0.9240	0.8980	0.8966	0.9044	0.8868	0.8626	0.9013

Case 6:

Slice	12	13	14	15	16	**3D**
Dice	0.8028	0.8675	0.8828	0.8623	0.8567	0.8574

Case 8:

Slice	11	12	13	14	15	**3D**
Dice	0.8887	0.9050	0.9109	0	0	0.7593

Case 9:

Slice	7	8	9	10	11	12	13	14	**3D**
Dice	0.9074	0.9470	0.9554	0.9541	0.9467	0.9275	0.9306	0.9309	0.9420

Table 6.10: Dice coefficients computed for comparison of the clinical segmentation and the automatic segmentation. Dice coefficents are calculated in 2D (on each slice) and in 3D (whole patient comparison) for each of the patients. The morphological operations have been performed in 3D. (Table is continued in Table 6.11.)

Case 10:

Slice	11	12	13	14	**3D**
Dice	0	0	0	0	0

Case 11:

Slice	25	26	27	**3D**
Dice	0	0	0	0

Case 12:

Slice	7	8	**3D**
Dice	0	0	0

Case 14:

Slice	7	8	9	10	11	12	13	**3D**
Dice	0.8911	0.9134	0.8995	0.9151	0.9098	0.9006	0.7935	0.8976

Case 15:

Slice	17	18	19	**3D**
Dice	0.7132	0.8634	0	0.6101

Case 16:

Slice	11	12	13	14	**3D**
Dice	0.8993	0.8984	0.8927	0.7801	0.8872

Case 18:

Slice	18	19	20	21	22	23	24	**3D**
Dice	0.9100	0.8938	0.8558	0.8895	0.8526	0.8540	0.8356	0.8737

Table 6.11: Dice coefficients computed for comparison of the clinical segmentation and the automatic segmentation. Dice coefficents are calculated in 2D (on each slice) and in 3D (whole patient comparison) for each of the patients. The morphological operations have been performed in 3D. (Continued from Table 6.10).

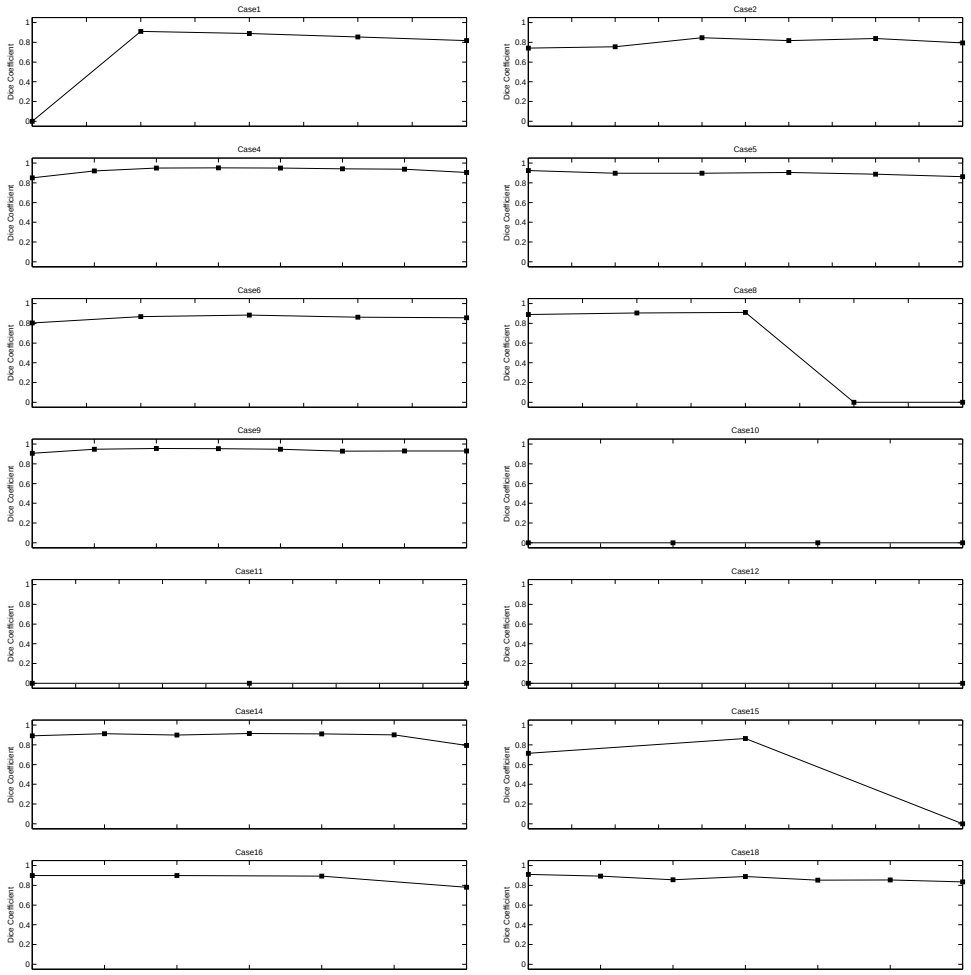


Figure 6.15: Dice coefficients computed for comparison of the clinical segmentation and the automatic segmentation. The morphological operations have been performed in 3D. Dice coefficents are calculated on each slice for all of the patients.

6.7 Discussion

This chapter has sought to identify an approach that is capable of automatically identifying the focal lesion disease within a prostate on MR data. To this end, texture features, machine learning and morphological processing have been combined to create a pipeline which may be able to fit into a clinical work flow and assist clinicians in the identification of focal disease. The identification of a focal lesion can be a difficult and time consuming process, so any method that can ease this burden may open up avenues for the improved treatment of patients with prostate cancer.

This work presents a method that is capable of showing significant agreement with contours outlined by an expert clinician. 11 of the 14 cases tested had a final Dice score greater than 0.65 and 8 of the 14 cases had a score greater than 0.9, indicating very good agreement with the

clinical segmentations. However 3 of the cases suffered catastrophic failure and the method was unable to successfully identify the diseased region within the prostate. The author believes that this is a very promising first step towards an automated approach for automatically identifying prostate focal disease.

The rest of this section will examine some of the areas were future work should focus and outline a potential way to integrate this approach with a modern clinical work flow.

6.7.1 Avenues for Further Investigation

The first issue with this data is that the reference is a contour outlined by a single expert clinician. This brings with it several key issues - there is great inter- and intra-clinician variability when identifying regions on clinical images. This is likely further compounded by the difficulty in accurately identify the focal disease. That is to say, the results presented here can only ever be as good as the reference data. Future work should use a larger corpus of data, ideally with segmentations by multiple clinicians and/or histological data (though registering histological information to the imaging frame of reference comes with its own challenges).

As explored above, some features (and the directions in which they are calculated) make a much larger contribution to the classification performance than others. The relevance of these important features should be investigated to determine what (if any) biological significance they may have. Removing the unimportant features can also provide an overall speed increase. It will take less time to calculate fewer texture features and the AdaBoost model will also require less time for training and testing.

The size of the subimages used to calculate the texture features does not seem to play much of a role in the classification performance. Further work should be carried out to stress test this conclusion - smaller and larger volume should be used until the classification performance breaks down. Additionally, the combination of features from different scales should be implemented. It should be possible to combine features calculated from different subimage sizes and determine if this provides a meaningful increase in performance.

