IMAGE PROCESSING AND MACHINE LEARNING APPLICATIONS IN LUNG CANCER TREATMENT

Matthew Gil

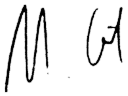
Submitted for the Degree of Doctor of Engineering

University of Strathclyde
Department of Electronic and Electrical Engineering

Academic Supervisors: Stephen Marshall and Paul Murray
Industrial Supervisors: Stephen Harrow

June 2024

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree. The copyright of this thesis belongs to the author under the terms of the United Kingdom Copyright Acts as qualified by University of Strathclyde Regulation 3.50. Due acknowledgement must always be made of the use of any material contained in, or derived from, this thesis.



Signed

14/06/2024

Date

Abstract

The continued advancement of image processing and machine learning techniques opens up the opportunity for their application in the medical setting. The aim of the work in this thesis was to apply these techniques from this broad field to lung cancer treatment with the aim of providing tools that can improve patient outcomes. The topics covered were; pulmonary and esophageal toxicity following radiotherapy, registration of PET/CT imaging to pathology and the automatic segmentation of tumour regions in gross pathology images.

For the prediction of pulmonary toxicity, predictive features were extracted from pre-treatment planning CT images using radiomic and deep learning based approaches. When combined with dose features, these models produced a large increase in predictive power compared to models using only dose and clinical features. For the ILD patients receiving SABR, predictive power was also shown on several metrics such as the FACT-L and EQ-5D-5L scales. For predicting esophageal toxicity, the data from the RTOG-0617 clinical trial was used. Here the focus was on improving predictions from the dose maps. It was found that using 3D-CNNs, regression based training, including additional toxicities and ensembling models improved model performance. Tests were also conducted to determine the robustness of boosted decision tree and artificial neural network based models for esophageal toxicity prediction by adding noise to the test data.

The PET/CT to pathology registration task followed on from a previous project that built the framework for registering CT to pathology but failed to include PET due to respiratory motion blurring. This was added to by including respiratory gating and the OncoFreeze algorithm in the workflow to reduce the effects of respiratory motion. A PET to pathology registration was evaluated using thresholding based registration of the PET image. Additionally, a deep learning based method for the automatic segmentation of gross pathology images was produced. This included training and testing various UNet and DeeplabV3+ models with both Dice and cross entropy based loss functions. The best performing model was an ensemble of several models with morphological post processing steps.

Acknowledgements

I would like to express my deepest gratitude to all those who have supported and contributed to the successful completion of this EngD thesis. To Stephen Marshall and Stephen Harrow for their support as the primary supervisors to this project and to Paul Murray and Craig Dick for their support as secondary supervisors. Thank you Gabriel Reines March for providing code and the IRAS documentation from their previous project. Thank you Bill Nailon for allowing me to get involved in the pneumonitis prediction work and all the support for all of the radiotherapy based work. Thanks also to Zhuolin Yang, Sarah Elliot and Karen Mactier. Thank you David Palma for including me in the ASPIRE-ILD follow-up study and for the support in achieving its outcomes. Thank you Mary Frances Dempsey for the support and advice for all the PET imaging involved in this thesis. Thanks also to Sandy Small, Ana Matos, Alastair Gemmall, Scott, Carole Maxfield and all of the staff at the West of Scotland PET Centre. Thank you Kai Rakovic for the support in the pathology processing for the PET-Path study. Thank you Rocco Bilancia for joining the PET-Path study which allowed the patient recruitment to progress. Thank you Rob Rulach for helping consenting the PET-Path patients. Thank you Jinchang Ren, Andrew Campbell and Efstathios Branikas for the technical advice. Thank you to Zander, David, Aoife and Luke from the HSI group.

Thank you to the Beatson Cancer Charity, EPSRC and the CDT in Applied Photonics for funding and facilitating this project.

Thank you to all my friends and family for the support. This thesis is dedicated to my parents, Oonagh and Bob for their continued support and love throughout my entire life.

Matthew Gil

July 2024

Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

18F-FDG Fluorine labelled fluorodeoxiglucose

3D-CRT Three-dimensional conformal radiation therapy

ADL Activities of daily life

AE Adverse event

ANN Artificial neural network

CASA Continuous bed motion automated spectral analysis

CE Cross entropy

CNN Convolutional neural network

CT Computed tomography

CTCAE Common criteria for adverse events

CTV Clinical target volume

DICOM Digital imaging and communications in medicine

DLCO Diffusing capacity of the lungs for carbon monoxide

DVH Dose volume histogram

EBRT External beam radiotherapy

EQ-5D-5L EuroQol 5-Dimension 5-Level

FACT Functional Assessment of Cancer Therapy

FVC Forced vital capacity

GLCM Grey level co-occurance matrix

GLRM Grey-level run length matrix

GLSZM Grey-level size zone matrix

GTV Gross tumour volume

IBSI Image biomarker standardisation initiative

ICP Iterative closest points

ILD Interstitial lung disease

IMRT Intensity modulated radiotherapy

IoU Intersection over union

IRAS Integrated Research Application System

Linac Linear accelerator

LKB Lyman Kutcher Burman

LOOCV Leave one out cross validation

MAE	Mean absolute error
MLP	Multi-layer perceptron
MRI	Magnetic resonance imaging
MSE	Mean square error
NGTDM	Neighbourhood grey tone difference matrix
NSCLC	Non-small cell lung cancer
NTCP	Normal Tissue Complication Probability
OaR	Organ at risk
PET	Positron emission tomography
PIS	Participant Information Sheet
PTV	Planning target volume
RE	Radiation esophagitis
REC	Research Ethics Committee
ReLU	Rectified linear unit
RoI	Region of interest
RP	Radiation pneumonitis
RT	Radiotherapy
SABR	Stereotactic ablative radiotherapy
SCLC	Small cell lung cancer
SMOTE	Synthetic minority oversampling technique
SNR	Signal to noise ratio
SUV	Standard uptake value
TN	True negative
TP	True positive
TRE	Target registration error
VMAT	Volume modulated arc therapy
WSI	Whole slide imaging

List of publications

Publications

Conference Paper

Matthew Gil, Craig Dick, Stephen Harrow, Paul Murray, Gabriel Reines March & Stephen Marshall , "A Deep Learning Based Approach to Semantic Segmentation of Lung Tumour Areas in Gross Pathology Images", *Medical Image Understanding and Analysis*, Aberdeen, UK, 19-21 July 2023
doi:10.1007/978-3-031-48593-0_2

Conference Paper

Matthew Gil, Stephen Marshall, Sorcha Campbell, Ai Wain Yong, John Murchison, Gillian Ritchie, Stephen Harrow, Paul Murray, William H Nailon, "Predicting Radiation Pneumonitis using Pre-Radiotherapy CT Scans by Radiomic and a Pre Trained CNN Based Feature Extraction", *International Conference on the Use of Computers in Radiotherapy*, Lyon, UK, 8-11 July 2024

Conference Paper

Matthew Gil, Stephen Marshall, Stephen Harrow, Paul Murray, William H Nailon, "Robust Models for Esopohageal Toxicity Prediction from Radiation Dose Maps", *International Conference on the Use of Computers in Radiotherapy*, Lyon, UK, 8-11 July 2024

Conference Oral

Matthew Gil, Stephen Harrow, William H Nailon, Stephen Marshall, Robert Doucet, Houda Bhaig, Jean-Pierre Bissonnette, Andrew Hope, Brock Debenham, Stewart Gaede, Andrew Warner, Chris Reyson, David Palma, "Predicting Pulmonary Toxicity in Lung Cancer Patients with Interstitial Lung Disease Receiving SABR Using

Machine Learning: A Secondary Analysis of ASPIRE-ILD", *International Conference on the Use of Computers in Radiotherapy*, Lyon, UK, 8-11 July 2024
doi:https://doi.org/10.1016/j.ijrobp.2024.07.2206

Conference Poster

Matthew Gil, Stephen Marshall, Sorcha Campbell, Ai Wain Yong, John Murchison, Gillian Ritchie, Stephen Harrow, Paul Murray, William H Nailon, "Prediction of Radiation Pneumonitis in NSCLC Patients Receiving External Beam Radiotherapy Using a Radiomic Analysis of Pre-Treatment CT Images", *CRUK-ARR Radiation Research Conference*, Glasgow, 4-6 June 2023

Conference Poster

Matthew Gil, Stephen Marshall, Sorcha Campbell, Ai Wain Yong, John Murchison, Gillian Ritchie, Stephen Harrow, Paul Murray, William H Nailon, "Prediction of Radiation Pneumonitis in NSCLC Patients Receiving External Beam Radiotherapy Using a Radiomic Analysis of Pre-Treatment CT Images", *Scottish Radiotherapy Research Forum*, Stirling, 9 November 2023

Chapter 1

Introduction

1.1 Motivations

Lung cancer is one of the most common forms of cancer in the UK with around 49,200 new lung cancer cases in the UK every year [1]. The treatment of lung cancer can be performed by a combination of; surgical removal of a tumour, radiotherapy and chemotherapy. The choice of treatment for any particular patient is determined by many factors including the cancers type, size and location. Treatment decisions, which are made by multidisciplinary teams of clinicians, are heavily influenced by medical scans which can determine all of these factors. Advancements in the fields of image processing and machine learning can therefore be applied to lung cancer imaging techniques to increase the amount of information available to clinicians from these images potentially improving patient outcomes. Two areas where image processing techniques can be used to potentially improve patient outcomes are radiotherapy outcome prediction and registration of PET/CT and pathology images.

1.2 Research Objectives

The aim of this work was to improve lung cancer treatment by employing medical image analysis techniques. The main objectives were to:

- **Improve radiotherapy toxicity prediction using image processing and machine learning based methods**

  Radiotherapy treatment requires a balance to be struck between maximising the control of cancerous regions while minimising the damage to healthy tissue. If

the dose to healthy tissue is too high a patient can develop toxicities, or adverse events, that can have serious consequences up to and including death. Currently, the dose to healthy tissue is limited by simple dose metrics that have been derived from statistical studies. Image processing and machine learning methods have the potential to increase the quality of predictions for how likely a patient is to develop these complications due to normal tissue toxicity. This would allow for more personalised treatment plans. One of the central research objectives of this is to assess how well pulmonary and esophageal toxicity can be predicted for lung cancer patients receiving radiotherapy using their pre-treatment information. This includes the use of radiomic and deep learning based approached to both CT images and radiotherapy dose images.

- **Produce a methodology for registering PET/CT images to pathology slices for lung cancer patients**

  PET/CT imaging is the most common method of non-invasive in-vivo imaging used in the diagnosis and treatment of most cancers including lung cancer. Currently, PET and CT imaging provide little information about the cellular make-up of a tumour and its environment. Increasing the knowledge that can be gained from these imaging modalities would aid clinicians in making treatment decisions which would improve the treatment and survival rates. To do this PET/CT images would need to be compared to pathological images of a tumour after it has been surgically removed. This would involve a registration of the different modalities of images so that they are as well aligned as possible. The research undertaken in this project follows on from a previous EngD project completed in 2020 by G. Reines-March [2]. Reines-March produced a registration framework for registering PET/CT images to histopathological slices of a surgically removed lung cancer tumour. While these methods worked well for CT imaging, it was found that due to motion induced by patients breathing during the scans that PET tumour volumes did not correlate to pathological volumes so the images could not be registered. One of the thesis aims was to improve on these techniques by including respiratory motion reduction techniques to the PET/CT scan and to produce a method for the segmentation of gross pathology photographs.

1.3 Contributions

The major contributions of the thesis are:

1. **Radiomic and deep learning features for radiation pulmonary toxicity prediction (Chapter 4).** Radiomic and deep learning features from CT images were used for the prediction of radiation pneumonitis in patients receiving IMRT. These features were additionally combined with dose and clinical features.

2. **Pulmonary toxicity prediction for SABR ILD patients (Chapter 4).** The use of dose, clinical, CT radiomic and CT deep learning features for the prediction of pulmonary toxicity in lung cancer patients with ILD receiving SABR is investigated. The use of SABR for patients with ILD is currently an emerging treatment option, this means that there has been no previous work on machine learning based prediction models for this specific radiotherapy cohort.

3. **3D-CNN for Esophageal toxicity prediction from dose maps (Chapter 5).** The novel application and development of a 3D-CNN for the prediction of esophageal toxicity from RT planning dose maps is completed. This includes a comparison to, and an ensemble with, more conventional approaches.

4. **Regression based training scheme for RT induced esophageal toxicity prediction (Chapter 5)**. A regression based machine learning training scheme was applied for the prediction of esophageal toxicity from radiotherapy planning dose maps.

5. **Registration of 4D-PET/CT to pathology (Chapter 6** The novel work for PET/CT to pathology image registration by Reines March, et al. [2] is advanced by the inclusion of respiratory motion reduction techniques for the PET images and by adapting the registration model to include PET tumour volumes.

6. **Automatic segmentation of gross pathology images (Chapter 6).** A methodology for the novel application of tumour segmentation gross pathology images is developed. This includes the first application of deep learning for the automatic segmentation of lung tumour in gross pathology. Previous works have focused on different applications and anatomical regions, such segmentation of endoscopy video [3], or have focused on non deep learning based approaches [4].

1.4 Thesis Structure

The outline of the thesis chapters is as follows:

- **Chapter 2** covers the background science necessary for the rest of the thesis. This includes sections on lung cancer, medical imaging and image processing.
- **Chapter 3** gives an overview of the literature relevant to the rest of the thesis.
- **Chapter 4** covers pulmonary toxicity prediction for lung cancer patients receiving radiotherapy.
- **Chapter 5** focuses on esophageal toxicity prediction for lung cancer patients receiving radiotherapy.
- **Chapter 6** includes the registration of 4D-PET/CT imaging to pathology for lung cancer patients and the development of a method for the automatic segmentation of gross pathology images.
- **Chapter 7** provides the final conclusions of the thesis and suggests some directions for future work.

Chapter 2

Background Science

2.1 Medicine, Anatomy and Treatment for Lung Cancer

2.1.1 The Lungs

The lungs are a pair of organs in the human body that facilitate the exchange of oxygen and carbon dioxide between the bloodstream and the external environment. This is essential for cells to receive the necessary oxygen for various metabolic processes while removing the waste product carbon dioxide.

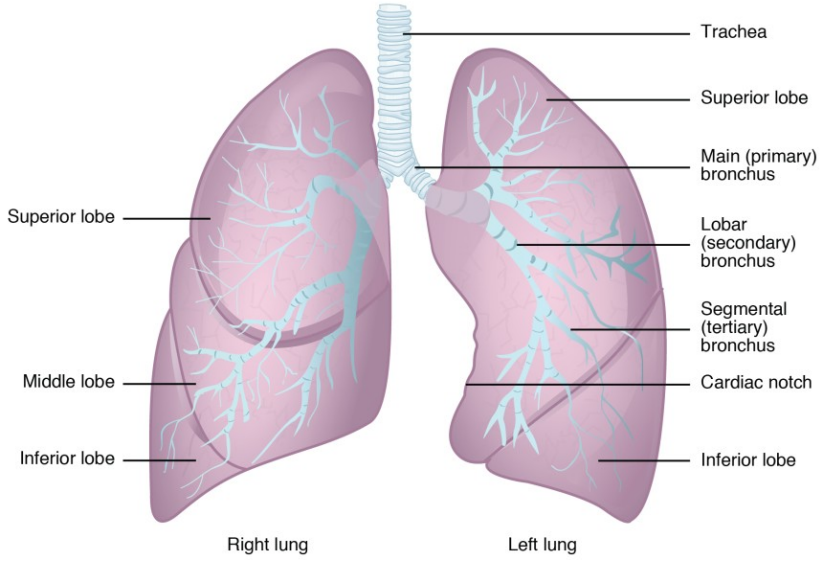


Figure 2.1: *Diagram showing the anatomy and major features of the lung. This image is taken from [5].*

Inhaled air travels through a branching system of tubes called the bronchial tree, eventually reaching tiny air sacs called alveoli. The walls of the alveoli are extremely

thin and surrounded by a dense network of capillaries allowing for the exchange of gases with the blood. All of the lung tissue involved in gas exchange is referred to as the lung parenchyma [6]. Oxygen from the inhaled air diffuses across the alveolar walls into the bloodstream, where it binds to hemoglobin in red blood cells. This oxygen-rich blood is then transported to various parts of the body. Carbon dioxide, a waste product produced by cells during metabolism, diffuses from the bloodstream into the alveoli. From there, it is expelled from the body during exhalation. The inhalation and exhalation process is controlled by the diaphragm and other respiratory muscles. The lungs are segmented into lobes that are separated by fissures with the right lung constituted of three lobes (superior, middle and inferior) and the left lung constituted of two lobes (superior and inferior) as it is slightly smaller due to the space required for the heart. The anatomy of the lung is shown in Figure 2.1.

2.1.2 Lung Cancer

Cancer is a broad term used to describe conditions where abnormal cell growth, with the potential to invade or spread to other parts of the body, is occurring. This abnormal cell growth results in a mass of cancer cells known as a tumour which is able to develop its own supporting blood supply. Cancer cells form through a process involving genetic mutations that change the normal regulation of cell growth and division. These mutations may be a result of gene inheritance or due to DNA damage of cells. DNA damage may happen naturally or be caused by external sources such as radiation, chemicals in food, air pollution, etc.

Lung cancer is the third most common cancer in the UK with around 48,500 people being diagnosed in the UK each year [7]. The 1-year and 5-year survival rates for lung cancer are around 40% and 15% respectively [7]. Depending on how early the cancer is diagnosed, the treatment options and prognosis drastically change. The level of development of a cancer is defined by what is known as its stage which can be stage I, II, III or IV. These stages can be split into subcategories but the general definitions are:

- **Stage 0:** This represents cancer that is in-situ, meaning it is localised and has not invaded nearby tissues. The cancer cells are present only in the layer of cells where they first developed and haven't invaded deeper tissues or spread to nearby lymph nodes. It is very rare to detect cancer at this stage and almost unheard of

in lung cancer

- **Stage I:** The cancer is localised and small in size. It has not invaded surrounding tissues extensively or spread to distant sites. The specific criteria for stage I can vary depending on the cancer type, such as tumour size and extent of invasion. For lung cancer, the maximum size for a stage 1 cancer tumour is 4cm.

- **Stage II:** Cancer in stage II might be larger or more invasive than in stage I. It may have spread to nearby tissues or lymph nodes though not extensively. Stage II is split into the subcategories A and B. For lung cancer, stage IIA means the tumour is less than 5cm while stage IIB means that the tumour is less than 5cm and it has spread to the most nearby lymph nodes within the lung.