In three of the cases the cancer makes up a large proportion of the prostate (Case 4: 69%, Case 5: 44% and Case 9: 29%). The validity of calling these regions “focal lesions” is questionable and perhaps they should be removed from the data set and treated separately. Additionally, the stage of the disease for each of these patients should be considered as it is likely that these patients’ MR scans were acquired at a much later stage of the disease’s development.

6.7.2 Clinical Work Flow

One key aspect of any research into improving radiotherapy with image analysis is the feasibility of incorporating the developed approach into a clinical work flow. The method presented in this chapter could easily fit into the clinical work flow after an initial training step with a large amount of labelled prostate data. An example integration is outlined below:

1. A diagnostic MR scan is acquired and the data is added to the hospital's PACS system.
2. The PACS system then calculates the required texture features for each pixel in an off line stage. This is feasible because there is usually a period of days or at least several hours before a clinician interrogates a scan. This step may take several hours depending on the features calculated and the size of the image.
3. The texture features for each pixel are passed through the precomputed classifier to predict the class label for each pixel - "healthy" or "focal disease". This is near instantaneous.
4. Once a clinician views the patient's MR scan and outlines the prostate volume, they request an automated segmentation of the focal disease. At this point, the predicted labels in the prostate (and optionally within a small radius of the prostate - in case the focal disease is protruding from the prostate volume) are extracted and the morphological processing steps are applied on the fly. This step should also be near instantaneous.
5. The clinician may now use the automatic segmentation as a starting point for identifying the focal disease.

6.7.3 Integration with CT

As already mentioned, in order to make use of the automatically segmented focal disease, the contour would need to be transferred to a CT planning scan. Recent work [41] by the research group I am a member of has demonstrated that this may be possible using registration techniques. Further refinements could include further evolving the registered contour on the CT scan by using texture features and level sets, for example. However the lower contrast on CT images makes the verfication of this approach extremely difficult and proper validation may require using gold standard histological information to determine the presence of focal disease.

6.7.4 Summary

This chapter has demonstrated that it is possible, with even a small cohort of patients, to develop a model that is capable of automatically identifying the prostate focal lesion in a number of cases. Further work is required in this area to study the importance of the features and to investigate the underlying reasons for the different performance of the features. Additionally, future work should use a much larger cohort of patients that would allow for the training of models based on disease stage and demographic information. Such a data set would allow

for a parametric study of the different texture features, different learning models and different morphological operations without the risk of over fitting to a small number of cases.

Chapter 7

Conclusions

7.1 Introduction

The overall aim of this thesis was to apply three dimensional texture analysis and machine learning methods to the field of medical imaging to improve radiotherapy. To this end, three dimensional implementations of texture features were implemented and applied to the prediction of radiation induced pneumonitis from CT planning scans (Chapter 5) and to identifying cancerous regions within the prostate from MRI scans (Chapter 6). Both of these applications may have the potential to improve the treatment of patients with lung or prostate cancer.

The aim of the radiation induced pneumonitis work was to determine if it was possible to use imaging data, acquired in the normal course of treatment, to predict if a patient was likely to develop pneumonitis post radiotherapy. Prediction of the occurrence of pneumonitis may allow for alternative treatment in an attempt to prevent the disease. Alternatively, high risk patients could be identified for a different follow up schedule. With the increasing use of SABR treatments for lung cancer which result in a higher dose delivered to the GTV, but also a significant dose wash from the VMAT such a predictive scheme may be timely. As explained in Chapter 2, it is thought that an increase in the low level dose wash (a little dose to a large region) may be correlated with radiation induced pneumonitis. If this proves to be the case, this would further underline the need for predicting a patient's risk of pneumonitis occurring as a larger number of lung cancer patients may be treated using radiotherapy plans likely to trigger the development of pneumonitis.

Identifying the focal disease within the prostate can be a time consuming and difficult undertaking for a clinician. The aim of this work was to define an automatic segmentation approach that could allow for easier and faster identification of the cancer within the prostate volume. Demonstration of the efficacy of such an approach could enable an easier path towards "focal boosting" by speeding up one of the more manual steps that would be required for such a treatment. This focal boosting could simultaneously facilitate an increase in the dose delivered to the cancer and a lower dose delivered to surrounding healthy tissues. Much work would be required to validate such an approach: the work in this thesis has focussed solely on the automatic identification of the disease on MRI scans.

The remainder of this chapter will cover how these aims have been met and lay out a potential pathway for a clinical deployment of the methods developed in this work. Finally, potential future work will be laid out.

7.2 Review of Contributions and Future Work

Chapter 2 laid out the relevant clinical background required as a foundation for understanding the applications presented in later Chapters. There was a particular focus on CT and MR imaging and on prostate cancer and lung cancer.

Chapter 3 introduced the concept of texture in an image before explaining texture's relation to human vision. A taxonomy of texture was then presented to place the methods used in this work in context. The remainder of the chapter described three dimensional texture and the specific methods used in the rest of this thesis.

Chapter 4 covers the machine learning techniques used in Chapters 5 and 6. There is an introduction to machine learning before the supervised learning methods used are covered. The chapter also includes the methods used for feature preprocessing and model evaluation used in this thesis.

Chapter 5 demonstrates that it is possible to predict if a patient with lung cancer will develop pneumonitis after radiotherapy by combining a small amount of clinical data with texture features calculated from a CT planning scan. An Area Under ROC of 0.873 was achieved using an SVM and further results showed the efficacy of multiple Decision Tree models using different regions of the lung volume. Furthermore, the use of multiple learners and multiple regions of the lung volume demonstrated the stability of the approach and reinforces the findings. These results seem to indicate that there is an inherent difference in the lung tissue of a patient at risk of pneumonitis and the lung tissue of a patient that does not develop pneumonitis post radiotherapy. This warrants further investigation with a much larger data set to test this hypothesis and the wider applicability of the approach. Additionally, future work should also focus on identifying the features that contribute most during the learning phase - if these features can be better understood, they may reveal something of the nature of the underlying tissue differences. That is, if it can be determined which features have the most discriminatory power, it may be possible to discern the underlying structures these features describe. Such information could lead to clues as to the biological nature of the differences of the lung tissues in the symptomatic and asymptomatic populations.

In Chapter 6 an automated pipeline for the identification of focal disease in patients with prostate cancer was developed. This pipeline made use of texture analysis to derive a number of features for each pixel in the prostate volume. These features were then used as inputs into

a classification scheme before being mapped back onto the MRI frame of reference and post-processed with morphological operations to produce a contour of the focal disease within the prostate. The results in Chapter 6 demonstrate the viability of this approach and display the need for more work in this area. A mean Area Under ROC of 0.713 was demonstrated on an aggregate data set (made up of the prostate volumes of all 14 subjects) using a 10-fold stratified cross validation strategy - this demonstrated that it is at least possible for a classifier to use texture features to discriminate between healthy prostate and focal disease tissues. A leave one case out strategy was then used to test the approach on a more clinically relevant footing. After the classification of the regions for each subject the class labels were mapped back onto the MRI space. Morphological operations were then used to produce a segmentation of the focal disease. The segmentations on all 14 subjects resulted in a mean DICE of 0.710. That is, there was a very significant overlap of the clinical contours and the predicted contours for the majority of the cases. This shows that texture descriptors are capable of discriminating between healthy and cancer tissues within the prostate. The methods outlined could easily fit into a clinical work flow: the texture analysis and prediction could be carried out off line on a hospital's PACS system so that the identification of the focal disease could be made in pseudo real time when a clinician is examining the MRI. If this approach could be incorporated with a method that allows for the transfer of the predicated focal disease contours to planning CT scans, this work could play an important role in the boosting of focal lesions.