- **Stage III:** Stage III lung cancer signifies an advanced phase where the cancer has extended beyond its initial site. It is divided into three subcategories: Stage IIIA, where the tumour is up to 5cm and has invaded central chest lymph nodes or is 5-7cm with multiple tumours, often involving nearby structures; Stage IIIB, where the tumour has grown larger (5-7cm), invaded multiple mediastinal lymph nodes, spread to collarbone lymph nodes, or affected vital structures; and Stage IIIC, the most advanced, involving extensive spread to multiple mediastinal lymph nodes, collarbone lymph nodes, or invasion of critical structures.

- **Stage IV:** Stage IV is the most advanced stage of cancer, indicating that it has spread to distant parts of the body, this is known as metastasis. Metastasis means that cancer cells have moved from the primary tumour site to other organs or tissues through the bloodstream or lymphatic system. Stage IV cancer can be split into the subcategories IVA and IVB. For lung cancer Stage IVA cancer has spread within the chest and/or has spread to one area outside of the chest and Stage IVB defines cancer that has spread outside of the chest to more than one place in one or to more organs.

In addition to the broad staging system, the TNM system defines parameters of the tumour, lymphatic nodes, and metastasis status using the T-stage, N-stage and M-stage respectively. The T-stage defines increasingly larger or more invasive tumours, often based on size, the extent of invasion, and involvement of nearby structures. The N-stage defines increasing involvement of regional lymph nodes, often based on the

number, size, and location of affected nodes. And the M-stage defines if metastasis is present.

There are several different types of lung cancer which are primarily categorised into two main groups: non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). SCLC is a more aggressive form of lung cancer that is closely associated with smoking and tends to grow rapidly and spread early. NSCLC is a grouping of cancers which includes adenocarcinomas, squamous cell carcinomas and large cell carcinomas. Adenocarcinoma is the most common subtype of NSCLC. It often develops in the outer parts of the lungs and is associated with both smokers and non-smokers. It tends to grow more slowly and is more likely to be found in an advanced stage. Squamous cell carcinomas usually arises in the central airways of the lungs and grow quickly. It's often linked to smoking and will generally cause symptoms earlier meaning it is usually picked up at an earlier stage. Large cell carcinoma is a group of cancers with large, abnormal-looking cells. They can occur in any part of the lung and tend to grow and spread quickly.

2.1.3 Interstitial Lung Disease (ILD)

Interstitial lung disease (ILD) is a term that defines a group of lung disorders affecting the interstitium (connecting tissue) and the space around the alveoli (air sacs) of the lungs. These regions can be damaged by autoimmune disorders, exposure to inhaled substances or may present idiopathically meaning that the damage presents spontaneously or there is an unknown cause. ILD generally presents as fibrosis, scarring and inflammation of the lung tissue and can often be diagnosed radiologically from computed tomography (CT) or magnetic resonance imaging (MRI) images. A summary of the different ILD subtype groupings is shown in Figure 2.2.

ILD can impact lung cancer treatment in multiple ways. Initially, the presence of ILD may make the diagnosis of lung cancer more challenging as many of the initial symptoms of ILD, such as coughing and shortness of breath, are the same as those of lung cancer. Once diagnosed with lung cancer, a patient with ILD may have fewer treatment options available due to the higher chance of treatment related complications. This can limit clinicians' ability to use chemotherapy and radiotherapy to treat lung cancer in patients with ILD and often a balance has to be found between effectively treating the lung cancer and preserving lung function.

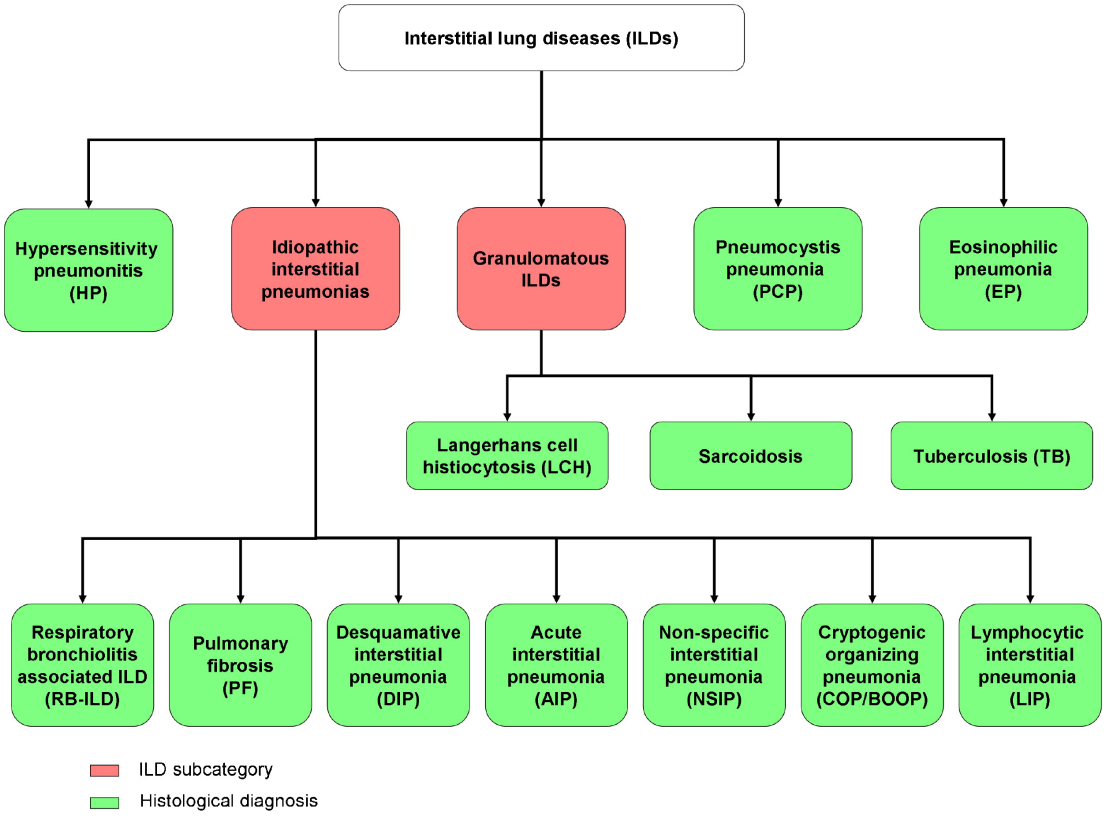


Figure 2.2: *Flowchart showing the groupings of ILD subtypes and subcategories. Image is taken from [8].*

2.1.4 Treatment Work Flow for Lung Cancer Patients

The choice of treatment for lung cancer depends on the cancer stage, the location of the tumour and a patient's general health. For patients with stage I or stage II lung cancer, the most common treatment is surgical removal of the tumour. For more advanced stage III and IV lung cancers, chemotherapy, radiotherapy or a combination of the two is more common. In general, the workflow for treating a patient with lung cancer involves the following steps:

1. **Diagnosis and Staging:** The treatment workflow starts with a thorough medical history, physical examination, and imaging tests (such as X-rays, CT scans, and PET scans) to diagnose and determine the stage of the cancer. Tissue samples are usually taken, to confirm the type of lung cancer and its specific characteristics.

2. **Multidisciplinary Review:** Once the type, location and stage of the cancer has been determined. A team of medical professionals including oncologists, sur-

geons, pathologists, radiologists and other specialists, review the patient's case and develop a personalised treatment plan from the available treatment options.

3. **Treatment:** The treatment plan is tailored to the patient's specific situation. For early-stage lung cancer, surgery may be the primary treatment option, while more advanced cases may require a combination of treatments such as radiation therapy, chemotherapy and immunotherapy. The order of treatment application can vary. For example, a patient might receive chemotherapy before surgery to shrink the tumour or after surgery to eliminate any remaining cancer cells. If the patient receives surgery, there will be a pathological analysis of their sample which provides more information about the cancer which may inform future treatment options.

4. **Monitoring and Followup:** After treatment, patients undergo regular follow-up appointments and scans to monitor their progress and check for any cancer recurrence.

The treatment options that are relevant to the work in this thesis are discussed in the following sections.

2.1.5 Radiotherapy as a Lung Cancer Treatment

Radiotherapy (RT) is a form of cancer treatment that uses ionising radiation to control or kill cancer cells. Depending on the cancer type, location and stage, it may be used as the primary treatment with curative intent or it may be used as an adjuvant treatment such as its use after surgery to reduce the chance of tumour recurrence [9]. The unit of radiation dose is the Gray (Gy) which quantifies the amount of energy deposited by ionizing radiation in a material per unit of mass. RT dose is most commonly applied by an external beam of particles but may also be applied by brachytherapy where a sealed radiation source is placed inside or adjacent to a cancerous region. Only external beam radiotherapy (EBRT) is relevant to the work in this thesis so it is the only mode of RT that will be discussed further.

External beam RT uses a beam of high energy particles to apply a radiation dose to a tumour. The particles used may be photons, electrons, protons or heavier nuclei such as carbon ions. To create the particle beam used in EBRT, generally, a linear accelerator (Linac) is used. A linac creates a particle beam by subjecting charged particles to

oscillating electric potentials, causing acceleration of the particles along a length. The particles used in a linac are generally electrons as heavier protons or nuclei usually require a circular accelerator to reach the required speeds. To produce photon beams from a linac, an x-ray target can be included before the beam output to convert the energy in the electrons into high energy photons through the Bremsstrahlung process [10].

The RT radiation dose is not applied in a single RT session, instead, dose fractionation is used to deliver the total prescribed radiation dose in smaller, divided doses over multiple treatment sessions or fractions. Fractionation allows healthy tissues surrounding the tumour to recover and repair between treatment sessions, reducing the risk of adverse effects. Additionally, dividing the total dose into fractions can improve the ability of radiation to kill cancer cells. This is because tumour cells with more access to oxygen are more sensitive to radiation so the previous fraction will kill the oxygen rich cells and allow the oxygen starved cells to access more oxygen between fractions. The timing of the fractionation is chosen to maximise this increased tumour sensitivity while minimising the potential for the tumour cells to repair and repopulate between RT sessions [11].

There are several different methods for the application of EBRT. All methods involve the radiation source, usually a linear accelerator (linac), mounted on a gantry that can rotate around the patient who is lying on a treatment table in the centre of the gantry. The gantry rotation allows for the precise delivery of radiation beams from various angles to target the tumour. Additionally, modern linacs are equipped with imaging systems, such as cone-beam CT, which provides close to real-time imaging of the patient's anatomy just before treatment. This imaging capability enhances accuracy by confirming the patient's position and ensuring the treatment beams are precisely aligned with the target. Some of the main EBRT methods and those relevant to this thesis are:

- Three-dimensional conformal radiotherapy (3DCRT)
- Intensity Modulated Radiotherapy (IMRT)
- Volume Modulated Arc Therapy (VMAT)
- Stereotactic Ablative Radiotherapy (SABR)

These are described in more detail in the following paragraphs.

Three-dimensional conformal radiotherapy (3DCRT) Three-dimensional conformal radiotherapy (3D-CRT) is a RT method that improves on conventional radiotherapy by adapting the beam to more accurately irradiate the tumour area that is defined by 3D delineation of CT or MRI scans. This is achieved by using a multi leaf collimator to shape the beam to the tumour and multiple beam angles that will line up on the tumour to apply a maximum dose to the tumour while sparing more healthy tissue. This allows for a higher tumour dose than conventional RT techniques would allow.

Intensity Modulated Radiotherapy (IMRT) Intensity modulated radiotherapy (IMRT) advances on the techniques of 3D-CRT by introducing an intensity modulation to the beam dose which allows for improved dose tailoring for a patients specific tumour shape. Computational methods and simulations are used to calculate the optimal dose to the tumour while minimising the dose to healthy tissue. Additionally, more beams are usually applied during IMRT than 3D-CRT, minimising high dose regions in healthy tissue. These benefits are most beneficial in complicated anatomical regions such as head, neck and lung as there are many OaRs in these regions that require dose minimisation.

Volume Modulated Arc Therapy (VMAT) Volume modulated arc therapy (VMAT) was first introduced clinically in 2007 [12] and has become the gold standard method of EBRT since then. VMAT improves on IMRT by using continuous beam angles, this means that the dose is applied continuously while a gantry rotates the beam around the patient essentially giving an infinite number of beam locations in a 2D plane and more scope for shaping the applied three dimensional dose applied to the patient. While the beam is rotated around the patient, three parameters are altered to produce the ideal dose, these are; the shape of the beam which is altered with a multi-leaf collimator, the speed of rotation and the rate of dose delivery. This again allows for an increase in healthy tissue sparing and dose uniformity over the tumour volume. While reducing regions of high dose in healthy tissue, VMAT techniques increase the volume of healthy tissue receiving a low, but non-zero, dose.

Stereotactic Ablative Radiotherapy (SABR) Stereotactic ablative radiotherapy (SABR) is a RT method used to deliver a high radiation dose to smaller tumours [13]. The term "stereotactic" refers to the use of a precisely calculated coordinate system to ensure accurate radiation delivery. Additionally, SABR delivers a high dose of radiation in just a few fractions as opposed to the many fractions of other RT treatments. This means that a higher dose is delivered per fraction. Due to the high precision of SABR, damage to healthy tissues is minimised which reduces the risk of side effects commonly associated with RT.

2.1.5.1 Radiation Induced Toxicities in Healthy Tissue

The biological damage that RT causes is not exclusive to cancerous cells. Healthy cells in all organs and areas of the human body are susceptible to this damage which may lead to toxicity or adverse events. This is especially a problem in external beam RT as the radiation beams must penetrate through healthy tissues in order to reach the cancerous cells meaning the dose to healthy tissue will be higher than other forms of RT. As defined in [14], normal tissue toxicity can be grouped into three categories based on the time it takes for them to occur, these are:

- **Early Effects:** These generally occur within 60 days of treatment and are due to acute cell death. This generally occurs in tissue with a quick cell turnover. These effects can be completely healed if the damage is not too great.
- **Late Effects:** These generally occur at least 60 days post-treatment and are due to mechanisms other than acute cell death such as fibrosis and vascular damage. These effects are rarely fully repaired.
- **Consequential late effects:** These are caused by early effects that are severe enough to cause permanent damage. An example of this is skin necrosis that requires a skin graft.

Not all healthy tissue is equally sensitive to the harmful effects of radiation. Different organs have different radio-sensitivities and develop different conditions. One differentiation between different organs is if they are serial or parallel organs [14]. A serial organ will entirely cease to function if there is a loss of function in one part of the organ. An example of a serial organ is the spinal cord. In serial organs, great care

has to be taken not to exceed a maximum dose threshold. Parallel organs on the other hand can lose function in part of their total volume while maintaining function in the rest of the organ. The lungs are an example of a parallel organ. The risk of injury in parallel organs is influenced most by the average dose as opposed to the maximum dose.

There are many different conditions caused by radiation exposure to different organs. These are exhaustively listed in the Common Terminology Criteria for Adverse Events (CTCAE) [15] with version 5.0 being the most recent version. Each toxicity that occurs can be given a grade to define its severity. Grades range from 1 to 5 with grade 1 being the least severe and grade 5 indicating death due to the specific toxicity. Additionally, grade 0 is often used to define no observed toxicity. The grade definitions are different for each specific toxicity and are defined in CTCAE. The general grading scheme is given in table 2.1.

Toxicity Grade	**Description**
1	Mild; asymptomatic or mild symptoms; clinical or diagnostic observations only; intervention not indicated.
2	Moderate; minimal, local or noninvasive intervention indicated; limiting age appropriate instrumental ADL.
3	Severe or medically significant but not immediately life-threatening; hospitalisation or prolongation of hospitalisation indicated; disabling; limiting self care ADL.
4	Life-threatening consequences; urgent intervention indicated.
5	Death related to adverse event.

Table 2.1: *Adverse event grade definitions from the Common Terminology Criteria for Adverse Events (CTCAE) ver 5.0 [16]. A semi-colon indicates 'or' within the description of the grade. ADL - activities of daily life.*

Radiation Pulmonary Toxicity For patients receiving external beam RT to target a lung lesion, a potential side effect of the treatment caused by the dose received by healthy lung tissue is the development of radiation pulmonary toxicity. The most common pulmonary toxicities are radiation pneumonitis (RP) and dyspnea. RP results from an inflammatory response within the lung tissue due to the radiation damage to the cells that line the alveoli in the lungs. RP is a serious condition causing breathing

issues in patients even with supportive measures for RP such as supplemental oxygen, steroids or mechanical ventilation, RP is a potentially fatal complication [17]. RP has also been observed in patients receiving RT treatment for other cancers in the chest region such as breast cancer [18].

Radiation Esophageal Toxicity Radiation esophageal toxicity is another group of radiation toxicities often experienced by lung Cancer patients due to the close proximity of much of the esophagus to the lungs. Radiation esophagitis (RE) is the most common radiation esophageal toxicity. RE is characterised by an inflammation and irritation of the esophagus. The main symptoms of RE are pain and nausea, in severe cases the symptoms will impact a patient's ability to eat and drink and complications from the most severe cases may be fatal [19].

2.1.5.2 Radiotherapy Planning

Before RT can be delivered an advanced treatment planning process is performed which aims to produce a plan for the application of the radiation dose that will give the patient the best outcome in terms of both tumour control and limiting toxicity in healthy tissues. The treatment planning process starts with a consultation between the patient and their radiation oncologist. The oncologist reviews the patient's medical history, performs a physical examination, and discusses their treatment options. Once EBRT is chosen as a treatment option, the next step is to conduct imaging scans to visualise the tumour and the surrounding structures. Imaging is performed using MRI or CT scans. During this imaging, the patient is positioned on a treatment table in the same way they will be during their actual treatment so that these pre-treatment scans match the treatment as closely as possible. After this, the radiation oncologist, in collaboration with other clinicians, delineates the target volume in the images, this includes the tumour and any nearby lymph nodes or areas at risk of containing cancer cells. Organs at risk are also identified to minimise radiation exposure to these areas. The radiation dose map is then calculated by treatment planning software in a semi-automatic approach using dose constraints for the target volume and organs at risk [20].

2.1.6 Surgery as a Lung Cancer Treatment

Surgical removal of a lung tumour is the most common form of treatment for patients with stage I and II lung cancers as these early stages indicate that it is unlikely that the cancer will have spread outside of the lung meaning the aim of surgery is often to remove all of the cancerous material [21]. It is also generally a goal of lung cancer surgery to remove any nearby lymph nodes that might be affected. There are several types of surgery that may be performed depending on factors such as the stage and location of the cancer, the patient's overall health and their lung function.

A lobectomy is the most common type of surgery for lung cancer. A lobectomy involves removing the entire lobe of the lung where the cancer is located. A segmentectomy removes a portion of the lung that is smaller than a whole lobe, this can be done in cases where the tumour is small and located in a peripheral area of the lung. This approach can preserve more lung function than a lobectomy while still removing the cancerous tissue. In some cases, when the cancer is located in the central part of the lung or involves a larger portion of the lung, a pneumonectomy might be necessary. This procedure involves removing the entire lung on one side. There are also some minimally invasive techniques available to surgeons such as video-assisted thoracoscopic surgery or robotic-assisted surgery. These approaches involve smaller incisions where a camera and surgical instruments are inserted for the surgery.

2.1.7 Quantifying Patient Health

During a patient's treatment, it can be beneficial to quantify features of their health in some way to aid with monitoring their treatment progression. This is particularly important during clinical trials where extra detail is usually required for patient monitoring and health based metrics may be the final outcome. Most of these patient monitoring methods involve getting the patients to fill out questionnaires that have been designed to produce health monitoring metrics. Two of the most commonly used health forms for lung cancer patients are the EuroQol 5-Dimension 5-Level (EQ-5D-5L) and Functional Assessment of Cancer Therapy (FACT) which are detailed below.

- **EuroQol 5-Dimension 5-Level (EQ-5D-5L)**: The EQ-5D-5L is a standardised questionnaire used to measure the health-related quality of life of patients [22]. It involves getting patients to rate five separate aspects of their health with a rating

of 1 to 5 where a rating of 1 signifies "no problems" or "no difficulty" in the assessed dimension and a rating of 5 indicates "extreme problems" or "extreme difficulty". The five assessed health aspects are; mobility, self-care, disruption of usual activities, pain/discomfort and anxiety/depression. These dimensions are often combined into a single metric with a range from 0 to 1 to summarise the patient's health.

- **Functional Assessment of Cancer Therapy (FACT)** The FACT scale is a collection of questionnaire based assessments that aim to quantify health-related quality of life in patients with cancer [23]. FACT-General (or FACT-G) is the general questionnaire and contains 27 questions relating to the patient's physical well-being, social/family well-being, emotional well-being, and functional well-being. In addition to this general form, there are several forms for specific cancer types. FACT-Lung (or FACT-L) is the FACT form specifically designed for lung cancer patients which includes 35 questions addressing lung cancer symptoms, respiratory issues, and concerns about breathing difficulties. The FACT forms are used to generate a single metric, with a range of 0 to 100, to describe the patient's health.

2.1.8 Histopathological Processing for Lung Cancer

Histological analysis of cancer specimens provides cellular information about a tumour which is the ground truth for a diagnosis to determine if a tumour is cancerous and what type of cancer it is. To perform a pathological analysis, a specimen has to be removed from the patient and examined ex-vivo. The removal of a tissue sample for pathological analysis is called a biopsy. Biopsies are usually taken early on in the patient's treatment process to confirm if a tumour is cancerous. For this purpose only a small sample is required and can be obtained through various methods such as bronchoscopy or a needle biopsy. In addition to the main tumour, the nearby lymph nodes are often biopsied to assess the spread of the cancer. The removal of a tumour by surgery for curative intent, as described in section 2.1.6, creates a large biopsy for analysis. The standard practice is to perform a pathological analysis of any surgically resected lung tumours.

Once a tissue sample is obtained, it is preserved in a formalin solution in a process called fixation. Fixation prevents tissue decay and prepares the sample for further processing. If the sample is large, such as one produced by curative surgical resection

(lobectomy, segmentectomy, etc.), it may then need to be sliced by a pathologist to expose the tumour and create multiple smaller samples instead of a single large sample. After this, the fixed tissue is dehydrated and embedded in paraffin wax which allows for thin tissue sections to be cut and placed on glass slides for microscopic imaging. The tissue sections are then stained with dyes that highlight different cell structures. There are many types of dyes with the most commonly used dye combination being hematoxylin and eosin. Hematoxylin stains cell nuclei blue-purple, providing information about cell density and arrangement, while eosin stains cytoplasm and extracellular structures pink, making them easier to differentiate.

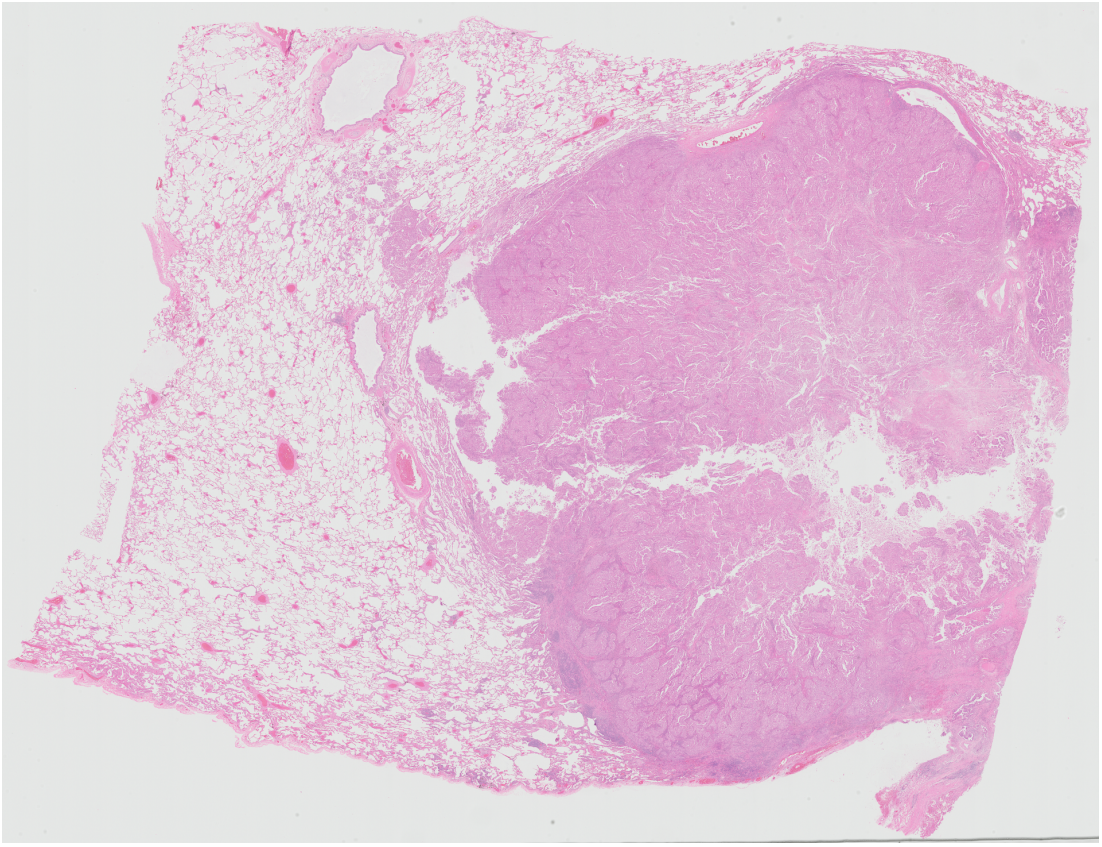


Figure 2.3: *A histology whole slide image (WSI).*

A pathologist would historically examine the stained slides under a microscope to assess the tissue's cellular characteristics and identify any abnormal or cancerous cells. Digital pathology is the modern approach where the samples are microscopically imaged as whole slide images (WSI) so that the pathologist can view the image on a screen, an example of a WSI is shown in Figure 2.3. When examining the sample, a pathologist will give the cancer a grade from well-differentiated to poorly-differentiated which describes

how closely cancer cells resemble normal cells. This is important as the more abnormal the cancer cells are, the more aggressive the cancer will be. In addition to microscopic image examination, the pathologist may also conduct molecular testing to look for specific genetic mutations or biomarkers that can influence treatment decisions.

2.2 Medical Imaging

Medical imaging is a broad field that includes imaging from any modality that is applied for a medical purpose. Medical imaging can be used to highlight anatomy or physiology (the function of organs). Medical images can be 2D, giving a planar image, 3D, giving a volume, or 4D which refers to a 3D image with an additional time dimension. Figure 2.4 shows the medical imaging planes, the horizontal, or transverse, plane is the most common plane for viewing 3D medical images.

2D medical imaging has some application when exposed surfaces are to be viewed such as during an endoscopy or through digital photography for skin lesions. Most medical imaging is generally 3D where the interior of the body can be imaged by non-invasive means. Examples of 3D imaging include computed tomography (CT), magnetic resonance imaging (MRI) and positron emission tomography (PET). 4D imaging is generally only used in specific cases for the removal of motion blurring effects over the duration of a scan that can be caused by a patients breathing or heartbeat among other sources. Medical imaging is almost always performed in-vivo. The main ex-vivo application in medical imaging is in digital pathology as discussed in Section 2.1.8.

The imaging modalities that have been used in the work detailed in this thesis are CT, PET and pathology. These are discussed further in the following sections.

2.2.1 Computed Tomography (CT)

Computed tomography (CT) imaging is a form of medical imaging that is able to create high-resolution 3D images of any part of the human anatomy. For this reason, it is one of the most widely used forms of medical imaging for diagnosis. CT imaging uses the attenuation of radiation, generally X-ray photons, through the human body to build an image. Areas of different levels of attenuation will correspond to different tissue types and can therefore be used to build a map of the human body. An example of a CT image of a lung cancer patient is shown in Figure 2.5. The radiation used for the imaging

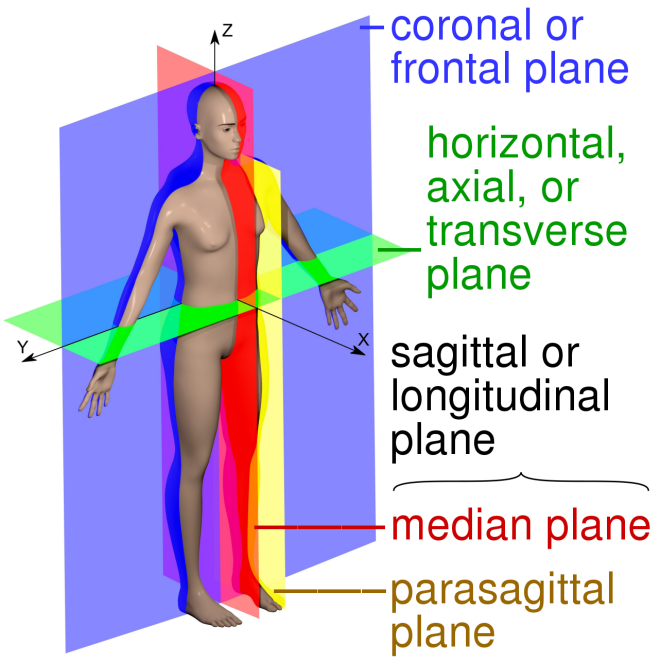


Figure 2.4: *The anatomical imaging planes used in medical imaging. Figure is taken from Wikipedia (*`wikipedia.org/wiki/Anatomical_plane`*).*

does produce an increased cancer risk when applied to the human body though the dose is never strong enough for acute effects such as those that occur during radiotherapy. There is, therefore, a trade-off between image quality and the radiation dose to the patient with a higher dose producing a higher SNR and allowing for increased spatial resolution.

An X-ray tube is generally used to produce the radiation which is detected by a row of detectors which will be on the other side of the patients body to the X-ray tube. The x-ray tube and detectors are held in a gantry that rotates around the patient during the scanning process allowing for a 3D image to be formed. The rotational imaging forms a sinogram which is the image in the gantry angle and time coordinate space. To convert the sinogram into cartesian coordinates, a reconstruction algorithm is applied, the most commonly applied currently are iterative reconstruction algorithms [24].

CT images are greyscale images where voxel intensity is displayed in terms of relative radiodensity, the unit of which is the Houndsfield unit. The Houndsfield unit is defined so that distilled water at standard pressure and temperature has HU value of 0 and air at standard pressure and temperature has an HU value of -1000. This means that the intensity of a voxel in a CT image with average linear attenuation coefficient μ will

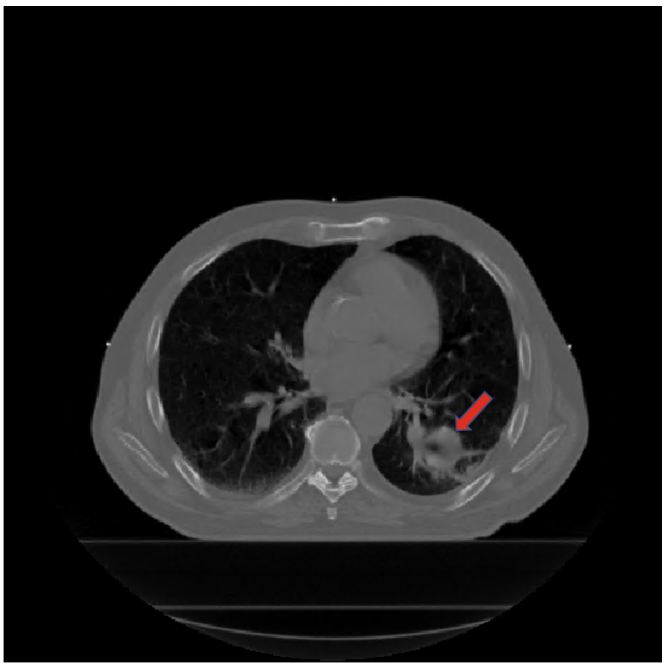


Figure 2.5: *An example CT image slice, imaged in the transverse plane, of the chest of a patient with lung cancer. The primary lung tumour is indicated by a red arrow.*

have a HU value defined by equation 2.2.1.

$$HU = 1000 \times \frac{\mu - \mu_{\text{water}}}{\mu_{\text{water}} - \mu_{\text{air}}} \tag{2.1}$$

Here μ_{water} and μ_{air} are the linear attenuation coefficients of water and air respectively. The HU values for some common tissues are shown in table 2.2 [25].

Tissue	**Houndsfield Units (HU)**
Air	0
Lung	-900 to -500
Fat	-100 to -50
Water	0
Muscle	10 to 40
Bone	700 to 3000

Table 2.2: *The HU values for some common tissues [25].*

As CT images can be high-quality images they are the most commonly used for diagnostic imaging. In the context of lung cancer, and generally in oncology, CT images are often used as the first diagnostic tool which may identify lesions and rule out other diagnoses. As CT images are maps of the attenuation of radiation of the

human body, they can provide useful information for other imaging modalities and treatment options. In nuclear medicine based imaging, such as PET imaging, a CT scan is often taken alongside the nuclear medicine modality to provide an attenuation map that can be used during the image reconstruction process. Additionally, during external beam radiotherapy, CT images are acquired before the patient receives their treatment so that anatomy can be identified and the radiation dose to all body parts can be calculated.

2.2.2 Positron Emission Tomography (PET)

Positron Emission Tomography (PET) is a 3D functional imaging modality that uses the radiation emitted by radionuclides to image regions of metabolic activity in the body. PET is a form of nuclear imaging which generally works by attaching a radionuclide to a drug, to form what is known as a radiotracer, and injecting it into the body. The body then processes the tracer which sends it to different parts of the body depending on which tracer is used. The concentration of the tracer at different anatomical locations can highlight the level to which a physiological process is occurring. The concentration of the radiotracer can be imaged by detecting the radiation produced by the radioactive decay of the radionuclide [26].

PET imaging relies specifically on β^+ decay which is a form of radioactive decay involving the conversion of a proton to a neutron resulting in the release of a positron. The released positron will travel until it collides with an electron at which point they annihilate and produce two identical gamma ray photons that are emitted in opposite directions. To detect these photons, a ring of detectors is placed around the patient. The β^+ decay process has some special properties that can be taken advantage of during PET imaging. The production of two gamma rays at the same time allows for the source of the annihilation to be located within a 2D plane. This is done by timing the photon detections which can be matched to the other photon produced by the same annihilation. The detector location and detection times for the two photons can then be used to determine the location of the annihilation event. A diagram of the detection of an annihilation event is shown in Figure 2.6.

The choice of radionuclide in PET imaging depends on its decay mode, positron energy and half-life. The half-life is perhaps the most important factor as too short of a half-life will make the imaging process more difficult but too long of a half life

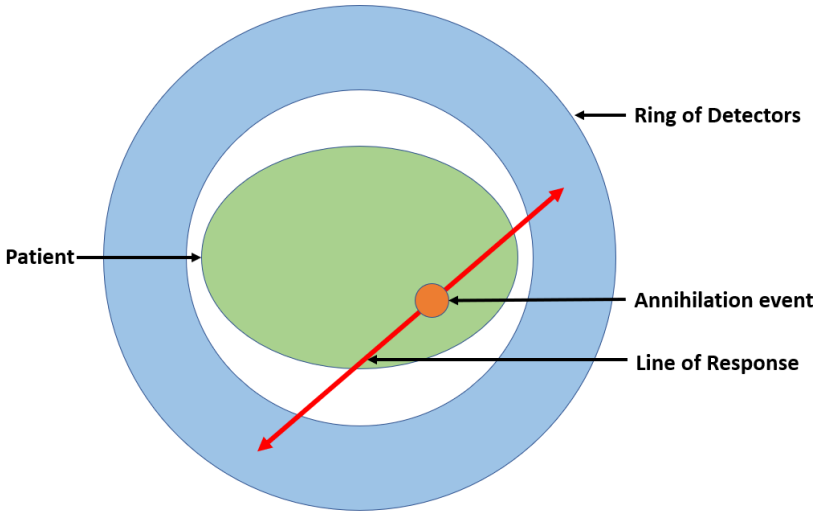


Figure 2.6: *Diagram of the patient position within a PET detector ring and the detection process for positron annihilation events.*

will expose the patient to an additional radiation dose. The choice of tracer depends on what physiological process is to be imaged. One of the most common tracers is a fluorine labelled glucose molecule called fluorodeoxyglucose (18F-FDG). 18F-FDG highlights regions of metabolic activity making it useful in detecting and localising cancerous tumours which are highly metabolically active. 18F-FDG PET is the only radiotracer used in any of the PET imaging in this thesis. An example of an 18F-FDG PET image is shown in Figure 2.7 (a) which highlights an area of high uptake in the lungs, indicating a tumour.

To acquire a 3D PET image, the patient is moved through the detector ring over a length of time. This is achieved by having the patient lie on a bed that moves as is done with CT imaging. The time it takes for a PET scan is in the order of minutes and depends on the anatomy being imaged and the resolution required with slower scanning speeds producing a higher SNR as more counts will be detected. Scans will generally be under 20 minutes to reduce the clinical demand and to limit the length of time the patient will have to remain motionless.