7.3 Clinical Pathways

Any new technology or innovation must have a clear pathway to clinical use. This pathway must involve extension auditing and proof that there is a real benefit to the patient or the clinician. Two such pathways are laid out below, one for the radiation induced pneumonitis work and one for the prostate focal disease application.

A first step towards the use of the radiation induced pneumonitis prediction method would be to validate the approach on a much larger data set. This could be achieved using retrospective data. It would only be necessary to collect the CT scans, the clinical features and the associated outcome data for a large number of lung cancer patients. Standard methods for the evaluation of the performance of machine learning methods could then be applied to verify the efficacy of the techniques. Ideally, this data would be a multi-centre study to allow for the method to be tested with different CT scanners and to ensure there are multiple clinicians involved in the determination of the gold standard class labels - the diagnosis of pneumonitis. If this stage were successful, the next step would be to undertake a prospective clinical trial wherein a protocol would be devised for the treatment of patients determined to be at a higher risk of pneumonitis. The trial would then seek to determine if altering treatment based upon the prediction of pneumonitis improves patient outcome.

A similar approach would be required for the prostate work. An initial clinical audit could be performed where the approach is implemented as part of a hospital's PACS system and contours automatically created. These contours would then be evaluated by clinicians to assess the viability of the approach. An alternative avenue would be to collect retrospective data and have a clinician (or multiple clinicians) outline the focal disease to allow testing on a much larger data set. After this initial validation of the approach, a next step could involve a clinical trial based around prostatectomy - if the MRI data from men scheduled for prostatectomy was collected this would provide histological ground truth for the location of the focal disease. This ground truth could be used to further tune and test the classification pipeline. In order to use this approach for a focal boosting treatment, the contours would need to be mapped to the planning CT scan and suitable plans developed. This would likely need to be a clinical trial to demonstrate that the boosted plans provide an improved outcome for the patient. If the focal boosting was not viable, the automatic identification of the disease may still prove useful to clinicians during diagnosis and if proved on a large cohort, may provide an automatic, non-invasive indication of the extent of the cancer in the prostate.

7.4 Summary

In summary, this thesis has combined texture analysis of medical images with machine learning to present two applications to radiotherapy that are clinically relevant.

Both applications exploit the information contained in the (ever increasing) image data available in modern radiotherapy departments. It is the author's belief that in the coming years and decades there will be manifold applications of this kind due to the wide availability of imaging data, computing power and the ever growing interest in machine learning and its applications.

The prediction of radiation induced pneumonitis and the identification of prostate focal disease are both possible using texture analysis and it is this author's belief that both avenues display promise and warrant further investigation and evaluation.

Appendix A

Morphological Operations

Mathematical morphology is a field that considers the analysis and processing of geometrical structures. In this section the more narrow field of morphological operations on binary images will be looked at, though it is trivial to expand this to grey level images (and more generally to complete lattices).

Erosion and dilation are the cornerstone of processing an image using morphology and the combination of these operations allows for more advanced operations.

Structuring element: A structuring element is a small reference shape which is applied to the image under consideration. In effect, any details smaller than the structuring element are removed from the image. Typical structuring elements are disks of various sizes or lines of various orientations. A particular structuring element must be chosen for the task at hand.

Dilation: is the process of overlaying the structuring element at every object point in the original image and filling in all empty points covered by the structuring element. This process results in the filling in of gaps and the enlargement of the objects in the original image. A large square being dilated by a disc structuring element of radius, a, would result in a new image of a square with rounded corners, the rounded corners also having a radius of a.

Erosion: is the process of overlaying the structuring element at every object point in the original image and removing the object points at which the structuring element cannot fit within the image object. This process results in the removal of pixels and a reduction in the size of the objects in the image. If we again consider a square of side, b, and a disc structuring element of radius a, the result of erosion would be a new square of length $(b-a)$.

Erosion removes details of an image smaller than the structuring element and dilation fills in details that are smaller than the structuring element. By combining these two operations we can keep details at one scale while removing them at another.

Opening: is when a dilation is applied after an erosion. This has the effect of removing the "hairs" of an image - sharp, distinct structures are removed and smoothed.

Closing: is when an erosion is applied following a dilation. This has the effect of closing up narrow holes in an image without increasing the size of the object (assuming the same structuring element is used for both operations).

More complex operations can be constructed by using different structuring elements during each phase of the opening or closing, by chaining multiple operations together or by addition and subtraction of images after the application of operations. One example of such an operation is the top-hat transform in which the opening of an image is subtracted from the original image, this results in structures smaller than the structuring element remaining and can be used to extract fine details in an image.

Appendix B

High Performance Computing Considerations

This appendix covers two HPC use cases extracted from the work presented in the main body of this thesis. Section B.1 describes the use of a small cluster of computers in order to speed up the calculation of texture features from the MRI data used in Chapter 6. Section B.2 lays out an additional piece of work that was carried out in addition to the cross validation described in Section 6.4.3 in order to complete the grid search over a range of model paramters.

B.1 Texture Feature Calculation

The calculation of texture features (described in Section 6.4.1) on a number of subregions of an image is an "embarrassingly parallel" problem: the texture from each subregion may be calculated completely independently of its neighbours. This allows for the computation to be split into discrete independent chunks to be run concurrently. In the ideal case, a problem such as this could be speed up by a factor equal to the number of computing resources used to process the individual calculations. However, in practice, there are other overheads such as IO that result in a smaller increase in computation speed.

A 32 node MATLAB cluster at the Western General Hospital was used calculate the texture features at each point in the prostate for each subject. This allowed for a much faster computation than would be possible on a single machine.

The computing pipeline was constructed as follows. Note that a master node was used to orchestrate the process and submit jobs to the cluster at which point the cluster nodes performed the following tasks in parallel:

- The imaging and contour data for each subject was loaded from a common data store.
- A series of subregions of varying size were extracted from each subject and saved to the common data store alongside the positional information required to reconstruct the images.

- Each set of subregions were loaded from the data store and texture features calculated. The resulting features were then saved to the common data store.

Finally, the master node gathered the texture features and assembled a feature matrix, the class labels and the positional information required for reconstruction. This data was then used for classification as described in Chapter 6.

B.2 Cross Validation Grid Search

As explained in Section 6.4.3, Table 6.3 is incomplete due to extremely long run times on the hardware resources that were available at the time. This section presents the result of calculating the missing values in Table 6.3 using a combination of a parallelised method and more powerful hardware.

Originally the computation of the results in Table 6.3 were abandoned if the run time exceeded 48 hours. Each row in the table represents a 5-fold cross validation of an AdaBoost model for different model parameters (number of estimators and the maximum tree depth). Due to the iterative nature of the AdaBoost algorithm, it is not an inherently parallel problem, like, for example, a Random Forest classifier. However, each fold in the cross-validation can be computed independently (ensuring the data chosen for each fold is not the same on independent nodes). As a result, there are 25 distinct units of work (five rows multiplied by 5 folds).

The original script to generate the data for this table was adapted to make use of the Python multiprocessing module which allowed for a simple parallelisation this problem on a multi-core machine. A 20 core (64Gb RAM) machine was then rented from a cloud provider [i] to run the modified script. This is a cheap and quick way to parallelise an operation. Each core of the machine was then able to run a cross validation fold.

The modified script took approximately 9 hours to complete. The longest fold took approximately 4 hours, but since there were more jobs than cores, the total time was greater than the longest individual time. The total cost of this operation was approximately $9.

In principle, each fold could be calculated on a different machine, each of which could be less powerful. This would result in an increased orchestration overhead (which, would be easily managed) and a lower cost as the overall time and cost per machine would decrease. If a separate instance was used for each fold, then the overall cost would have been approximately $4.

The updated results are presented in Table B.1. The classification performances for the rows are fall short of the existing best classifier. This is most likely because the larger tree depths and the large number of learners causes the model to overfit the data at the training stage.

---

i. Digital Ocean was chosen for ease of use, per hour billing and low cost. At the time of writing the rented machine cost approximately $1 per hour or $640 per month.

Number of Estimators	Maximum Depth	Sensitivity	Specificity	Area Under ROC
10	1	0.6052	0.5799	0.6302
10	10	0.5422	0.6755	0.6501
10	50	0.5150	0.7014	0.6506
10	100	0.5150	0.6999	0.6496
100	1	0.6014	0.6412	0.6604
100	10	0.5181	0.7152	0.6612
100	50	0.5259	0.7497	**0.69105**
100	100	0.5278	0.7472	0.6884
500	1	0.5916	0.6566	0.6652
500	10	0.5265	0.7478	0.6895
500	50	0.5548	0.7282	0.6790
500	100	0.5534	0.7249	0.6793
1000	1	0.5831	0.6607	0.6619
1000	10	0.5544	0.6959	0.6674
1000	50	0.5561	0.7293	0.6840
1000	100	0.5540	0.7296	0.6827

Table B.1: Performance metrics from the grid search over the number of estimators and the maximum depth parameters in the AdaBoost model.

B.3 Conclusions

Techniques from the field of HPC can be very useful in early stages of research as they may allow for increased development, testing and iteration of novel techniques and enabling more data to be processed at an earlier stage of the research. However a there is a large amount of specialist knowledge required to fully exploit HPC architectures and this is beyond the scope of this current thesis. For example, it may be possible to port the texture calculations to a GPU and greatly reduce the execution time using only a single machine.

Furthermore, this thesis has not explored this topic in depth as the author believes that there is great value in developing techniques that can be run on today's commodity hardware as it reduces at least one barrier to future adoption of new methods and techniques. For example, it is easier for a hospital to invest in a technology if it only involves a software license than it would be for a hospital to purchase a large scale computing infrastructure or incur a possibly large monthly bill from a cloud computing provider.

Appendix C

Publications

1. Identifying the Dominant Prostate Cancer Focal Lesion using Image Analysis and Planning of a Simultaneous Integrated SABR Boost. Feng Y, Welsh D, McDonald K, Carruthers L, Cheng K, Montgomery D, Lawrence J, Argyle DJ, McLaughlin S, McLaren DB, Nailon WH,
   Acta Oncologica 54(9), 1543-1550 (Oct 2015).
2. Identifying Changes in The Gross Tumour Volume after Radiotherapy by Image Analysis. Feng Y, Cheng K, Montgomery D, Lawrence J, Forrest L, McLaren DB, McLaughlin S, Argyle DJ, Nailon WH,
   European Society of Radiotherapy and Oncology (ESTRO 2015).
3. Identifying Focal Prostate Cancer by Image Analysis. Nailon WH, Feng Y, Welsh D, Cheng K, Montgomery D, Lawrence J, McLaughlin S, Argyle JD, McLaren DB,
   Bio-guided Adaptive Radiotherapy (BiGART 2015).
4. Improving Radiotherapy by Informatics and Image Analysis. Nailon WH, Welsh D, Kehce T, Erridge S, Price A, Campbell S, McLaren D, Montgomery D, Feng Y, Cheng K, McLaughlin S, Lawrence J, Argyle D.
   Presented at the NHS Lothian R&D Conference (2015).
5. Image Analysis for Improving Radiotherapy in Prostate Cancer: Topical Review. Nailon WH, Feng Y, Welsh D, Cheng K, Montgomery D, Liao H, Lawrence J, McLaughlin S, Argyle JD, McLaren DB.
   Proposal submitted to Nature Oncology Reviews (2015).
6. Identifying Radiotherapy Target Volumes in Brain Cancer by Image Analysis. Cheng K, Montgomery D, Feng Y, Steel R, Liao H, McLaren DB, Erridge SC, McLaughlin S, Nailon WH.
   Healthcare Technology Letters 2(5), 123-218 (2015).
7. Image Registration in Veterinary Radiation Oncology: Indications, Implications and Future Advances. Feng Y, Lawrence J, Cheng K, Montgomery D, Forrest L, McLaren DB...
   Submitted to Veterinary Radiology & Ultrasound (2014).
8. Active Shape Models for Prostate Cancer Planning with Optimized Features. Cheng K, Feng Y, Montgomery D, Steel R, Liao L, McLaren DB...

The International Society for Optics and Photonics (2014).

9. Identifying Changes in The Gross Tumour Volume after Radiotherapy by Image Analysis. Feng Y, Cheng K, Montgomery D, Lawrence J, Forrest L, McLaren DB...
   ESTRO (2014).
10. Predicting the Occurrence of Radiation Induced Pneumonitis by Texture Analysis of CT Images from Lung Cancer Patients. Montgomery D, Cheng K, Feng Y, McLaren DB, Erridge SC, McLaughlin S...
    The Fifth International Work Shop on Pulmonary Image Analysis at the 16th International Conference on Medical Image Computing and Computer Assisted Intervention (2013).

Other Publications

1. Data clustering methods for the determination of cerebral autoregulation functionality. Montgomery D, Addison PS, Borg U.
   Journal Clinical Monitoring and Computing (2015).