2.2.2.1 The Standard Uptake Value (SUV)

The voxel intensity in a PET image is directly proportional to the radioactivity concentration within the body at that voxel location. This concentration is commonly expressed in terms of the standard uptake value (SUV). The SUV is a quantitative measure of the radiotracer uptake in a region of a PET image that aims to reduce

variability between different patients due to differences in their weight. The SUV is defined by equation 2.2 [27].

$$SUV = \frac{r}{a'/w} \tag{2.2}$$

Here, r is the radioactive activity concentration (kBq/ml) that the PET scanner measures, a' is the decay corrected activity (kBq) of the injected radiotracer and w is the patient's weight (kg) which is used to create an estimate of the distribution volume of the tracer.

While the SUV remains a popular quantification tool, it has many sources of error that are mostly unavoidable in PET imaging. Firstly, there can be a large variation in patient-based variables aside from weight. This includes patient anatomy, the natural glucose levels of the patient at the time of imaging and a patient's renal function. The imaging physics of the PET scanner also contributes to the error in the SUV. The usual sources of noise such as background and scatter noise, as well as properties of the detector such as the dead time of the detectors, increase the uncertainty of any individual voxel's SUV.

2.2.2.2 PET/CT Imaging

A PET scan is often acquired in conjunction with a CT scan in a PET/CT combined scan. This allows for the functional information available in the PET image to be mapped to the anatomical information available in the CT image [28]. Additionally, the CT scan provides an attenuation map that can be used to create an attenuation corrected PET image producing a more accurate map of the concentration of the radiotracer. An example of a fused PET/CT image is shown in Figure 2.7 (c) where a lung tumour is highlighted by the PET image.

2.2.3 PET Respiratory Motion Reduction

Generally, the largest source of error in PET and CT imaging is patient movement [29]. Patient movement can cause misregistration artefacts such as blurring or streaking to occur in the final image as part of the anatomy has moved during the imaging process. This is an issue for imaging any part of the human body though this problem is especially significant in imaging of the chest or abdomen due to movements arising

Figure 2.7: *Example of* ***(a)*** *a PET scan,* ***(b)*** *a CT scan and* ***(c)*** *a fused PET/CT image of the chest and torso region of a lung cancer patient. A cancerous lung tumour is highlighted by the PET image as an area of high uptake. All images are displayed in the coronal plane and were acquired on a joint PET/CT scanner during the same scanning process for a single patient.*

from a patient's breathing cycle. This is usually not an issue in CT imaging as the images are taken over a few seconds but for PET imaging, where each field of view is imaged over several minutes, many breathing cycles will have occurred over the time taken to produce a full image. This can be counteracted in some imaging modalities by making use of quick scan times or by implementing breath-hold instructions for the patient to follow at certain times during a scan. Neither of these options are implementable with a standard PET chest scan as the scan length is too long and can not be reduced as this would reduce the SNR and breath-hold methods have been found to be ineffective over long scans due to patient error.

Respiratory motion reduction is a group of techniques developed to reduce the image blurring due to patient breathing by tracking the patient's breathing cycle over the length of the scan. The image is generally then reconstructed using only data from one repeating section of the breathing cycle for the final image. A review of respiratory gating methods can be found in [29]. To apply respiratory gating a method of tracking a patient's breathing cycle during a scan must first be used.

2.2.3.1 Device Based Tracking Methods

Many methods of tracking respiratory motion during a scan have been developed. Some of these methods involve using an additional device during the scan to produce a respiratory signal, systems have been designed to do this by measuring pressure changes on a belt wrapped around a patient's chest, the volume of air inhaled and exhaled or changes in temperature of the patients inhaled and exhaled air [29]. An example of a deviceless

method is the revolutionary gating for scanners (RGSC) system, developed by Varian Medical Systems (Varian Medical Systems, Palo Alto, California, USA) which involves tracking the movement of markers placed on a box on the patient's chest. No device based gating methods have been used in this thesis so there is no more detail given here.

2.2.3.2 Deviceless Tracking Methods

Deviceless tracking methods, sometimes called data-driven tracking, estimate a respiratory motion signal directly from the imaging data. These methods remove the requirement for hardware in respiratory tracking which has the advantage of reducing the setup time of the patient before imaging. Another disadvantage of device-based tracking methods that deviceless methods can solve is that they record a single signal for the respiratory motion of the patient which is taken to correspond to the respiratory motion at every anatomical location. This is an inherent limitation of most device-based tracking systems as they usually rely upon the physical location of a device which can only be placed at a single anatomical location. It has been shown that respiratory motion is highly dependent on anatomical location and that there can be time-varying phase differences in the respiratory motion at different locations [30]. Deviceless methods that measure the respiratory motion of a region from the data acquired in only that region can therefore produce more accurate estimations of localised motion.

Many deviceless methods for PET imaging rely upon a PET specific spectral analysis method detailed in [31]. This spectral analysis involves taking the fast Fourier transform (FFT) of the PET data in 4D sinogram space to produce a peak at the frequency of respiratory motion. Areas that are subject to respiratory motion can be defined by the presence of a peak within a frequency range that could correspond to human breathing. A respiratory signal is then produced for the areas where motion is present by recording the variation of the integrated counts within a region. To remove contributions from motion that does not correspond to the respiratory signal, a weighting based upon the phase of the motion is added to only include motion in phase with the main frequency component found by the FFT.

Deviceless methods for PET imaging have an inherent problem where the relationship between the polarity of the signal and the change in intensity at a given location is ambiguous. This is because exhalation and inhalation could cause an increase in in-

tensity at one location and a decrease at another location as an inhalation could move an anatomical point of high intensity both closer to or further from an image voxel location depending on the position of the point. P. Schleyer et al. propose a solution for this in their 2018 paper describing a method called continuous bed motion automated spectral analysis (CASA) [32]. This method is the only gating method applied in this thesis so it is reviewed in detail here.

CASA uses continuous bed motion during the imaging to introduce a motion with known speed and direction into the imaging process that can be used to remove the ambiguity in the signal for regions of inhalation and exhalation. This is different from the usual acquisition process for PET imaging where the bed moves in a step-and-shoot manner, although newer scanners are more likely to use continuous bed motion. For CASA, a signal generated by spectral analysis, referred to as signal B, is compared to an initial global signal estimate generated by observing the standard deviation of the activity distribution variation in the anterior-posterior direction over time, referred to as signal A. This standard deviation (signal A) can generate a global respiratory signal estimate because an increasing standard deviation will correspond to inhalation as the total anatomical area being imaged will increase, spatially spreading out the areas of intensity. Signal B then has a phase shift of either 0° or 180° applied to it so that it most closely matches signal A. This defines the global phase for signal B and the timing of inhalation and exhalation is known. At this point signal B is an accurate local estimate of the respiratory motion. The CASA method has been clinically evaluated in [33] and found to give results at least as good as hardware based methods. Figure 2.8 shows an example from [32] of the results from both hardware and deviceless based gating methods for a PET scan of a lung tumour.

(a) Hardware (b) CASA

Figure 2.8: *A sagital slice of a respiratory gated PET scan of a lung tumour using (a) an Anzai respiratory gating belt device and (b) CASA data driven gating. Image taken from [32].*

2.2.3.3 Respiratory Gating

The general method to process the respiratory signal along with the image data to produce a motion corrected image is to gate (or group) the data collected into separate bins that correspond to different sections of the respiratory cycle. The number of bins defined produces a trade-off between the deblurring and the signal-to-noise ratio of the image data in each bin. A higher number of bins reduces the blurring associated with breathing but reduces the signal-to-noise ratio of each bin. A reduction in the blurring results in more accurate lesion volumes while a reduction in the SNR results in a decrease in lesion detectability. It has been reported that six bins per respiratory cycle is optimal for cardiac scans [34].

The different parts of the respiratory cycle can be grouped as amplitude-based or time-based methods. For time-based approaches, there are many methods to separate the different phases of the respiratory cycle. The simplest method is to create bins of a fixed time for all repetitions of the respiratory cycle [35]. Due to the irregular frequency of the respiratory cycle, this method causes some of the data to be discarded. Alternatively, the timing length of the bins can vary from each respiratory cycle so that each cycle contains the same number of bins [36], this is shown in Figure 2.9 (a). For this method, data from respiratory cycles outside of a user-defined frequency range would be discarded to remove irregular data.

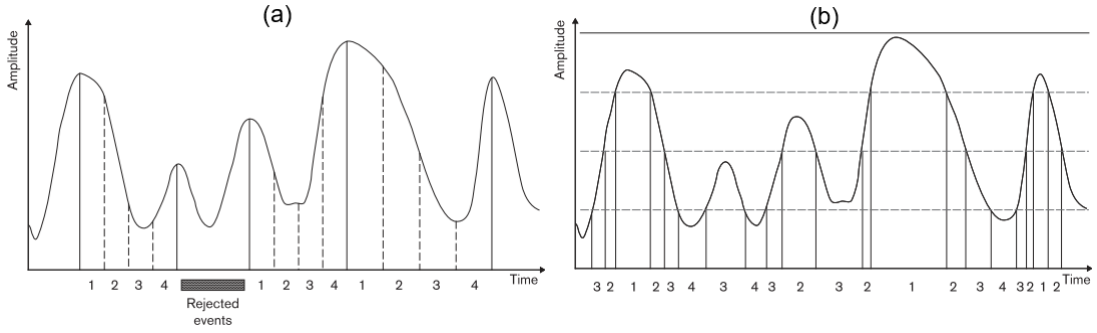


Figure 2.9: *Respiratory gating of an example respiratory signal using four bins per cycle with (a) a time-based method where bin timings vary between respiratory cycles and (b) an amplitude based method where the bins contain equal ranges of amplitudes. Figure taken from [29].*

It has been shown that amplitude-based gating methods generally produce better results than time-based methods [35]. The simplest method of amplitude-based gating is when all the bins contain an equal range of amplitudes, this is shown in Figure 2.9 (b). An issue with this method is that the bins are unlikely to contain similar coincidence

numbers so the SNR will vary across the bins. Variable bin processing can be used to keep the SNR constant across the bins at the cost of varying the motion amplitude in each bin [35]. There is no ideal number of bins to use in amplitude-based methods as Bettinardi et al. [37] suggest that the ideal number of bins is related to the size and displacement of a lesion.

Once the PET data has been grouped into bins using the timing information recorded during the scan, an image can be produced for each bin. Each image corresponds to a different part of the respiratory cycle so any lesions will have a different location in each image. The gated images can be analysed on their own or some additional processing may be applied to combine the gated images into a single image. One method to combine the gated images is to deformably register all of the images to one particular image. This method uses all of the collected data which ensures the best possible SNR at the cost of reducing the resolution due to the uncertainties in location introduced by the deformable registration. Another method is simply to use only one of the gated images, generally the image with the lowest level of motion blur. This method produces a high spatial resolution due to a low spatial uncertainty but produces a low SNR as a large portion of the data collected has not been used in the final image. This method may require increased scan times to improve the SNR.

2.2.4 Digital Imaging and Communications in Medicine (DICOM)

Digital Imaging and Communications in Medicine (DICOM) is the international standard for the communication, management and storage of medical images and related data [38]. DICOM encompasses both the format of the medical images themselves and the communication protocols used to exchange these images and related information between different systems. DICOM includes a full set of metadata that accompanies each image, providing details about the patient, imaging device, acquisition settings, image orientation, etc. This metadata is required for accurate interpretation and diagnosis in the clinical setting as well as in the research setting for image processing. Additionally, DICOM ensures interoperability between different imaging equipment and systems from various manufacturers. DICOM also covers many security concerns with implemented encryption, authorisation and access control.

2.3 Image Processing and Machine Learning for Medical Computer Vision

Medical computer vision is a large field with a diverse set of techniques and applications that are based on clinical and technical endpoints. The specific task and endpoint will impact the workflow and methods that should be implemented. One of the main characterisations of computer vision techniques is whether a technique is learning-based or not. Non-learning-based approaches are generally more traditional and use mathematical models or algorithms where data is only required to validate the performance of the model. Learning-based techniques are usually more computationally intensive and require a separate dataset to train a model for a specific task. The choice of whether to use learning or traditional techniques often comes down to the size of the dataset available as learning-based techniques perform poorly when the training dataset is small.

Applications of medical computer vision include image enhancement, segmentation, classification and registration. Image enhancement aims to improve the quality of an image through various methods such as reducing noise or removing artefacts. Segmentation aims to automatically delineate particular structures of regions of anatomy in a medical image. Classification tasks aim to diagnose or predict clinical outcomes of patients from their medical images. Registration tasks aim to spatially align medical images from different modalities or taken at different times so that the information in each image can be viewed in the same coordinate system.

Many medical computer vision tasks can be considered to have two stages, which are feature extraction and model development. Feature extraction describes the process of converting an image or a region of interest (RoI) within an image into a set of numeric features that describe various qualities of the image. The extracted features detail information about the image's texture, colour distribution, edges, and other distinct attributes that contribute to the images visual content. The aim of feature extraction is to find the most relevant information from the original image and represent that information in a lower dimensionality space [39]. The process of extracting these features depends on the specific requirements of the task and involves selecting relevant techniques for feature detection. Once these numeric features are derived, they provide a structured representation of the visual data that machine learning algorithms and computational models can use to make informed decisions. This transformation from

raw visual input to a meaningful feature representation allows computer vision systems to interpret images in a way that aligns with the desired task. There are many techniques for feature extraction from images such as image texture analysis or the use of convolutional neural networks (CNNs).

Once an image has been converted to a format that is interpretable by computational methods, the model or algorithm for a specific task can be applied. There are many choices of methods available, with deep learning techniques often achieving state-of-the-art performance, especially when there is a sufficient amount of data available. Deep learning models usually combine the feature extraction and classification tasks in a single model. The choice of method depends on factors such as the size of the dataset, the computational resources available and the specific requirements of the medical application. In addition to deep learning, traditional non-learning image processing techniques continue to be valuable tools in medical computer vision. These non-learning methods are particularly useful when data availability is limited or when the interpretability and explainability of the model's decisions are crucial.

This section aims to give an overview of the image processing and machine learning techniques that are relevant to this thesis.

2.3.1 General Image Processing Techniques

There are several general image processing techniques used in this thesis that do not fit well in any other section, these are discussed here.

2.3.1.1 Image Interpolation

Image interpolation is a technique used in image processing to estimate pixel values at geometrical locations that do not lie on the grid of the original image. This is done to increase the resolution of an image, often for the purpose of matching the resolution to the resolution of another image. Image interpolation involves generating new pixel values based on the existing pixel values in the image. The main application of interpolation is to resize or transform an image. The accuracy of the interpolation depends on the quality of the original image and the length of consecutive missing data samples. There are several approaches to image interpolation, some of which are detailed below and also included in the review paper [40].

Nearest Neighbour Interpolation Nearest neighbour interpolation is the simplest form of image interpolation. The pixel intensity value for the target pixel is taken as the intensity value of the closest pixel of the original image. This leads to an interpolated image that has a blocky, or pixelated, appearance.

Linear Interpolation Linear interpolation methods rely on linear polynomials to calculate the intensity values for new pixels. When applied to a 2D or 3D image, this is referred to as bilinear and trilinear interpolation respectively. Linear interpolation works by first identifying the nearest pixels to the off-grid location of the new pixel to be produced. The distances from neighbouring pixels to the new pixel location are found and weights are calculated based on these distances which determine how much each neighbouring pixel will contribute to the new pixel's intensity value. A weighted mean of the neighbouring pixels is then taken to determine the value of the new pixel. This is repeated at all off-grid points where a new pixel is to be created. This results in a smoother image than nearest neighbour quantisation though this can cause edges to be poorly reconstructed.

Cubic interpolation Bicubic interpolation uses a larger number of neighbouring pixels and more complex cubic polynomials than bilinear interpolation to calculate new pixel values. The nearest 2 pixels in all directions are used to calculate the new pixel value which in 2D is a 4 by 4 grid. Weights are generated from these neighbouring pixels based on the distance of the pixels to the location of the new pixel using cubic polynomials. Several methods are available for the calculation of these weights such as the use of Lagrange polynomials, cubic splines or cubic convolution algorithms [41]. Once the weights are found, a weighted average is again taken. An example of how nearest neighbour, bilinear and bicubic interpolation work is shown in Figure 2.10.

2.3.1.2 Mathematical Morphology

Mathematical morphology is a branch of image processing that deals with the analysis and manipulation of geometric structures in images. Mathematical morphology operations are widely used for tasks like image filtering, noise reduction, feature extraction, and object segmentation [42]. Often these operations will require using a structuring element which is a small binary or grayscale pattern that defines the neighbourhood around a pixel, this is essentially a shape that can be used to interact with an image.

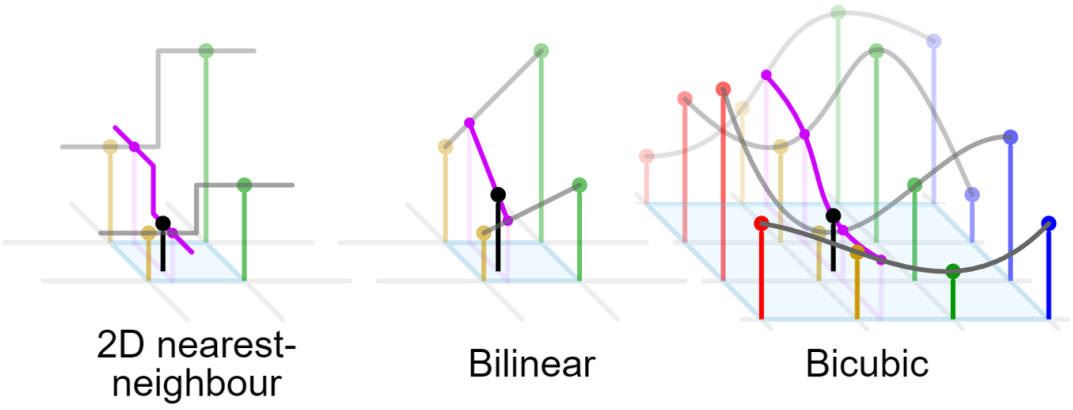


Figure 2.10: *Diagram detailing the process of nearest-neighbour, bilinear and bicubic interpolation. This figure is taken from Wikipedia (https://en.wikipedia.org/wiki/Bilinear_interpolation).*

The choice of structuring element depends on the task but a common choice would be a disk.

Some of the main operations in mathematical morphology for binary image processing are the erode, dilate, close and open functions. The erode function shrinks regions in an image. It works by placing a structuring element at each pixel and checking if all corresponding pixels are positive in the image. If they are, the central pixel remains positive otherwise, it is set to zero. This works to remove pixels in an object at its boundaries. A dilate operation works entirely opposite to the erode function and will set all pixels to be active if a single pixel within the structuring element is positive, expanding any structures from their borders. The opening operation is a sequence of erosion followed by dilation and it is used for tasks like noise reduction and removing small objects or structures. Closing is a sequence of dilation followed by erosion and is used for filling small gaps and connecting broken structures. An example of an erosion followed by a dilation operation applied to a binary 2D image is shown in Figure 2.11.