Bibliography

[1] Abo-Madyan, Y., Aziz, M.H., Aly, M.M., Schneider, F., Sperk, E., Clausen, S., Giordano, F.A., Herskind, C., Steil, V., Wenz, F., Glatting, G.: Second cancer risk after 3d-crt, imrt and vmat for breast cancer. Radiotherapy and Oncology 110(3), 471–476 (Mar 2014), http://dx.doi.org/10.1016/j.radonc.2013.12.002

[2] Şahan, S., Polat, K., Kodaz, H., Güneş, S.: A new hybrid method based on fuzzy-artificial immune system and -nn algorithm for breast cancer diagnosis. Computers in Biology and Medicine 37(3), 415 – 423 (2007), http://www.sciencedirect.com/science/article/pii/S001048250600076X

[3] Ahmad, W., Fauzi, M.: Comparison of different feature extraction techniques in content-based image retrieval for CT brain images. Multimedia Signal Processing (Jan 2008), http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4665130

[4] Aird, E., of Radiology, B.I., of Physics, I., in Medicine, E., Biology: Central Axis Depth Dose Data for Use in Radiotherapy, 1996: A Survey of Depth Doses and Related Data Measured in Water Or Equivalent Media. BJR: Supplement, British Institute of Radiology (1996)

[5] Bağcı, U., Bray, M., Caban, J., Yao, J., Mollura, D.J.: Computer-assisted detection of infectious lung diseases: A review. Computerized Medical Imaging and Graphics 36(1), 72–84 (2012)

[6] Barbu, T.: Content-based image retrieval using gabor filtering. In: Database and Expert Systems Application, 2009. DEXA '09. 20th International Workshop on. pp. 236–240 (Aug 2009)

[7] Bauman, G., Haider, M., der Heide, U.A.V., M'enard, C.: Boosting imaging defined dominant prostatic tumors: A systematic review. Radiotherapy and Oncology 107(3), 274–281 (Jun 2013), http://dx.doi.org/10.1016/j.radonc.2013.04.027

[8] Bennett, K., Demiriz, A., et al.: Semi-supervised support vector machines. Advances in Neural Information processing systems pp. 368–374 (1999)

[9] Bigun, J.: Speed, frequency, and orientation tuned 3-d gabor filter banks and their design. In: Pattern Recognition, 1994. Vol. 3 - Conference C: Signal Processing, Proceedings of the 12th IAPR International Conference on. pp. 184–187 vol.3 (Oct 1994)

[10] Bitar, R., Leung, G., Perng, R., Tadros, S., Moody, A.R., Sarrazin, J., McGregor, C., Christakis, M., Symons, S., Nelson, A., Roberts, T.P.: Mr pulse sequences: What every radiologist wants to know but is afraid to ask. RadioGraphics 26(2), 513–537 (2006), http://dx.doi.org/10.1148/rg.262055063, pMID: 16549614

[11] Blanzieri, E., Bryl, A.: A survey of learning-based techniques of email spam filtering.

Artificial Intelligence Review 29(1), 63–92 (2008), http://dx.doi.org/10.1007/s10462-009-9109-6

[12] den Boer, J.A., Vlaardingerbroek, M.T.: Magnetic Resonance Imaging. Theory and Practice. Springer (2003)

[13] Bomford, C.K., Miller, J., Kunkler, H., Sherriff, I., Bomford, S., IH Kunkler, S., et al.: Walter and Miller's Textbook of radiotherapy: radiation physics, therapy, and oncology. No. Sirsi) i9780443028731 (1993)

[14] Bongers, E.M., Botticella, A., Palma, D.A., Haasbeek, C.J., Warner, A., Verbakel, W.F., Slotman, B., Ricardi, U., Senan, S.: Predictive parameters of symptomatic radiation pneumonitis following stereotactic or hypofractionated radiotherapy delivered using volumetric modulated arcs. Radiotherapy and Oncology 109(1), 95–99 (2013)

[15] Breiman, L.: Bagging predictors. Mach. Learn. 24(2), 123–140 (Aug 1996), http://dx.doi.org/10.1023/A:1018054314350

[16] Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001), http://dx.doi.org/10.1023/A%3A1010933404324

[17] Burner, A., Donner, R., Mayerhoefer, M., Holzer, M., Kainberger, F., Langs, G.: Texture bags: anomaly retrieval in medical images based on local 3d-texture similarity. In: Medical Content-Based Retrieval for Clinical Decision Support, pp. 116–127. Springer (2012)

[18] Burnet, N.G.: Defining the tumour and target volumes for radiotherapy. Cancer Imaging 4(2), 153–161 (2004), http://dx.doi.org/10.1102/1470-7330.2004.0054

[19] Byrne, D., O'Halloran, M., Jones, E., Glavin, M.: Support vector machine-based ultrawideband breast cancer detection system. Journal of Electromagnetic Waves and Applications 25(13), 1807–1816 (2011), http://dx.doi.org/10.1163/156939311797454015

[20] Campbell, F.W., Robson, J.G.: Application of fourier analysis to the visibility of gratings. The Journal of Physiology 197(3), 551–566 (1968), http://jp.physoc.org/content/197/3/551.abstract

[21] Cao, Z., Liu, X., Peng, B., Moon, Y.S.: DSA image registration based on multiscale gabor filters and mutual information. In: Information Acquisition, 2005 IEEE International Conference on. pp. 6 pp.– (June 2005)

[22] Chen, S., Zhou, S., Yin, F.F., Marks, L.B., Das, S.K.: Investigation of the support vector machine algorithm to predict lung radiation-induced pneumonitis. Medical physics 34, 3808 (2007)

[23] Chen, S., Zhou, S., Zhang, J., Yin, F.F., Marks, L.B., Das, S.K.: A neural network model to predict lung radiation-induced pneumonitis. Medical physics 34, 3420 (2007)

[24] Chicklore, S., Goh, V., Siddique, M., Roy, A., Marsden, P.K., Cook, G.J.: Quantifying tumour heterogeneity in 18f-fdg pet/ct imaging by texture analysis. European journal of nuclear medicine and molecular imaging 40(1), 133–140 (2013)

[25] Chu, A., Sehgal, C., Greenleaf, J.: Use of gray value distribution of run lengths for texture analysis. Pattern Recognition Letters 11(6), 415–419 (Jun 1990), http://dx.doi.org/10.1016/0167-8655(90)90112-F

[26] Claude, L., Pérol, D., Ginestet, C., Falchero, L., Arpin, D., Vincent, M., Martel, I., Hominal, S., Cordier, J.F., Carrie, C.: A prospective study on radiation pneumonitis following conformal radiation therapy in non-small-cell lung cancer: clinical and dosimetric factors analysis. Radiotherapy and oncology 71(2), 175–181 (2004)

[27] Clausi, D., Deng, H.: Design-based texture feature fusion using gabor filters and co-occurrence probabilities. Image Processing, IEEE Transactions on 14(7), 925–936 (July 2005)

[28] Clausi, D.A.: Comparison and fusion of co-occurrence, gabor and mrf texture features for classification of sar sea-ice imagery. Atmosphere-Ocean 39(3), 183–194 (2001), http://dx.doi.org/10.1080/07055900.2001.9649675

[29] Cortes, C., Vapnik, V.: Support-vector networks. Machine learning 20(3), 273–297 (1995)

[30] CRUK: CRUK cancer statistics. cruk.org/health-professional/cancer-statistics, accessed: 2015-09-22

[31] CRUK: CRUK lung cancer statistics. cruk.org/cancer-info/cancerstats/types/lung/, accessed: 2014-05-29

[32] CRUK: CRUK prostate cancer statistics. cruk.org/cancer-info/cancerstats/types/prostate/, accessed: 2014-05-29

[33] Dang, J., Li, G., Ma, L., Diao, R., Zang, S., Han, C., Zhang, S., Yao, L.: Predictors of grade$\geq$ 2 and grade$\geq$ 3 radiation pneumonitis in patients with locally advanced non-small cell lung cancer treated with three-dimensional conformal radiotherapy. Acta Oncologica 52(6), 1175–1180 (2013)

[34] Dasarathy, B.V., Holder, E.B.: Image characterizations based on joint gray level-run length distributions. Pattern Recognition Letters 12(8), 497–502 (Aug 1991), http://dx.doi.org/10.1016/0167-8655(91)80014-2

[35] Daugman, J.G.: Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. J. Opt. Soc. Am. A 2(7), 1160 (Jul 1985), http://dx.doi.org/10.1364/JOSAA.2.001160

[36] Depeursinge, A., Foncubierta-Rodriguez, A., Van De Ville, D., Müller, H.: Three-dimensional solid texture analysis in biomedical imaging: Review and opportunities. Medical image analysis 18(1), 176–196 (2014)

[37] Dettori, L., Bashir, A., Hasemann, J.: Texture classification of normal tissues in computed tomography using gabor filters. vol. 6512, pp. 65120Q–65120Q–10 (2007), http://dx.doi.org/10.1117/12.710316

[38] Emami, B., Lyman, J., Brown, A., Cola, L., Goitein, M., Munzenrider, J., Shank, B., Solin, L., Wesson, M.: Tolerance of normal tissue to therapeutic

irradiation. International Journal of Radiation Oncology*Biology*Physics 21(1), 109 – 122 (1991), http://www.sciencedirect.com/science/article/pii/036030169190171Y, three-Dimensional Photon Treatment Planning Report of the Collaborative Working Group on the Evaluation of Treatment Planning for External Photon Beam Radiotherapy

[39] Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. pp. 226–231. AAAI Press (1996)

[40] Fehr, J.: Rotational invariant uniform local binary patterns for full 3D volume texture analysis. In: Finnish signal processing symposium (FINSIG) (2007)

[41] Feng, Y., Welsh, D., McDonald, K., Carruthers, L., Cheng, K., Montgomery, D., Lawrence, J., Argyle, D., McLaughlin, S., McLaren, D., Nailon, W.: Identifying the dominant prostate cancer focal lesion using image analysis and planning of a simultaneous integrated sabr boost. Acta Oncologica (2015)

[42] Frankel, S., Smith, G.D., Donovan, J., Neal, D.: Screening for prostate cancer. The Lancet 361(9363), 1122 – 1128 (2003), http://www.sciencedirect.com/science/article/pii/S0140673603128905

[43] Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55(1), 119 – 139 (1997), http://www.sciencedirect.com/science/article/pii/S002200009791504X

[44] Freund, Y., Schapire, R.E., et al.: Experiments with a new boosting algorithm. In: ICML. vol. 96, pp. 148–156 (1996)

[45] Gabor, D.: Theory of communication. part 1: The analysis of information. Electrical Engineers-Part III: Radio and Communication Engineering, Journal of the Institution of 93(26), 429–441 (1946)

[46] Galloway, M.M.: Texture analysis using gray level run lengths. Computer graphics and image processing 4(2), 172–179 (1975)

[47] Good, D., Khan, A., Hammer, S., Scanlan, P., Shu, W., Phipps, S., Parson, S., Stewart, G., Reuben, R., McNeill, S.: Tissue quality assessment using a novel direct elasticity assessment device (the e-finger): a cadaveric study of prostatectomy dissection. PLoS One 9(11), e112872 (11 2014)

[48] Groenendaal, G., Borren, A., Moman, M.R., Monninkhof, E., van Diest, P.J., Philippens, M.E., van Vulpen, M., van der Heide, U.A.: Pathologic validation of a model based on diffusion-weighted imaging and dynamic contrast-enhanced magnetic resonance imaging for tumor delineation in the prostate peripheral zone. International Journal of Radiation Oncology*Biology*Physics 82(3), e537 – e544 (2012), http://www.sciencedirect.com/science/article/pii/S0360301611030914

[49] Gumus, E., Kilic, N., Sertbas, A., Ucan, O.N.: Evaluation of face recognition techniques

using pca, wavelets and svm. Expert Systems with Applications 37(9), 6404–6408 (Sep 2010), http://dx.doi.org/10.1016/j.eswa.2010.02.079

[50] Haralick, R.M., Shanmugam, K., Dinstein, I.H.: Textural features for image classification. Systems, Man and Cybernetics, IEEE Transactions on (6), 610–621 (1973)

[51] Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., Tibshirani, R.: The elements of statistical learning, vol. 2. Springer (2009)

[52] He, C., Zheng, Y., Ahalt, S.: Object tracking using the gabor wavelet transform and the golden section algorithm. Multimedia, IEEE Transactions on 4(4), 528–538 (Dec 2002)

[53] Hill, R., Healy, B., Holloway, L., Kuncic, Z., Thwaites, D., Baldock, C.: Advances in kilovoltage x-ray beam dosimetry. Physics in Medicine and Biology 59(6), R183 (2014), http://stacks.iop.org/0031-9155/59/i=6/a=R183

[54] Holmström, B., Johansson, M., Bergh, A., Stenman, U.H., Hallmans, G., Stattin, P.: Prostate specific antigen for early detection of prostate cancer: longitudinal study. BMJ 339 (2009)

[55] Hope, A., Lindsay, P., Venugopal, A.: 281 increased acute symptoms of radiation pneumonitis with concurrent chemoradiotherapy vs. radiotherapy alone in a murine model of fractionated sub-total thoracic igrt. Radiotherapy and Oncology 102, S147 (2012)

[56] Hope, A.J., Lindsay, P.E., El Naqa, I., Alaly, J.R., Vicic, M., Bradley, J.D., Deasy, J.O.: Modeling radiation pneumonitis risk with clinical, dosimetric, and spatial parameters. International Journal of Radiation Oncology* Biology* Physics 65(1), 112–124 (2006)

[57] Jain, A., Farrokhnia, F.: Unsupervised texture segmentation using gabor filters. In: Systems, Man and Cybernetics, 1990. Conference Proceedings., IEEE International Conference on. pp. 14–19 (Nov 1990)

[58] Jemal, A., Bray, F., Center, M.M., Ferlay, J., Ward, E., Forman, D.: Global cancer statistics. CA: a cancer journal for clinicians 61(2), 69–90 (2011)

[59] Jin, H., Tucker, S.L., Liu, H.H., Wei, X., Yom, S.S., Wang, S., Komaki, R., Chen, Y., Martel, M.K., Mohan, R., et al.: Dose–volume thresholds and smoking status for the risk of treatment-related pneumonitis in inoperable non-small cell lung cancer treated with definitive radiotherapy. Radiotherapy and Oncology 91(3), 427–432 (2009)

[60] Julesz, B.: Visual pattern discrimination. Information Theory, IRE Transactions on 8(2), 84–92 (February 1962)

[61] Julesz, B., Gilbert, E.N., Shepp, L.A., Frisch, H.L.: Inability of humans to discriminate between visual textures that agree in second-order statistics – revisited. Perception 2(4), 391–405 (1973), http://dx.doi.org/10.1068/p020391