2.3.1.3 Thresholding

Thresholding is the simplest method of segmenting an image. The most basic form of thresholding creates a binary mask over an image where a mask pixel will equal 1 if the intensity of the pixel in the original image is above a fixed threshold and 0 if it is below. This can be reversed to threshold below a fixed value. For CT images, as the voxel intensity corresponds to different tissue types thresholding can be a useful segmentation technique. In PET imaging, thresholding is often used to isolate a tumour which can

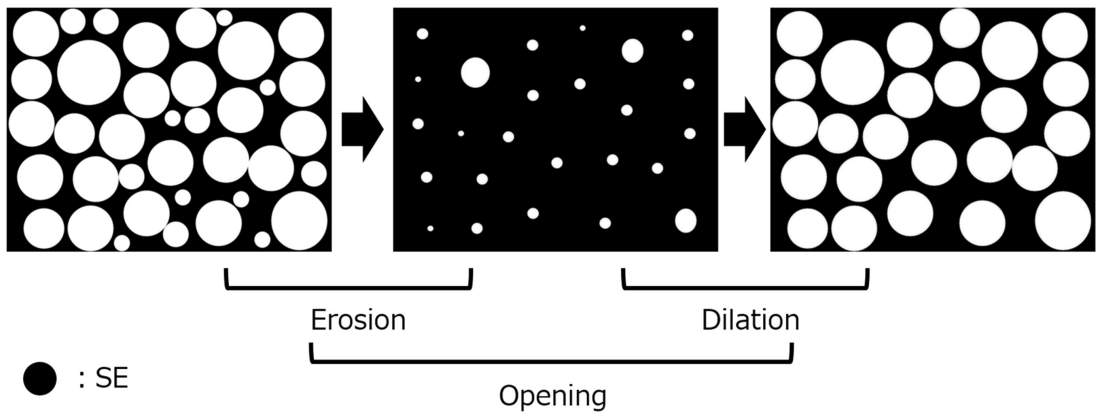


Figure 2.11: *Example of an erosion and dilation of a binary 2D image using a circular structuring element (SE). Figure is taken from [43]*

be delineated by selecting a percentage of the maximum tumour SUV to threshold [44]. In addition to these global methods of thresholding where the thresholding value stays constant over the full image, there are many methods of adaptive thresholding where the threshold value at a particular location is informed by image features [45].

2.3.1.4 K-Means Clustering

K-means clustering is an unsupervised method used to segment an area of interest in an image [46]. The standard algorithm (naive K-means clustering) works by initially selecting the number of clusters, K, for the image to be segmented into, this is a user-defined variable. Each cluster will represent a distinct segment. Initial centroids for these clusters are then randomly chosen across the image. Next, each pixel in the image is assigned to the nearest cluster centroid based on a distance metric, usually the Euclidean distance in the colour space. Pixels that are closer in colour to a particular centroid are grouped into the same cluster. The centroid for each cluster is then re-selected as the centre of mass of each cluster. The pixels are then re-assigned to the new cluster centroids and the process iterates until it converges on a solution. K-means clustering can be performed on greyscale or colour images. In addition to naive k-means, there are several adaptations to the algorithm such as Fuzzy C-Means Clustering [47] and K-means++ [48].

2.3.2 Radiomics (Image Texture Analysis)

Radiomics involves the extraction of quantitative features from medical images using data characterisation algorithms. These features can then be used to develop predictive models for diagnosis, prognosis, and treatment. In recent years, the field of radiomics has received increasing attention for medical image research. This has been partially prompted by the development of software packages such as PyRadiomic [49], LifeX [50] and MATLAB [51] (as of the R2023b update) that improve the ease of the calculation of IBSI radiomic features and the application of full radiomics workflows. Much of the more recent literature aims to use radiomics for computer-aided diagnosis as well as using it as a tool for predicting patient outcomes.

A large portion of radiomic features are image texture features. Image texture is commonly defined as the spatial variation of pixel intensities in an image [52]. Image texture can be used to segment or classify different regions of an image, an example of this is shown in Figure 2.12.

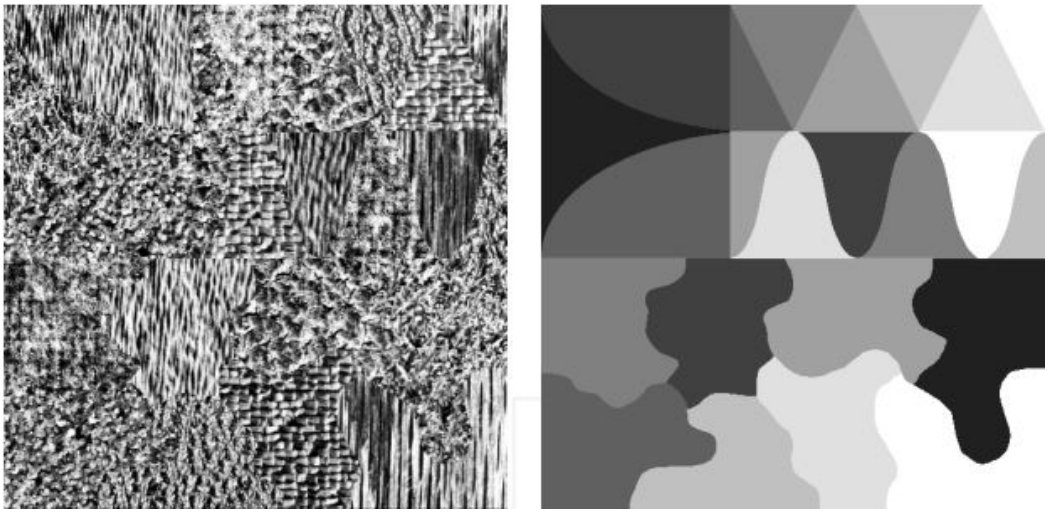


Figure 2.12: *Example of how image texture can be used to segment an image. The left image is a mosaic of eight different image textures. The right image is a grey-level texture map detailing an ideal segmentation of the texture image. Figure taken from [52].*

Image texture analysis can be applied in either 2D or 3D with the methods and measures being extended to accommodate the extra directions available in 3D. In 2D there are 8 pixels neighbouring any pixel, and four direction vectors with a Chebyshev distance [53] of 1, these direction vectors are (1,0), (0,1), (1,1) and (-1,1). In 3D there are 26 neighbouring voxels to any voxel, this gives 13 direction vectors with a Chebyshev

distance of 1, these are; (1,0,0), (0,1,0), (0,0,1), (1,1,0), (0,1,1), (1,0,1), (1,-1,0), (0,1,-1), (1,0,-1), (1,1,1), (1,1,-1), (1,-1,1) and (1,-1,-1) [54]. Variations of image textures may be imperceivable to human vision but are still easily detected by image texture analysis methods. This means that for certain textures, image processing methods may be able to highlight image features that would be missed by a simple visual inspection of the image. For medical imaging, this has applications in discerning changes in anatomy or types of tissue in an image that would not normally be perceivable.

Radiomic features have been standardised by the collaborative initiative "the image biomarker standardisation initiative" (IBSI). Details of most image textures, including those discussed in the following subsections, can be found in the IBSI manual [54] which is currently the best source for detailed information on radiomic features.

2.3.2.1 General Radiomic Workflow

The workflow for radiomic analysis of medical imaging often follows a similar workflow. The main aspects of this are summarised below.

1. **Image pre-processing** - The first stage of a radiomic workflow is any image pre-processing that is to be applied. There is often nothing to be done at this stage but, depending on the task, it may be beneficial to apply denoising or artefact removal algorithms.

2. **RoI segmentation** - The RoI that is to be analysed must be segmented so that the radiomic methods can be applied to only the pixels of that region. In some situations, segmentations may be available from the clinical process such as the delineations of organs at risk produced during radiotherapy planning, that may be used.

3. **Image interpolation** - There is no standard imaging resolution in most medical imaging modalities including CT, MRI and PET. Therefore, to make sure radiomic features are equivalent over a dataset that is being analysed, image interpolation must be applied to match the resolutions of all images that are being analysed. This can involve both downsampling and upsampling of images. Additionally, it is important to match image resolutions in all planes of an image so that image texture features calculated in different directions are directly comparable. For 3D imaging modalities, it is generally the z-axis (head to toe) that

will have a lower resolution. A downside of interpolation here is that the image resolution will be reduced in some directions. It is therefore sometimes more beneficial to keep the higher resolution and perform a 2D analysis instead of a 3D analysis.

4. **Image quantisation** - Before radiomic features are calculated from a RoI, image quantisation is usually applied to reduce the number of grey levels in the image. This is necessary to reduce the dimensionality of the feature space.

5. **Radiomic Feature Calculation** - Finally, radiomic features are calculated. There are many methods for this with the relevant methods to this thesis summarised in the following sections.

For a visual representation of this, the IBSI recommended flowchart for radiomic-based projects is shown in Figure 2.13. Part of the purpose of this pipeline is to increase the reproducibility of any results [55].

2.3.2.2 Statistical Features

Statistical texture analysis methods use the spatial distribution of pixel values to find a set of statistics from the pixel locations and values. Statistical approaches are defined as first-order if a single pixel is used to define a local feature, second-order if two pixels are used and so on [56]. The most common texture methods are:

- First-Order Texture Features
- Grey Level Co-occurrence Matrix
- Grey Level Run Length Matrix
- Grey Level Size Zone Matrix
- The Neighbourhood Grey Tone Difference Matrix

These are described in the following paragraphs.

First-Order Texture Features The difference between first-order and higher order statistics is that first-order statistics ignore the spatial interaction between pixels. This means they can be calculated from the grey-level histogram of a region of interest.

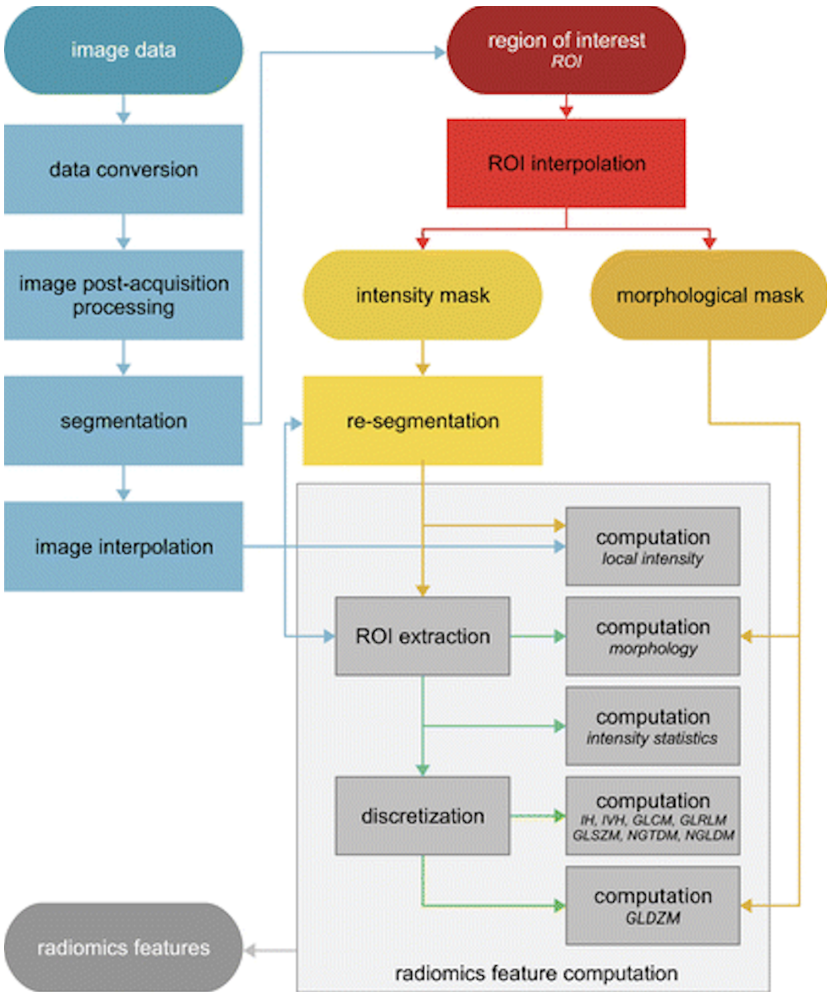


Figure 2.13: *IBSI Standard workflow for radiomic projects, figure taken from [54]*

The most common first-order texture measures are intensity-based statistical features, examples include the mean, variance, entropy, energy, skewness, coarseness and kurtosis of an image [52]. The advantage of first-order measures is their simplicity allowing for faster computational times as well as an increased ease in interpreting any results.

Grey Level Co-occurrence Matrix There are many methods for statistical image texture analysis that involve the conversion of an image into a matrix that highlights certain aspects of the image texture. From these matrices, statistical features can be calculated that describe specific texture qualities. The grey level co-occurrence matrix (GLCM) is one of these matrix based methods. It is a second-order texture analysis method meaning that it relies on immediately adjacent voxels to calculate texture information. A GLCM shows how often each grey level occurs at a fixed distance from

a voxel. Element (i, j) in a GLCM describes how often voxel grey-level value i is located immediately before voxel grey-level value j in the discretised original ROI. GLCMs have to be calculated for each individual direction vectors, giving 13 separate GLCMs. To combine the GLCMs they are commonly averaged to produce a single GLCM.

Once a GLCM has been ccreated it can be used to calculate image features. Harelick et al. proposed fourteen different texture measures based upon GLCM that are commonly used today, these features are known as the Haralick texture features [57]. These features include contrast, correlation and angular second momentum among others.

Grey Level Run Length Matrix The grey-level run length matrix (GLRLM) is a common higher-order texture analysis method first developed by M. Galloway [58]. A GLRLM contains information corresponding to the number of consecutive voxels along a direction vector that are a particular grey level. Element (i, j) in a GLRLM corresponds to the number of times a run length, j, occurs for a certain grey level, i, in a particular direction vector. Just like the GLCM, a separate GLRLM is produced for all 13 vector directions in 3D which can be combined by averaging the values. GLRLM can differentiate between course and fine textures as course textures will have longer run lengths for any grey level than fine textures. Various metrics can be produced from GLRMs such as the short-run emphasis, long-run emphasis and grey-level distribution [52].

Grey Level Size Zone Matrix The grey-level size zone matrix (GLSZM) is a high-order texture analysis method similar in function to GLRLM. A GLSZM counts the number of linked voxels that have the same grey level. For a voxel to be linked to another voxel it must have the same grey level and be one of the 26 neighbouring voxels (in 3D). Element (i, j) of a GLSZM corresponds to the number of zones with a grey level of i and size of j. Unlike the GLCM and GLRLM, this technique will produce only one GLSZM from a 3D ROI so no averaging has to be performed. Metrics calculated from the GLSZM include small zone emphasis, large zone emphasis and grey level non-uniformity.

The Neighbourhood Grey Tone Difference Matrix The neighbourhood grey tone difference matrix (NGTDM) is a grey level based texture matrix that quantifies the relationship between the grey level of a voxel and the difference between that voxel

and its neighbouring voxels. A NGTDM has more parameters than some other grey level based matrices. A NGTDM will consist of three variables per grey level, i. These variables are the total number of grey levels in the ROI, n_i, the probability of a voxel having grey-level i, p_i and the neighbourhood greytone difference, s_i. s_i is defined by equation 2.3 where $\bar{X}_k$ is the average grey level in the neighbouring voxel of a particular voxel and N_v is the number of voxels in the ROI. n_i, p_i and s_i are used to calculate various image metrics such as busyness, contrast and coarseness.

$$s_i = \sum_{k}^{N_v} \left| i - \bar{X}_k \right| \tag{2.3}$$

2.3.2.3 Deep Learning in the Context of Texture Analysis

Deep learning methods, for computer vision, may also be considered to be performing a form of image texture analysis. Deep-learning based methods work by learning features from a dataset used to train the layers of a network. Some of the features learned by the deep-network are analogous to classic image texture features as they will be detecting changes in textures in regions of the image. Unlike classic texture analysis, these features cannot be easily represented mathematically and are unique to a network that has been trained on a particular dataset.

2.3.3 Convolutional Neural Networks (CNN)

Increases in computational power in the past 15 years have allowed for the application of deep learning methods for image feature extraction. These techniques have revolutionised image feature extraction by leveraging complex neural network architectures, such as convolutional neural networks (CNNs). CNNs can automatically learn intricate features from raw image data. As opposed to many traditional methods for feature extraction, CNNs must learn image features from a dataset using a training-based approach. This allows for the CNN to learn abstracted features. The successful application of AlexNet [59] in 2012 to the ImageNet [60] classification challenge saw the start of a shift towards the use of deep learning and CNNs as the standard method for image processing applications.

A Convolutional Neural Network is structured with multiple trainable stages that are stacked on top of each other. These successive stages are subsequently followed by a supervised classifier. Throughout the network's architecture, feature maps are

utilised as arrays to depict the input and output at each stage of computation. This arrangement enables a CNN to progressively learn hierarchical features from raw data, making it a powerful tool for tasks such as image recognition, object detection, and various other forms of pattern analysis. CNNs are specifically designed to process and analyse images (or other grid-like data). The information in the following sections on CNNs can be found in greater detail in the books [61] and [62]. A CNN consists of multiple layers that are detailed in this subsection, these are:

- Convolutional Layer
- Pooling Layer
- Fully Connected (Dense) Layer
- Activation Function
- Dropout

Convolutional Layer The convolutional layer is the main feature of CNNs that make them well-adapted to process images. A convolution involves sliding kernels across the input data, the weights of these kernels are learned during training, allowing the network to learn features that are relevant to a particular task. The output of the convolution operation is a set of feature maps representing the response of a particular kernel across the input. CNNs stack convolutional layers so that a subsequent convolutional layer is applied to the feature map output of a previous layer allowing for more abstracted and detailed features to be learnt as the depth of the network increases. This causes earlier convolutional layers in a CNN to capture simple features like edges while deeper layers will capture more complex features. There are several parameters that can be changed in a convolutional layer depending on the application with the kernel size, stride and padding being commonly applied parameters. The kernel size determines the dimensions of the receptive field used to extract features from the input image (or feature map). The stride of a convolutional layer dictates how the kernel moves during convolution by making the kernel move in steps equal to the stride, essentially skipping pixels. Padding can be added to control the size of the feature maps by adding pixels to the edge of an image allowing for the kernel to be applied to the edge pixels which allows the input dimensions to be maintained after convolutions.