[62] Julesz, B.: Experiments in the visual perception of texture. Sci Am 232(4), 34–43 (Apr 1975), http://dx.doi.org/10.1038/scientificamerican0475-34

[63] Julesz, B.: A theory of preattentive texture discrimination based on first-order statistics

of textons. Biol. Cybern. 41(2), 131–138 (Aug 1981), http://dx.doi.org/10.1007/BF00335367

[64] Kak, A.C., Slaney, M.: Principles of computerized tomographic imaging. IEEE press (1988)

[65] Korfiatis, P.D., Karahaliou, A.N., Kazantzi, A.D., Kalogeropoulou, C., Costaridou, L.I.: Texture-based identification and characterization of interstitial pneumonia patterns in lung multidetector CT. Information Technology in Biomedicine, IEEE Transactions on 14(3), 675–680 (2010)

[66] Kumar, S., Moni, R., Rajeesh, J.: Liver tumor diagnosis by gray level and contourlet coefficients texture analysis. In: Computing, Electronics and Electrical Technologies (ICCEET), 2012 International Conference on. pp. 557–562. IEEE (2012)

[67] Lakshmi, D., Santhosham, R., Ranganathan, H.: Comparison of texture analysis in the differentiation of carcinoma from other lung abnormalities using low-dose CT images. In: Point-of-Care Healthcare Technologies (PHT), 2013 IEEE. pp. 271–274. IEEE (2013)

[68] Lee, C.C., Chen, S.H., Tsai, H.M., choo Chung, P., Chiang, Y.C.: Discrimination of liver diseases from ct images based on gabor filters. In: Computer-Based Medical Systems, 2006. CBMS 2006. 19th IEEE International Symposium on. pp. 203–206 (2006)

[69] Li, Y., Verma, R.: Multichannel image registration by feature-based information fusion. Medical Imaging, IEEE Transactions on 30(3), 707–720 (March 2011)

[70] Liu, C.L., Nakashima, K., Sako, H., Fujisawa, H.: Handwritten digit recognition: benchmarking of state-of-the-art techniques. Pattern Recognition 36(10), 2271–2285 (Oct 2003), http://dx.doi.org/10.1016/S0031-3203(03)00085-2

[71] Lloyd, S.: Least squares quantization in pcm. IEEE transactions on information theory 28(2), 129–137 (1982)

[72] Mamourian, A.C.: CT Imaging: Practical Physics, Artifacts, and Pitfalls. Oxford University Press (2013)

[73] Marčelja, S.: Mathematical description of the responses of simple cortical cells∗. J. Opt. Soc. Am. 70(11), 1297–1300 (Nov 1980), http://www.opticsinfobase.org/abstract.cfm?URI=josa-70-11-1297

[74] Menon, M., Shrivastava, A., Tewari, A., Sarle, R., Hemal, A., Peabody, J.O., Vallancien, G.: Laparoscopic and robot assisted radical prostatectomy: establishment of a structured program and preliminary analysis of outcomes. The Journal of urology 168(3), 945–949 (2002)

[75] Mitchell, T.M.: Machine Learning. McGraw-Hill, Inc., New York, NY, USA, 1 edn. (1997)

[76] Mitchell, T.M.: The discipline of machine learning. Carnegie Mellon University, School of Computer Science, Machine Learning Department (2006)

[77] Nava, R., Escalante-Ramírez, B., Cristóbal, G.: Texture image retrieval based on log-

gabor features. In: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, pp. 414–421. Springer Science + Business Media (2012), http://dx.doi.org/10.1007/978-3-642-33275-3_51

[78] Nava, R., Escalante-Ramírez, B., Cristóbal, G., Estépar, R.S.J.: Extended Gabor approach applied to classification of emphysematous patterns in computed tomography. Medical & Biological Engineering & Computing 52(4), 393–403 (Apr 2014), http://link.springer.com/10.1007/s11517-014-1139-9

[79] Ojala, T., Pietikainen, M., Harwood, D.: Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In: Pattern Recognition, 1994. Vol. 1 - Conference A: Computer Vision amp; Image Processing., Proceedings of the 12th IAPR International Conference on. vol. 1, pp. 582–585 vol.1 (Oct 1994)

[80] Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. Pattern Analysis and Machine Intelligence, IEEE Transactions on 24(7), 971–987 (2002)

[81] OpenStax: Anatomy & Physiology. OpenStax-CNX (2014), accessed: 2014-06-19

[82] Palma, D.A., Senan, S., Tsujino, K., Barriger, R.B., Rengan, R., Moreno, M., Bradley, J.D., Kim, T.H., Ramella, S., Marks, L.B., et al.: Predicting radiation pneumonitis after chemoradiation therapy for lung cancer: an international individual patient data meta-analysis. International Journal of Radiation Oncology* Biology* Physics 85(2), 444–450 (2013)

[83] Park, M., Jin, J., Wilson, L.: Fast content-based image retrieval using quasi-gabor filter and reduction of image feature dimension. In: Image Analysis and Interpretation, 2002. Proceedings. Fifth IEEE Southwest Symposium on. pp. 178–182 (2002)

[84] Parry, R.M., Jones, W., Stokes, T.H., Phan, J.H., Moffitt, R.A., Fang, H., Shi, L., Oberthuer, A., Fischer, M., Tong, W., Wang, M.D.: k-nearest neighbor models for microarray gene expression analysis and clinical outcome prediction. The Pharmacogenomics Journal 10(4), 292–309 (Aug 2010), http://dx.doi.org/10.1038/tpj.2010.56

[85] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)

[86] Peng, H., Long, F., Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. Pattern Analysis and Machine Intelligence, IEEE Transactions on 27(8), 1226–1238 (Aug 2005)

[87] Petrella, L., Gómez, W., Alvarenga, A.: Gabor filter for the segmentation of skin lesions from ultrasonographic images. INTERNATIONAL (Jan 2012), http://scitation.aip.org/content/aip/proceeding/aipcp/10.1063/1.3703201

[88] Petrou, M., Garcia Sevilla, P.: Introduction, pp. 1–10. John Wiley & Sons, Ltd (2006), http://dx.doi.org/10.1002/047003534X.ch1

[89] Prasanna, P., Tiwari, P., Madabhushi, A.: Co-occurrence of local anisotropic gradient orientations (collage): Distinguishing tumor confounders and molecular subtypes on mri. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014, pp. 73–80. Springer (2014)

[90] Pudil, P., Ferri, F., Novovicova, J., Kittler, J.: Floating search methods for feature selection with nonmonotonic criterion functions. In: Pattern Recognition, 1994. Vol. 2-Conference B: Computer Vision & Image Processing., Proceedings of the 12th IAPR International. Conference on. vol. 2, pp. 279–283. IEEE (1994)

[91] Rancati, T., Ceresoli, G.L., Gagliardi, G., Schipani, S., Cattaneo, G.M.: Factors predicting radiation pneumonitis in lung cancer patients: a retrospective study. Radiotherapy and oncology 67(3), 275–283 (2003)

[92] Rodrigues, G., Lock, M., D'Souza, D., Yu, E., Van Dyk, J.: Prediction of radiation pneumonitis by dose–volume histogram parameters in lung cancer—a systematic review. Radiotherapy and oncology 71(2), 127–138 (2004)