Pooling Layer A pooling layer is a component used in CNNs to reduce the spatial dimensions of feature maps while retaining important information. Pooling layers help simplify the networks computations, reduce overfitting, and improve translation invariance. A pooling layer performs spatial downsampling by partitioning the input feature map regions (pools) and selecting a representative value from each pool. Max pooling and average pooling are two popular choices for this. As pooling downsamples the spatial dimensions, successive convolutional layers essentially gain an increased receptive field as the convolution kernels will cover a larger area in the original image dimensions. A global average pooling layer, that averages feature maps over the entirety of their spatial dimensions, is often applied for classification problems to generate global features that can be used for classification.

Fully Connected (Dense) Layer Fully connected layers connect each neuron from the previous layer to every neuron in the current layer so that each neuron in a dense layer receives input from all the neurons in the preceding layer. To process the information from a feature map to a fully connected layer, the feature map must first be flattened to a one dimensional vector. The purpose of a fully connected layer is to allow the network to learn relationships between all of the features. In an artificial neural network (ANN), also known as a multilayer perceptron (MLP), fully connected layers are the main layers used to process the features. In a CNN for a classification or segmentation task, fully connected layers are generally only used prior to the output layer to establish relationships between all of the features to the output classes. When used as the output layer of a neural network, a fully connected layer will have the same number of neurons as the number of classes in a classification problem.

Activation Function An activation function introduces non-linearity to a neural network, enabling it to capture complex relationships between inputs and outputs. Additionally, activation functions limit the output size of each neuron which is beneficial especially when dealing with vanishing gradients. In a CNN, activation functions are applied element-wise to the output of each neuron in the convolutional layer, shaping the network's ability to model intricate patterns and features. It is desirable for an activation function to be zero-centred, computationally cheap and it must be differentiable.

The most used activation function is the rectified linear unit (ReLU) activation

function which is defined by equation 2.4 [63]. This function is also displayed in Figure 2.4.

$$ReLU(x) = \max(0, x) = \frac{x + |x|}{2} = \begin{cases} x & \text{if } x > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{2.4}$$

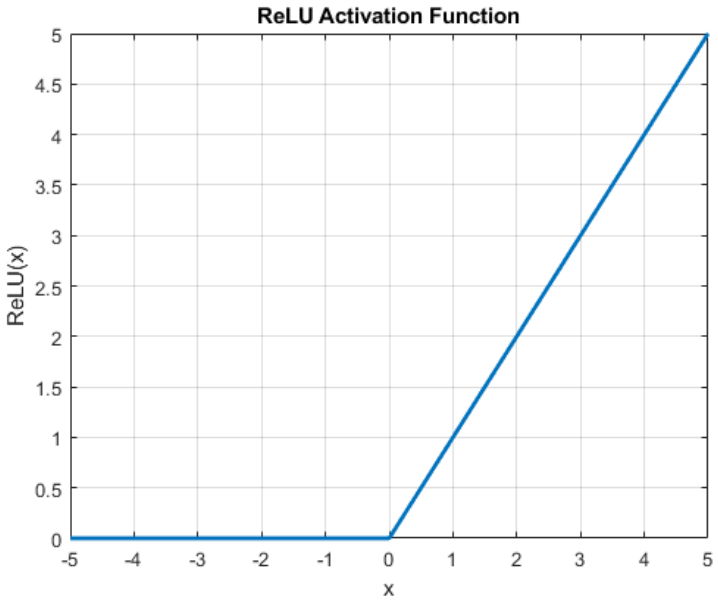


Figure 2.14: *The ReLU activation function*

In addition to the ReLu function, there are several other activation functions available such as the sigmoid, hyperbolic tangent, leaky ReLU and softmax functions. The softmax activation function has a particular use in the output layers of a CNN as it is used to convert a vector of numbers into a probability distribution which can be used as a classification with the input producing the highest probability being the defined class. The softmax function is defined in equation 2.5.

$$S(\mathbf{x})_i = \frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_j}} \tag{2.5}$$

Dropout A dropout layer is a regularisation technique used in neural networks, the purpose of which is to prevent overfitting by randomly deactivating a portion of neurons during each training iteration. This stops the neural network from relying too heavily on a single feature for whatever task it is being applied to which reduces overfitting and makes the network more robust. In a CNN, a dropout layer will be applied after

most convolutional and fully connected layers [64].

2.3.3.1 CNN Training

To train a CNN to function for a specific task, a training scheme will have to be used. This usually involves splitting the data into a training, validation and test set. The training set is used to train the model and is the data from which the CNN will learn features. The validation set is used to select the point in training that the CNN is performing best and to tune the hyperparameters of the model. Using a validation dataset is an optional step in the training process though if the dataset is large enough, it will always provide a benefit to the model. If the dataset is not large enough, a validation dataset may require a reduction in the test and training datasets that is too large to be worthwhile. Additionally, a small validation dataset may bias any hyperparameter tuning. The test dataset is a dataset that is held back during the training process and is used to blindly test the performance of the final model. It is important here that the model has not been trained on any of the testing data as this would bias the final results. It is also important in the medical imaging domain to hold back entire patients in the test dataset so as to not bias the results. The main features of CNN training are detailed in the following paragraphs, these are:

- Loss function
- Training Loop
- Optimiser
- Data augmentation
- Hyperparameter tuning

Loss Function The loss function is a metric that is calculated after every mini-batch, which is a subset of the training data, to determine the performance of the network. The goal of the network training process is to minimise (or, depending on the loss function, maximise) the loss function by learning the best features and relationships between the features from the test dataset. The choice of loss function depends on the task as tasks such as classification, regression or segmentation will require different loss functions. For classification tasks, the most common loss functions are based on

cross entropy. There are several variations of cross entropy based loss functions but the weighted cross entropy (WCE) loss function is the most popular. This adds a weighting to the loss from different classes and is usually applied to unbalanced datasets so that the network is not biased towards correctly predicting cases of the most sampled class. The WCE loss is described by equation 2.6.

$$WCE = -\frac{1}{N}\sum_{n=1}^{N} w_i \ln \hat{y}_{ni} \tag{2.6}$$

Here, N is the mini-batch size w_i is the weighting applied to cases of class i and $\hat{y}_{ni}$ is the probability that the network associates the n^{th} input sample with class i.

For regression based problems, where the output is a continuous variable, the mean square error (MSE) is the most common loss function which is described in equation 2.7 where y_n is the ground truth and $\hat{y}_n$ is the predicted output.

$$\text{MSE} = \frac{1}{N}\sum_{n=1}^{N} (y_n - \hat{y}_n)^2 \tag{2.7}$$

For segmentation problems, cross entropy can be applied as a loss function on a pixel-wise basis or metrics such as the DICE metric [65] can be applied which is described by equation 2.8.

$$Dice = \sum_{n=1}^{N}\left(1 - \frac{\sum_{p=1}^{P} 2y_p\hat{y}_p}{\sum_{p=1}^{P} y_p + \sum_{p=1}^{P} \hat{y}_p}\right) \tag{2.8}$$

Here y_p is the ground truth pixel label and $\hat{y}_p$ is the predicted pixel label where p is a specific pixel. For further reading, a review of many common loss functions used in deep learning applications is available in [66].

Training Loop The process of training a CNN involves looping through the training data over many iterations to minimise a loss function. To apply this training loop, the training data is usually split into mini-batches. A training iteration begins with a forward propagation where a mini-batch is passed through the network to calculate predictions. The loss is then computed by comparing the predicted values to the ground truth labels using the chosen loss function. Backpropagation is then performed where the gradients of the loss with respect to the model's parameters are calculated. Finally, the model's parameters are updated using an optimiser to minimise the loss. This process then iterates on a new mini-batch. Once all the mini-batches have been

processed an epoch has been completed. The process generally then repeats multiple times with training schemes sometimes reaching over 1000 epochs. At set multiples of epochs, certain training parameters such as the learning rate may be updated which can provide a more efficient learning process. The model may also be tested over the validation dataset to monitor its performance performance at set intervals. This can be an important step as models can overfit to the training data if trained for too many epochs. This means the model will learn features specific to examples in the training set that do not generalise well causing a reduction in performance on unseen data. The version of the model with the best performance on the validation dataset during training is generally used as the final model.

Optimiser An optimiser is an algorithm that adjusts a model's parameters during training to minimise the loss function. Different optimisers use various strategies to update the model's parameters, and the choice of optimiser can impact the convergence speed and final performance of the model. Optimisers work by calculating gradients of the loss surface and adjusting the parameters to move in the direction of the steepest gradient. This means that all optimisers are gradient descent algorithms. One of the main hyperparameters of any optimiser is the learning rate. This determines the step size taken along the gradient direction in each iteration. A higher learning rate can lead to faster convergence but risks overshooting the minimum. A lower learning rate can lead to more stable convergence but slower progress.

Stochastic gradient descent (SGD) is one of the most common optimisation algorithms in deep learning. The general gradient descent algorithm is summarised by equation 2.9 [67].

$$w_{t+1} = w_t - \eta \cdot \nabla_{w_t} L(w_t) \tag{2.9}$$

Where w is the model parameters, t is the iteration step, η is the learning rate, L is the loss function and $\nabla_{w_t} L(w_t)$ is the gradient of the loss function with respect to the model parameters. Here the parameters are shifted in the direction of fastest decrease in the loss function. Stochastic gradient descent is the version of the gradient descent algorithm that is applied after every sample of the training process as opposed to after every epoch. This allows for faster convergence times and allows for local minima in the loss function to be escaped more easily by the algorithm. Additionally, a momentum

term can be applied to the SGD model to improve the speed of convergence.

The adaptive moment estimation (Adam) optimiser is another popular optimisation algorithm used to train machine learning models. It combines an adaptive learning rate method with momentum to achieve efficient updates to the model parameters during training. The Adam optimiser adapts the learning rate for each parameter based on the first and second moments of the gradients. This means that the learning rate for each parameter can be different. The first and second moments of the gradient are m_w and v_w as defined by equations 2.10 and 2.11 [68].

$$m_w^{(t+1)} = \beta_1 m_w^{(t)} + (1 - \beta_1) \nabla_w L^{(t)} \tag{2.10}$$

$$v_w^{(t+1)} = \beta_2 v_w^{(t)} + (1 - \beta_2) \left(\nabla_w L^{(t)} \right)^2 \tag{2.11}$$

β_1 and β_2 are the exponential decay rates for the moment estimates with initial values of 0.9 and 0.999 respectively being reported to produce good results in the original Adam paper [68]. To stop m_w and v_w from being biased towards 0 as β_1 and β_2 are close to 1, the biased corrected values $\hat{m}_w$ and $\hat{v}_w$ can be calculated by the equations below.

$$\hat{m}_w = \frac{m_w^{(t+1)}}{1 - \beta_1^t} \tag{2.12}$$

$$\hat{v}_w = \frac{v_w^{(t+1)}}{1 - \beta_2^t} \tag{2.13}$$

From these, the parameter update calculation is

$$w^{(t+1)} = w^{(t)} - \eta \frac{\hat{m}_w}{\sqrt{\hat{v}_w} + \epsilon} \tag{2.14}$$

Here ϵ is a small scalar used to prevent division by zero, suggested to be 10^{-8} in the original Adam paper [68].

Data Augmentation Data augmentation is any method used to increase the effectiveness of model training by manipulating the data before presenting it to the model. The aim of this is usually to reduce overfitting and increase the robustness of the model by artificially increasing the diversity of the training data. For CNNs, data augmenta-

tion is a necessary step in any training pipeline. The most well-used data augmentation techniques involve simple transformations to the training images. Images can be randomly rotated, translated and flipped so that the CNN will learn features at all angles and image locations. This is important as CNNs are not rotationally invariant so a feature that is learned from a training image will not be recognised if it is rotated by 90 degrees (assuming there are no other examples in the training dataset). Additional image augmentation techniques include cropping, resizing, adding random Gaussian noise, applying an elastic transform, brightness and contrast shifts. An example of some of these techniques is shown in Figure 2.15.

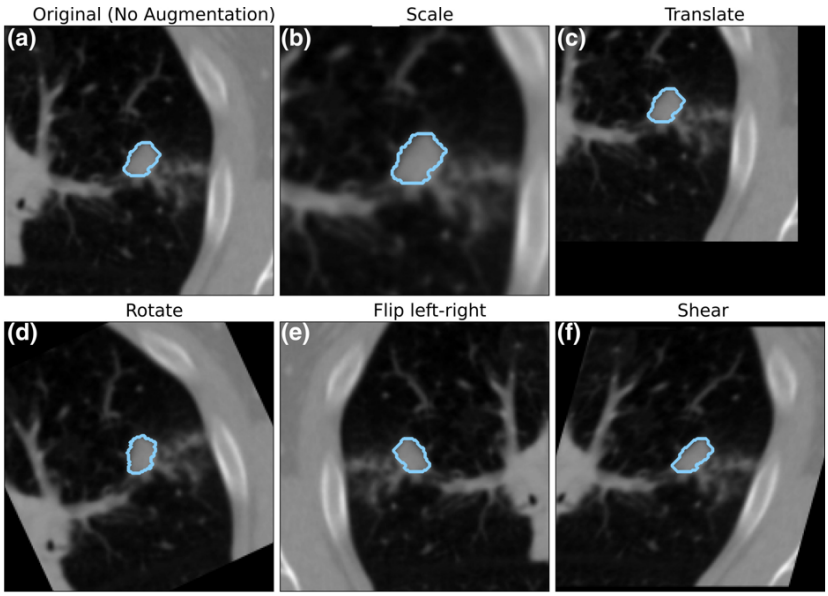


Figure 2.15: *Examples of some different image augmentation techniques applied to a CT image of a lung tumour. Image is taken from [69].*

The choice of augmentation techniques applied is dependent on the dataset and the task as some augmentation techniques may be detrimental in some situations. For example, for a task based on the CT or PET modalities it would generally be detrimental to apply random brightness (or intensity) shifts to the images as the voxel intensity values have been calibrated and are expressed in known units. These simple image augmentation techniques are usually applied to the data after every epoch of training. In addition to these simple techniques, image generation is currently a large field in data augmentation with techniques such as generative adversarial networks being used to generate synthetic images that can be then used to train a CNN. For further detail on image augmentation techniques for deep learning applications the

2023 survey on image augmentation [70] and the 2021 review of image augmentation for medical images [69] are valuable sources of information.

Hyperparameter Tuning Hyperparameters are parameters that are set before the training process begins. These parameters can have a large impact on the model training and final performance. Examples of common hyperparameters to be tuned are the learning rate, batch size, number of layers, dropout rate and the optimiser type. To find the best hyperparameters for a given task, hyperparameter tuning can be performed. This involves training the model many times with different hyperparameters and selecting the hyperparameters that produce the best model performance.

The first step of hyperparameter tuning is to define the search space by choosing which hyperparameters to optimise and to define the range of values that these hyperparameters can take. A validation dataset then has to be created which will be used to calculate the model performance. A hyperpameter search method must then be used to test different hyperparameter combinations. A grid search exhaustively tries all combinations of hyperparameters from the search space. This can be computationally expensive especially as the number of hyperparameters increases. A random search randomly samples combinations of hyperparameters from the search space. A Bayesian optimisation approach uses probabilistic models to predict the performance of different hyperparameter values and then selects the next values to try based on these predictions. These searches are an iterative process that will repeat until a stopping criterion is met. Once the optimal set of hyperparameters are selected, the final model can be trained using those hyperparameters and tested using the test dataset that the training or hyperparameter tuning process has not seen. For further reading a detailed review of hyperparameter tuning can be found in [71].

k-fold cross-validation can be applied during the hyperparameter tuning process. The training data is split into k-folds where one of the folds will be used as the validation dataset and the rest of the folds are used to train the model. A set of optimal hyperparameters are found using the validation fold and the process is repeated while taking a different fold as the validation dataset. This means the entire training dataset can be used as the validation dataset, increasing the size of the validation dataset and reducing the overfitting of the hyperparameters.

2.3.3.2 Specific CNN Architectures

There are many specific CNN model architectures that have been designed to improve performance at specific tasks. The model architectures used in this thesis are detailed in this subsection, these are:

- ResNet
- UNet
- DeepLab

ResNet ResNet (Residual Network) is a CNN architecture first introduced by Ka. H, et al in their 2015 paper "Deep Residual Learning for Image Recognition" [72]. This paper introduced residual blocks, which use skip connections to bypasses one or more layers in the network. The skip connection in a residual block works by simply passing the input directly to the output, this is often referred to as identity mapping. The layout of a residual block is shown in Figure 2.16. ResNet models stack many residual blocks to create deep neural networks with many skip connections. The number of residual blocks in the model is usually included in the model naming scheme with ResNet-18 and ResNet-50 having 18 and 50 residual blocks respectively.

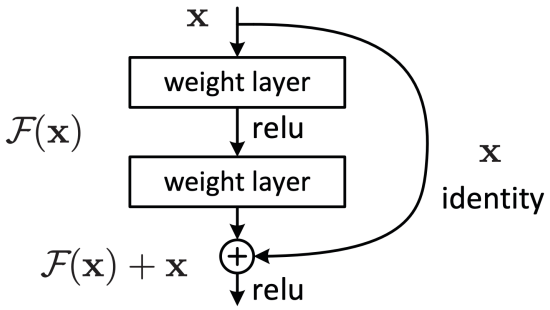


Figure 2.16: *The residual block. Image taken from [72].*

The purpose of the skip connection is to reduce the effects of vanishing gradients during training. Vanishing gradients occur when the derivative of the loss function with respect to the model's parameters becomes extremely small as they are propagated backwards through the layers of the network. This slows down training and leads to sub-optimal model performance which becomes more of an issue as network depth increases. The residual blocks introduced in ResNet allow for the gradient to pass directly through the network without diminishing.

UNet UNet is a CNN architecture designed for semantic segmentation. It was originally introduced by Olaf R, et al, in their 2015 paper "U-Net: Convolutional Networks for Biomedical Image Segmentation" [73]. Since then, UNet has become one of the most popular CNN architectures for image segmentation, especially in the biomedical image domain. UNet uses an encoder-decoder architecture where the encoder compresses the input image into a lower dimensional feature representation and the decoder expands this feature representation to produce a pixel-wise segmentation map. The encoder consists of multiple convolutional and pooling layers that progressively reduce the spatial dimensions of the input image while increasing the number of feature channels. The encoder is the part of the network that extracts features from the input image. The decoder uses transposed convolutional layers to upsample the feature maps to their original resolution so that pixel-wise features are available to perform a pixel-wise classification. UNet additionally uses skip connections between corresponding layers of the encoder and decoder. This is done to conserve the features from the earlier layers of the encoder, allowing them to be used in the final segmentation. An example UNet architecture with an encoder depth of 4 is shown in Figure 2.17.