[93] Scalco, E., Fiorino, C., Cattaneo, G.M., Sanguineti, G., Rizzo, G.: Texture analysis for the assessment of structural changes in parotid glands induced by radiotherapy. Radiotherapy and Oncology 109(3), 384–387 (dec 2013), http://dx.doi.org/10.1016/j.radonc.2013.09.019

[94] Scarfe, W.C., Farman, A.G.: What is cone-beam ct and how does it work? Dental Clinics of North America 52(4), 707–730 (2008)

[95] Seeram, E.: Computed tomography: physical principles, clinical applications, and quality control. Elsevier Health Sciences (2015)

[96] Segawa, Y., Takigawa, N., Kataoka, M., Takata, I., Fujimoto, N., Ueoka, H.: Risk factors for development of radiation pneumonitis following radiation therapy with or without chemotherapy for lung cancer. International Journal of Radiation Oncology* Biology* Physics 39(1), 91–98 (1997)

[97] Shah, V., Turkbey, B., Mani, H., Pang, Y., Pohida, T., Merino, M.J., Pinto, P.A., Choyke, P.L., Bernardo, M.: Decision support system for localizing prostate cancer based on multiparametric magnetic resonance imaging. Medical Physics 39(7), 4093 (2012), http://dx.doi.org/10.1118/1.4722753

[98] Shen, L., Auer, D., Bai, L.: 3d gabor wavelets for evaluating medical image registration algorithms. In: Yang, G.Z., Jiang, T., Shen, D., Gu, L., Yang, J. (eds.) Medical Imaging and Augmented Reality, Lecture Notes in Computer Science, vol. 4091, pp. 261–268. Springer Berlin Heidelberg (2006), http://dx.doi.org/10.1007/11812715_33

[99] Shen, L., Jia, S.: Three-dimensional gabor wavelets for pixel-based hyperspectral imagery classification. Geoscience and Remote Sensing, IEEE Transactions on 49(12), 5039–5046 (Dec 2011)

[100] Sidky, E., Chartrand, R., Duchin, Y., Ullberg, C., Pan, X.: High resolution image reconstruction with constrained, total-variation minimization. In: Nuclear Science Symposium Conference Record (NSS/MIC), 2010 IEEE. pp. 2617–2620 (Oct 2010)

[101] Somol, P., Pudil, P., Novovicova, J., Paclik, P.: Adaptive floating search methods in feature selection. Pattern Recognition Letters 20(11–13), 1157–1163 (1999), http://www.sciencedirect.com/science/article/pii/S0167865599000835

[102] Sørensen, L., Shaker, S.B., De Bruijne, M.: Texture classification in lung ct using local binary patterns. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2008, pp. 934–941. Springer (2008)

[103] Sørensen, L., Shaker, S.B., De Bruijne, M.: Quantitative analysis of pulmonary emphysema using local binary patterns. Medical Imaging, IEEE Transactions on 29(2), 559–569 (2010)

[104] Sorzano, C., Vargas, J., Montano, A.P.: A survey of dimensionality reduction techniques. arXiv preprint arXiv:1403.2877 (2014)

[105] Sutton, R.S., Barto, A.G.: Reinforcement learning: An introduction, vol. 1. MIT press Cambridge (1998)

[106] Tan, X., Triggs, B.: Enhanced local texture feature sets for face recognition under difficult lighting conditions. IEEE Trans. on Image Process. 19(6), 1635–1650 (2010), http://dx.doi.org/10.1109/TIP.2010.2042645

[107] Tang, Y.: Deep learning using support vector machines. CoRR abs/1306.0239 (2013)

[108] Tarabalka, Y., Fauvel, M., Chanussot, J., Benediktsson, J.: Svm- and mrf-based method for accurate classification of hyperspectral images. Geoscience and Remote Sensing Letters, IEEE 7(4), 736–740 (Oct 2010)

[109] Tesař, L., Shimizu, A., Smutek, D., Kobatake, H., Nawano, S.: Medical image analysis of 3d ct images based on extension of haralick texture features. Computerized Medical Imaging and Graphics 32(6), 513–520 (Sep 2008), http://dx.doi.org/10.1016/j.compmedimag.2008.05.005

[110] Thibault, G., Fertil, B., Navarro, C., Pereira, S., Cau, P., Levy, N., Sequeira, J., Mari, J.L.: Texture indexes and gray level size zone matrix: application to cell nuclei classification. Pattern Recognition and Information Processing (PRIP), Minsk, Belarus pp. 140–145 (2009)

[111] Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58, 267–288 (1994)

[112] UK, C.R.: CRUK (2014), cruk.org/about-cancer/type/prostate-cancer/treatment/radiotherapy/internal-radiotherapy-for-prostate-cancer, accessed: 2014-06-19

[113] Wang, D., Shi, J., Liang, S., Lu, S., Qi, X., Wang, Q., Zheng, G., Wang, S., Zhang, K., Liu, H.: Dose–volume histogram parameters for predicting radiation pneumonitis using

receiver operating characteristic curve. Clinical and Translational Oncology 15(5), 364–369 (2013)

[114] Westbrook, C., Roth, C.K.: MRI in Practice. John Wiley & Sons (2011)

[115] Xu, Y., Zhu, J.Y., Chang, E.I., Lai, M., Tu, Z., et al.: Weakly supervised histopathology cancer image segmentation and classification. Medical Image Analysis 18(3), 591–604 (2014)

[116] Xu, Y., Sonka, M., McLennan, G., Guo, J., Hoffman, E.A.: MDCT-based 3-D texture classification of emphysema and early smoking related lung pathologies. Medical Imaging, IEEE Transactions on 25(4), 464–475 (2006)

[117] Xu, Z., Allen, W.M., Baucom, R.B., Poulose, B.K., Landman, B.A.: Texture analysis improves level set segmentation of the anterior abdominal wall. Medical Physics 40(12), 121901 (Jan 2013), http://scitation.aip.org/content/aapm/journal/medphys/40/12/10.1118/1.4828791

[118] Ylioinas, J., Hadid, A., Pietikäinen, M.: Combining contrast information and local binary patterns for gender classification. In: Heyden, A., Kahl, F. (eds.) Image Analysis, Lecture Notes in Computer Science, vol. 6688, pp. 676–686. Springer Berlin Heidelberg (2011), http://dx.doi.org/10.1007/978-3-642-21227-7_63

[119] Zavaletta, V.A., Bartholmai, B.J., Robb, R.A.: High resolution multidetector CT-aided tissue analysis and quantification of lung fibrosis. Academic radiology 14(7), 772–787 (2007)

[120] Zhang, D., Wong, A., Indrawan, M., Lu, G.: Content-based image retrieval using gabor texture features. In: IEEE Transactions PAMI. pp. 13–15 (2000)

[121] Zhang, X.J., Sun, J.G., Sun, J., Ming, H., Wang, X.X., Wu, L., Chen, Z.T.: Prediction of radiation pneumonitis in lung cancer patients: a systematic review. Journal of cancer research and clinical oncology 138(12), 2103–2116 (2012)