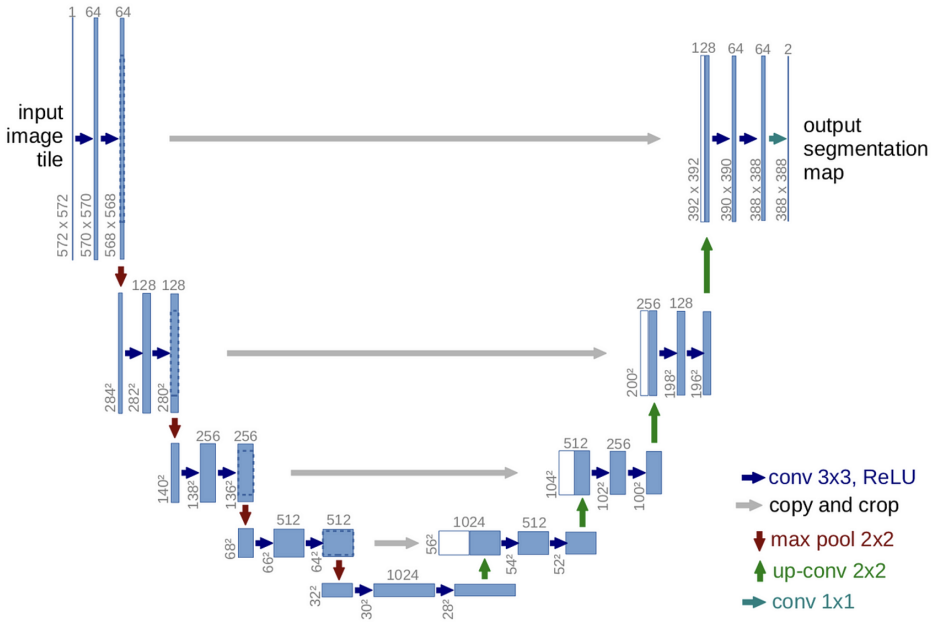


Figure 2.17: *The UNet architecture with an encoder depth of 4. Image taken from* `https://towardsdatascience.com/unet-line-by-line-explanation-9b191c76baf5`.

DeepLab DeepLab is a CNN used for semantic image segmentation of which there have been several versions that have added features to improve the performance of the network. The first version of DeepLab introduced Conditional Random Fields or CRFs for segmentation CNN to include the relationships between nearby pixels in the segmentation [74]. DeepLabV2 introduced atrous convolutions (dilated convolution) for semantic segmentation [75]. Atrous convolutions include gaps in the filters so that they skip over pixels. The size of the gaps is known as the dilation rate. This allows the network to capture information from a larger receptive field without increasing the number of parameters. An example of a dilated convolution compared to a standard convolution can be seen in Figure 2.18.

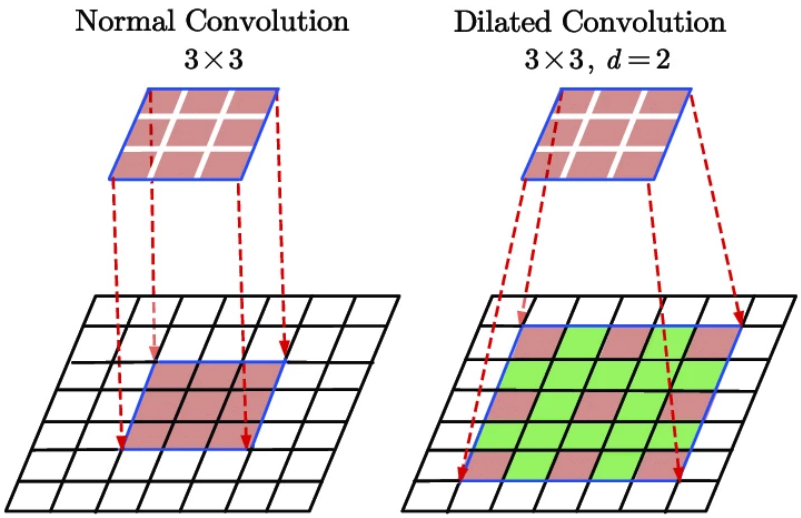


Figure 2.18: *Example of a normal and dilated (atrous) convolution kernal. Image is taken from Hasty.ai (`https://hasty.ai/docs/mp-wiki/model-architectures/deeplabv3`).*

DeepLabV3 advanced this by applying cascades of atrous convolutions to progressively extract features from an image [76]. Additionally, atrous spatial pyramid pooling (ASPP) was used where multiple parallel atrous convolutions with different dilation rates were applied allowing the network to gather information at multiple scales. DeepLabV3+ then improved upon DeepLabV3+ by introducing an encoder-decoder architecture with skip connections, similar to UNet [77].

2.3.4 Decision Trees

Decision trees are a supervised method for classification or regression that can be used in medical image problems. Decision trees take a set of features, which are generally

expressed as numeric values, and apply a set of consecutive rules based on these features in a flowchart-like design to reach a final prediction. A decision tree is constructed of nodes which represent a decision point after which the data is split into subsets. The first node in the decision tree is called the root node, subsequent nodes are called intermediate nodes and the nodes at the end of the tree that provide the final prediction or decision are called leaf nodes. Branches are defined as subsections of the decision tree consisting of multiple nodes. At each internal node, the decision tree selects a feature and a threshold value to split the data into subsets based on the values of that feature in a process known as splitting. The splitting is defined by decision rules that determine which subsequent node to proceed to. At the end of a chain of nodes, the leaf node defines the prediction.

The process of constructing a decision tree starts with a root node that contains all of the training data. The algorithm then recursively selects the best feature and threshold to split the data into subsets at the root and all internal nodes. The tree-growing process stops when certain criteria are met. These can be a predefined maximum depth of the tree, a minimum number of samples at a node or when splitting stops increasing accuracy. During the construction process, it is often beneficial to limit the tree depth as deeper trees are more likely to overfit to the training data. This process is described by the classification and regression trees (CART) algorithm [78]. Adaptations to the tree construction process are available such as the ID3 [79] and CHAID [80] algorithms.

Optimisation Criterion For Decision Trees Ginis impurity is a common optimisation metric for classification tasks that quantifies the likelihood that a randomly selected element from the set would be incorrectly labelled if it were labelled randomly and independently according to the distribution of labels in the set. Ginis impurity index is expressed by equation 2.15.

$$I_{Gini} = 1 - \sum_{i=1}^{J} p_i^2 \tag{2.15}$$

Here I_{Gini} is Ginis impurity, J is the number of classes and p_i is the probability of correctly randomly selecting a sample from class i. Gini impurity is used in decision tree algorithms to evaluate potential splits in the data that will create the nodes of the tree. When a decision tree is constructed, it selects the split that minimises the Gini impurity in the resulting child nodes [78].

For regression trees, the mean square error (MSE) is a commonly used metric for constructing decision trees. To do this, the MSE is calculated for each potential split by comparing the predicted values within each child node to the actual target values. The feature and threshold for a node split are selected by choosing the values that minimise the MSE.

Boosted Decision Trees and AdaBoost The performance of decision trees can often be improved by applying gradient boosting. Gradient boosting is a method in machine learning that creates an ensemble of weak learners to make a final prediction [81]. Here, weak learners are simple models that do not perform substantially better than random chance. For boosted decision trees, these weak learners are trees with a small depth. By including these weak classifiers in the training process and then combining their predictions to make a final prediction, improved accuracy is generally observed over a single complex decision tree.

There are many methods for gradient boosting decision trees. One of the most popular methods is the AdaBoost (Adaptive Boosting) algorithm [82]. AdaBoost is an iterative algorithm that trains a weak classifier every iteration. AdaBoost initialises by assigning uniform weights, $w_i^{(m)}$, to all of the samples. A weak classifier is then created which minimises the error, W_m, defined by equation 2.16.

$$W_m = \sum_{y_i \neq k_m(x_i)} w_i^{(m)} \tag{2.16}$$

Here, m is the iteration number, i is the sample, y_i is the ground truth and k_m is the model output. W_m is therefore the sum of the weights, which are uniform in the first instance, of the cases that were misclassified. The error rate is then the sum of the misclassified sample weights over the sum of all the sample weights which is defined by equation 2.17.

$$\epsilon_m = \frac{\sum_{y_i \neq k_m(x_i)} w_i^{(m)}}{\sum_{i=1}^{N} w_i^{(m)}} \tag{2.17}$$

After each iteration, a weight, α, is assigned to the weak classifier based on how well it performed on the training dataset. ϵ_m is used to calculate α by equation 2.18.

$$\alpha_m = \frac{1}{2} \ln \left(\frac{1 - \epsilon_m}{\epsilon_m} \right) \tag{2.18}$$

The weak learner is then added to the final classifier with its contribution defined by its weighting as defined in equation 2.19 where C_m is the ensemble of weak learners.

$$C_{m-1} \text{ to } C_m = C_{(m-1)} + \alpha_m k_m \tag{2.19}$$

The sample weights are then updated based on whether they were misclassified or not so that the misclassified samples are given more importance in the subsequent iterations. If the sample was misclassified the updated sample weight, $w_i^{(m+1)}$, is

$$w_i^{(m+1)} = w_i^{(m)} \mathrm{e}^{\alpha_m} = w_i^{(m)} \sqrt{\frac{1 - e_m}{e_m}} \tag{2.20}$$

otherwise,

$$w_i^{(m+1)} = w_i^{(m)} \mathrm{e}^{-\alpha_m} = w_i^{(m)} \sqrt{\frac{e_m}{1 - e_m}} \tag{2.21}$$

This whole process is repeated until a maximum user-defined number of iterations is reached at which point the final classifier is taken as a weighted ensemble of the weak learners that were created over the iterative process. While boosted decision trees generally produce more accurate models than non-employee-based methods, they are less explainable as following the path that a single decision tree takes to make its decision is simple and following the paths of many trees is more complex.

2.3.5 Image Registration

Registration of medical images is the alignment of two or more images so that the features of the images are as spatially aligned in the same coordinate system. This is a complex problem with applications in image-guided surgery and disease diagnosis. Registration can be between images from one image modality, taken at different times or positions, or between images from different imaging modalities. Registration may be performed in 2D or 3D depending on the specific application. Registration methods involve taking one image as the target or fixed image and the image, or images, to be aligned with the target as the moving image. The target image is not altered during the registration process and a geometrical transform is applied to the moving image so that it aligns with the target image as closely as possible.

The goal of any registration algorithm or model is to find the transformation function ($\hat{W}$) that optimises the following functional [83]:

$$\hat{W} = \underset{W}{argmax}\mathcal{M}(T, S \circ W) + \mathcal{R}(W) \tag{2.22}$$

Where T is the target image, S is the moving image and W is the transformation applied to S. $\mathcal{M}$ defines the level of alignment between T and S and $\mathcal{R}$ is a regularisation term that enforces any user-defined properties of the solution. From this, it can be seen that a registration algorithm is composed of three main components which are (i) a transformation model, (ii) a similarity measure and (iii) an optimisation method. There are many methods available for solving this function from classic image-processing methods to deep learning based methods. Deep learning registration based methods were not applied in this thesis so they are not summarised here, for further reading on deep learning based medical image registration, [84] and [85] may be referred to.

2.3.5.1 Transformation Model

The transformation model defines the deformation applied to the moving image. The most basic transformations that can be applied are linear transformations such as rotation, scaling and translation. These are global transformations that deform the whole image and cannot resolve local discrepancies between the fixed and moving images. Elastic transformations are defined as transformations that can change the local structure of the moving image. The transformation model dictates the number of parameters that need to be estimated.

A similar distinction in registration algorithms is whether they use a rigid or non-rigid transformation. Rigid registration involves no deformation of the features of the images during the registration process so it is generally best suited to the registration of hard structures such as bones or problems where the original structures contained within the images are required to be unaltered. Rigid registration involves only a rotation and translation of the moving image to fit the fixed image so it can be simply defined by equation 2.23 [86].

$$T = R * S + l \tag{2.23}$$

Here R and l represent the rotation and translation parameters respectively.

Non-rigid registration, on the other hand, allows for the deformation of features making it most suited to the registration of soft tissues, especially in areas where defor-

mations of the tissue are known to take place such as in the lungs due to the breathing cycle. Non-rigid registration methods involve many more parameters than rigid registration. Due to the simplicity of rigid transformation models compared to non-rigid methods, they are often used as a pre-registration step in a non-rigid registration to reduce the computation time of the non-rigid registration.

2.3.5.2 Similarity measures

The similarity measure, or cost function, is a measure used to quantitatively compare the alignment of two images. Similarity measures can be grouped into two categories which are intensity based and feature based methods. Intensity based methods rely upon the pixel intensities of the images to produce a measure of the difference in alignment. Feature based methods rely upon extracting common features between the images and constructing a measure based on the distances between the common features of each image.

Intensity Based Measures The mean squared error (MSE) or the sum of squared differences is one of the most common intensity based measures. Here the average difference in pixel intensity values of the fixed and moving images are found. This is described by equation 2.24.

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(I_{T,i} - I_{S,i})^2 \tag{2.24}$$

Here $I_{T,i}$ and $I_{S,i}$ are the intensity values of the fixed and moving image respectively with pixel number i. A low value of the MSE error corresponds to a well registered image. The assumption with the MSE measure is that corresponding structures in the images should have identical intensities. This means that, for medical images, the MSE measure is usually only suited to mono-modality registration or the registration of binary image masks as the intensity values of structures within both images should be close to matching. Further intensity-based similarity measures include cross-correlation and mutual information [87].

Feature Based Measures Feature based methods of registration aim to minimise the distance between common features of the images to be registered. These features can be defined as points, curves or surfaces. The reduced number of image elements

used in a feature based measure compared to an intensity based measure means the computational complexity is also reduced. In their 2018 review paper, C Y Guan et al. describe in detail many methods of feature based medical image registration [86].

Feature based registration methods have an additional component compared to intensity based methods which is the acquisition of the features. The simplest method of feature acquisition is for clinicians to manually define features. This approach has a good accuracy but the time taken to define the features is too long for most applications so automatic feature acquisition is usually required. Automatic methods for feature selection include the Laplacian of Gaussian (LoG) [88] and the scale invariant feature transform (SIFT) [89] algorithms.

Once features have been extracted for the images to be registered, corresponding features need to be found. This is done by matching features with similar descriptors where the spatial location of the features can also be taken into account to improve the robustness of the feature matching. One of the most common feature matching methods is the iterative closest points (ICP) algorithm, first introduced by Besl and McKay [90]. The similarity measure of the algorithm can be described by equation 2.25 [86].

$$d(R,t) = \frac{1}{N}\sum_{i=1}^{n}(Rt_i + l - s_i)^2 \tag{2.25}$$

Here a rigid registration of the point set of S to the point set of T is computed, with $s_i \in S$ and $t_i \in T$, where R and l are the rotation and translation parameters and d is a Euclidean distance similarity measure to be minimised. Due to the popularity of the ICP algorithm, many adapted versions exist such as the EM-ICP [91] and LM-ICP algorithms [92].

2.3.5.3 Optimisation Algorithms For Image Registration

Once a similarity measure has been defined it is necessary to perform an iterative optimisation of this measure to reach a maximum or minimum so that the images are as closely registered as the algorithm will allow. Optimisation algorithms can be split into two categories which are continuous and discreet [83]. Continuous optimisation algorithms involve real-valued variables with differentiable cost functions and discreet algorithms involve variables that are discreet and have a non-differentiable cost func-

tion.

One of the most widely used continuous optimisation algorithms is the gradient descent method. This algorithm iteratively changes the parameters of the similarity measure to move the measure in the direction opposite to its steepest gradient. This is described by equation 2.26 [67].

$$x_{n+1} = x_n - \gamma \nabla(F(x_n)) \tag{2.26}$$

Here x is a vector of the deformation parameters, n is the iteration number, γ is the step size and ∇ is the first order differential applied to the similarity measure F. The step size determines how far along the direction opposite to the gradient the similarity measure is moved. An appropriate step size should be large enough to reduce computation time by lowering the number of iterations and low enough so that the solution is not overshot which would cause oscillations. The number of iterations is capped, often by defining a minimum distance between steps, as the algorithm will not produce a final value for the similarity measure without this. Variations in applying the step size are available such as reducing the step size with the iteration number to improve the accuracy of the result and reduce computation time [93]. Depending upon the similarity measure and transform model, there may be multiple local minima that the gradient descent algorithm can reach which would cause a non optimal solution to be produced. To solve this, a pre-registration step may be included that roughly registers the images so that when the gradient descent starts, it converges to the correct local minimum. The gradient descent algorithm can become computationally intensive if the parameters to be estimated have high dimensionality. To overcome this problem stochastic gradient descent methods may be used where only a random subset of the parameters are used to find a solution. A review of the many adaptations of the gradient descent algorithm can be found in [67].

2.3.6 Performance Metrics

To assess the performance of a computer vision model or algorithm, it is usually necessary to generate metrics that describe how well it is accomplishing a given task. These metrics are calculated on a test dataset which, in the case of learning-based models, has been held back from the training process. The choice of performance metric depends

on the task and what aspect of the performance is to be highlighted. Classification, regression, segmentation and registration tasks require mostly different performance metrics, some of these metrics are discussed in the following sections.

2.3.6.1 Classification Performance Metrics

Confusion Matrix The confusion matrix is a popular method for assessing the performance of a classification model on a test dataset by clearly displaying the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). This is most useful in the simplest case of binary classification. A confusion matrix for a binary classification problem is shown in Figure 2.19.

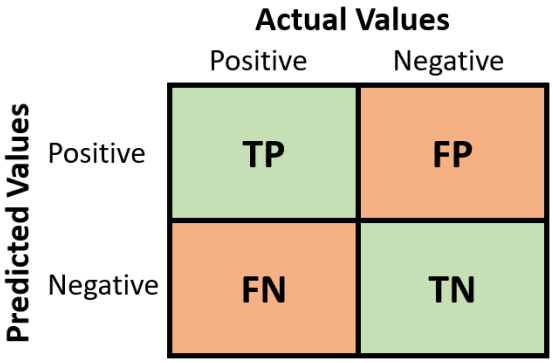


Figure 2.19: *A confusion matrix for a binary classification problem.*

Accuracy, Sensitivity and Specificity Accuracy is the most simple metric for classification tasks and is simply the proportion of correct predictions to the total number of samples. This can be expressed by equation 2.27

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.27}$$

Issues arise when using the accuracy for unbalanced classification problems where a high accuracy can be reported for a model that fails to correctly predict every example from the underrepresented class. In this case, and generally, for more information on model performance, it is beneficial to look at the sensitivity and specificity which is the accuracy when only considering the positive and negative cases respectively. Note that sensitivity is also often referred to as recall. The sensitivity and specificity can be calculated by equations 2.28 and 2.29 respectively.

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.28}$$

$$Specificity = \frac{TN}{TN + FP} \tag{2.29}$$

Precision and F1-score Precision, also known as positive predictive value, measures the proportion of true positive predictions among all positive predictions made by the model. This is a useful metric when minimising false positives is important. Precision is defined by equation 2.30.

$$Precision = \frac{TP}{TP + FP} \tag{2.30}$$

The F1-Score is a metric that combines precision and recall (sensitivity) into one value. It is the harmonic mean of precision and recall that provides a measure of a model's performance. The F1-Score is defined by Equation 2.31.

$$F1Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{2.31}$$

Receiver Operator Characteristics (ROC) The receiver operator characteristic (ROC) curve is a graphical representation used to assess the predictive power of a binary classifier and the trade-off between the true positive and false positive rates. A ROC curve is constructed by varying the probability threshold for predicted cases to be considered positive from 0 to 1 and recording the true and false positive rates. A plot is then made of the true positive rate vs the false positive rate at the threshold values from 0 to 1. Usually a diagonal line from (0,0) to (1,1) in the ROC plot is also included which represents the performance of a random classifier i.e. a classifier with no predictive power. A ROC curve that is closer to the upper-left corner of the plot is indicative of a well-performing model as it achieves a higher TPR while keeping a lower FPR. An example of a ROC curve is shown in Figure 2.20.

To summarise the ROC Curve and the predictive power of the model in a single metric, the area under the curve (AUC) of the ROC can be calculated. An AUC ROC of 1 indicates a perfect classifier and an AUC ROC of 0.5 indicates a classifier with the same predictive power as a random guess. The AUC ROC is a very popular tool for binary classifier performance evaluation as it summarises the predictive power of the

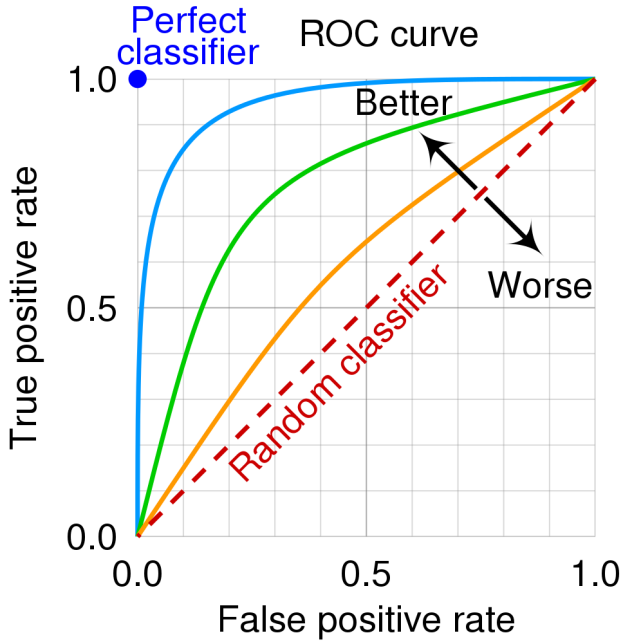


Figure 2.20: *Example ROC curves. Image taken from Wikipedia (https://en.wikipedia.org/wiki/Receiver_operating_characteristic)*

model in a single metric and is not biased by imbalanced datasets.

2.3.6.2 Regression Performance Metrics

In regression-based problems it cannot be defined if a prediction is correct or incorrect as with classification problems, instead, the size of the error of a prediction has to be found. This error is a measure of how far from the ground truth the predicted value is.

Mean Square Error and Mean Absolute Error The most simple and commonly applied regression error metrics are the mean square error (MSE) and mean absolute error (MAE) which are defined by equations 2.32 and 2.33 respectively.

$$\text{MSE} = \frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2 \tag{2.32}$$

$$\text{MAE} = \frac{1}{N}\sum_{n=1}^{N}|y_n - \hat{y}_n| \tag{2.33}$$

The main difference between the MSE and the MAE is that the MSE more harshly punishes predictions that are far from the ground truth value due to its square relationship.

2.3.6.3 Segmentation Performance Metrics

Semantic segmentation is a pixel-wise classification task, this means that all classification-based performance metrics are still relevant when considered in a pixel-wise context. For example, to calculate sensitivity for a segmentation task with two classes, the number of pixels of the positive class in the image that are correctly labelled when compared to the ground truth mask is used. Metrics such as accuracy, sensitivity and specificity are commonly used in segmentation tasks. In addition to these classification metrics, there are some segmentation-specific metrics.

Dice and Intersection Over Union (IoU) The Dice coefficient (also known as the Sørensen-Dice coefficient) and Intersection over Union (IoU) (also known as the Jaccard Index) are common metrics for assessing the performance of a segmentation model. The IoU measures the intersection of the predicted segmentation mask and the ground truth mask relative to their union. The IoU is defined by equation 2.34.

$$IoU(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{2.34}$$

Here, A represents the predicted segmentation mask, B represents the ground truth mask, $|A \cap B|$ represents the number of pixels common to both A and B while $|A \cup B|$ represents the number of pixels in either or both A and B. The IoU therefore emphasises how much the predicted and ground truth masks overlap.

The Dice coefficient is a similar metric that measures the overlap between the predicted segmentation mask and the ground truth mask. The Dice coefficient is defined by equation 2.35 [94].

$$Dice(A, B) = \frac{2|A \cap B|}{|A| + |B|} \tag{2.35}$$

For both the Dice and IoU metrics, the maximum value is 1 which represents a complete overlap of the two image masks, while the minimum value of 0 represents no overlap between the two sets. For tasks with multiple classes, a dice score can be calculated for each class or a global average can be taken.

2.3.6.4 Registration Performance Metrics

To analyse the performance of a registration model, the alignment of the areas or volumes to be registered in the test dataset must be compared. Most registration metrics are applicable to either segmentation or regression problems so they have been defined in previous sections. The metrics used depend on what aspect of the registration is to be highlighted and on what information is available. When the area or volume masks are available, the overlap of the masks can be used as a performance metric in the same way as for segmentation metrics. This means that the Dice and IoU metrics, defined by equations 2.35 and 2.34 respectively, can be directly applied to these registration problems.

As well as overlap based metrics, distance based metrics can be applied to calculate performance based on the average error in terms of distance between equivalent points. distance based metrics are generally used when feature points or landmarks are available. The target registration error (TRE) is a metric that can be used to quantify the alignment of feature points. To calculate the TRE, only points that were not used to register the two sets can be used so that any bias is avoided. The TRE of a point set is defined by equation 2.36 [95].

$$TRE = \frac{1}{N}\sum_{i=1}^{N}|m_i - r_i| \tag{2.36}$$

Here, N is the total number of feature points, m_i and r_i are the i^{th} feature points of the moving and reference images respectively. This is equivalent to the MAE for regression tasks, defined by equation 2.33.

Finally, a qualitative assessment of registration performance by expert clinicians is a valuable tool for assessing performance. This can help to highlight features that are not well aligned or any potential issues with a registration model.

Chapter 3

Literature Review

This chapter gives a summary of the published literature that is relevant to the research topics of this thesis, detailed later in chapters 4 to 6. The research topics covered are mostly self-contained and isolated pieces of work with in the lung cancer image processing space as as such, require their own separate literature reviews. These are detailed in their own sections which are:

- **Section3.1** Radiotherapy outcome prediction (relevant to chapters 5 and 4.
- **Section 3.3** Registration of pathology slides to PET/CT images (relevant to chapter 6).
- **Section 3.4** Gross pathology image segmentation (relevant to chapter 6).

3.1 Radiotherapy Outcome Prediction

Radiation therapy has seen ongoing advancements since its inception. While much of this work has been focused on developing the radiation delivery method, such as through the implementation of IMRT and VMAT devices, another active area of research, which has seen increased attention due to advances in AI and machine learning techniques in the past 15 years, is in the prediction of radiotherapy outcomes including the prediction of toxicities in healthy tissues.

Currently, the dose to organs at risk during RT is limited by simple dose metrics that have been determined from studies such as the seminal 2010 study titled "quantitative analysis of normal tissue effects in the clinic" (QUANTIC) [96]. The primary goal of the QUANTIC study was to establish dose-volume effects in all regions of human anatomy.

This study produced dose limits for OaRs, such as the lungs and the esophagus, using statistical methods to predict the maximum values of various dose metrics that should be adhered to so that the prevalence of toxicity in these OaRs is kept below a set threshold.

There have been many studies aiming to predict toxicities with different aims. Earlier studies from around the mid-1990s to the early 2010s generally aimed to investigate dose metrics such as the percentage of an OaRs volume receiving a dose above 20 Greys and clinical metrics such as a patients age for the prediction of various toxicities. These studies usually calculate dose volume histogram (DVH) metrics from the OaR being studied and then apply statistical models to find the metrics with the most predictive power.

More recent studies generally aim to produce a model that predicts if a patient will or will not develop a particular toxicity above a certain CTCAE grade, usually grade 2 or 3, with the AUC ROC being the standard metric that is used to determine a models performance. Often, additional information beyond dose and clinical metrics such as radiomic features calculated from CT or MRI scans are included. A differentiator in these studies is whether they use pre-treatment information or both pre and post-treatment information. Studies that use post-treatment information could only be applied clinically for monitoring and early detection of OaR toxicity. Studies that use pre-treatment information are potentially more beneficial as they could be applied in all the same ways as post-treatment information based methods but even earlier in treatment as well as for the prevention of toxicity through altering dose plans. The work presented in Chapters 4 and 5 use only pre-treatment information for the outcome predictions.

A challenge when reviewing the literature on radiotherapy toxicity prediction is that there is a lack of gold-standard datasets that can used across studies. Most studies, therefore, use data that is not publically accessible making the comparison of methods challenging. This is further complicated by the fact that many studies use data from clinical trials which often have patient populations that are not statistically equal to the general clinical setting due to the protocols of the clinical trial. These clinical trials often investigate multiple parameters such as changing radiation dose or using concurrent chemotherapy at the same time further adding to the challenge of comparing prediction studies. For this reason, it is not possible to directly compare the results

of any two RT outcome prediction studies unless they use the same dataset. An aim of this thesis is to improve the comparison of different RT toxicity prediction methods by using the largest publicly available dataset for esophageal toxicity prediction to test several prediction models. With these challenges in mind, previous works on RT outcome prediction are reviewed in the following subsections.

3.1.1 The Lyman-Kutcher-Burman Model

The first radiotherapy prediction models to be developed are mathematical models designed to relate the dose an organ receives during RT to the probability that a certain toxicity outcome is observed. One of the most popular normal tissue complication probability (NTCP) models is the Lyman Kutcher Burman (LKB) NTCP model [97] [98] [99] [100] which was used in the QUANTEC study for estimating the likelihood of toxicity endpoints. The LKB NTCP model assumes that the likelihood of an organ developing a particular toxicity is dependent on the total dose received by that organ as well as the volume of the organ irradiated by specific dose levels. The LKB model has three parameters that must be determined empirically from clinical data with known patient outcomes. These parameters are the TD50, m and n. The TD50 represents the dose that would cause 50% of patients to develop a specific toxicity within 5 years, this assumes a homogeneous dose applied to the OaR. The variable m determines how steep the dose response curve is and n determines the volume effect of the organ being studied. The LKB NTCP model is described by equation 3.1.

$$\text{NTCP} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{x^2}{2}} dx \tag{3.1}$$

$$t = \frac{\text{EUD} - TD_{50}}{m^* TD_{50}} \tag{3.2}$$

$$\text{EUD} = \left(\sum_i v_i D_i^{\frac{1}{n}} \right)^n \tag{3.3}$$

Here v_i and D_i are linked volume and dose values respectively where v_i is the volume receiving a dose above D_i.

3.1.2 Dosimetric Features

For the prediction of toxicity occurrence post-RT, all models will include information regarding the dose to the organ or organs relevant to that particular toxicity. To include this dose information, dose features need to be extracted from the planning dose image which is the most accurate representation of the applied dose. This becomes an image feature extraction task where the methods described in section 2.3 can be applied. The most commonly applied dose based features are handcrafted metrics which are popular due to their ease of calculation and, most notably, their high level of explainability which is favoured by clinicians. These features are usually calculated from teh dose volume histogram of a particular organ at risk, an example of the dose vilume histogram for the lung region of a RT patient with NSCLC is given in figure 3.1. Some of the most standard dose metrics are given below:

- D_{mean}: The mean dose to an organ at risk.
- D_{max}: The max dose to an organ at risk.
- V_x: The total volume of an OaR that is receiving a dose of x Grays or more. This is generally reported in steps of 5 Gy and can be calculated in terms of an absolute volume as well as a percentage.
- D_{2cm}: The maximum dose at 2 cm from the PTV as percentage of the prescribed dose.
- R_{50} & R_{100}: The R_{100} metric is defined as the volume of region receiving 100% of the prescribed dose divided by the PTV volume, R_{50} is the same measure but using the volume of region receiving 50% of the prescribed dose. These are used as a measure of conformality of the dose plan
- D_{2cc}: The minimum dose to the $2cm^3$ highest dose region of an OaR.

In addition to these common metrics, there are many more dose based metrics that have been applied for outcome prediction, some of which are specific to individual OaRs. It has been shown that dose features are highly correlated [101], meaning it is usually unnecessary to exhaustively calculate every potential dose metric for a specific problem.

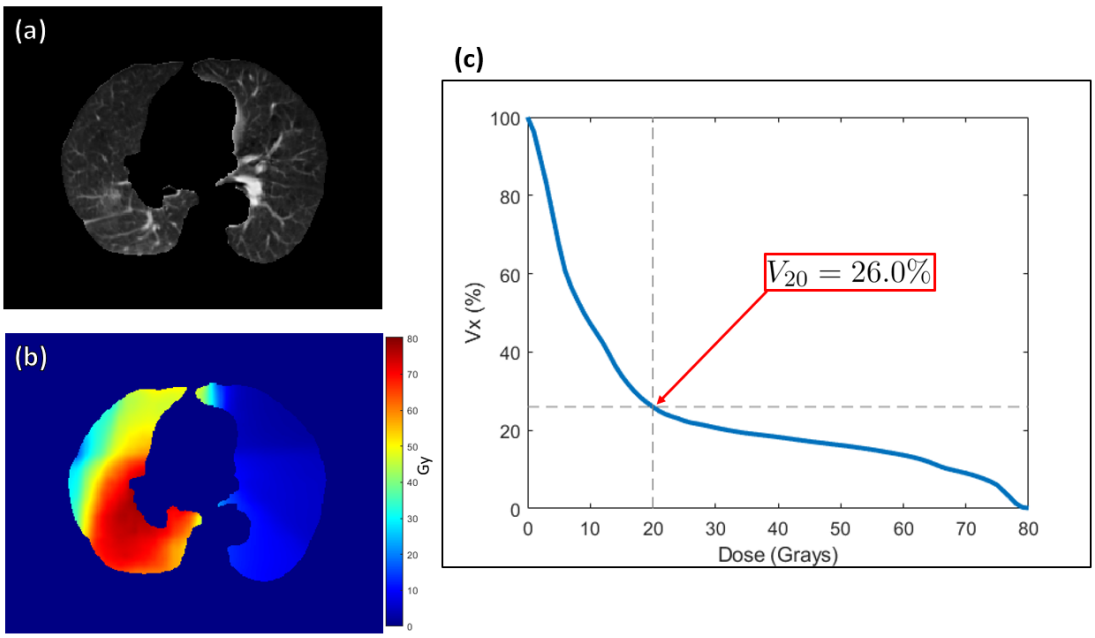


Figure 3.1: ***(a)*** *a CT image segmented to only include the lungs,* ***(b)*** *the corresponding dose map, segmented to only include the lungs and* ***(c)*** *the corresponding lung dose volume histogram with the V_{20} highlighted.*

3.1.3 Pulmonary Radiation Toxicity Prediction

In this subsection, papers relating to the prediction of pulmonary toxicity are discussed. The approaches for toxicity prediction can be grouped into three categories:

- Dose feature based
- Clinical feature based
- Radiomic feature based

These approaches directly relate to the work in chapter 4. Much of the previous work has been focused on the prediction of radiation pneumonitis specifically.

Dose Feature Approaches Prediction of pulmonary toxicity from pre-RT data has previously been shown to be possible in the literature. Much of the focus has been on dosimetric and dosiomic approaches. Dose volume histogram features calculated for the lung including, the mean lung dose (MLD), V_5 ,V_{10} and V_{20}. have been shown to act as a predictor for radiation pneumonitis [102], [103]. The lung specific portion of the QUANTEC study [104] recommended limiting the V_{20} to $\leq$30–35% and the mean lung dose (MLD) to $\leq$20–23 Gy, if conventional fractionation is used, to limit the risk

of radiation pneumonitis to ≤20% in patients with NSCLC. This was based on a review of the data available from clinical trials at the time. There is some evidence that the predictive power of certain dose metrics is dependent on whether they are calculated for contralateral or ipsilateral lung [105]. The LKB NTCP model has been applied for the prediction of pulmonary toxicity in [106].

More recently, spatial features extracted from the RT dose distribution, known as dosiomics, have been included in predictive models. Image texture methods have been applied to predict radiation pneumonitis from RT planning dose maps [107], [108]. As well as classic texture approaches, deep learning-based approaches have been applied to extract features from dose images for pneumonitis prediction [109], [110].

Clinical Feature Approaches After dose features, the most well researched features for the prediction of pulmonary toxicity are clinical features. Clinical features such as age [111], gender [112, 113], the presence of chronic lung disease [114, 113] and the use of concurrent chemotherapy [112, 111] have been reported to be predictive features for the task of radiation pneumonitis prediction. Núñez-Benjumea, et al. applied 300 predictive models for the task of predicting acute cough, dyspnea and pneumonitis as well as chronic dyspnea and pneumonitis in an attempt to benchmark these models for this task [115]. A total of 875 patients datasets were used, though not every patient was used for each endpoint prediction. For the prediction, dose and clinical features were used. Different models and features were chosen to be optimal for different endpoints with the ANN model combined with the minimum redundancy maximum relevance (mRMR) method for feature reduction being selected for the prediction of chronic pneumonitis, achieving an AUC of 0.77 on the external validation set which contained only 7 positive cases.

Radiomic Approaches Radiomic approaches often have been applied to determine if a patient has developed pneumonitis from their post-RT CT scans. It has been shown that changes in radiomic features between pre-and post operative lung CT images can be used to determine if a patient has radiation pneumonitis [116]. An increase in CT image density post-RT has been shown to correlate with regions of higher dose and PTV size [117]. CT image markers extracted at different fictionalisation time points were shown to predict lung density changes [118]. A strong correlation has also been observed between RT dose and post-RT changes in CT image density of normal lung

tissue [119].

Less commonly, radiomic approaches have been applied to the pre-RT CT images to predict radiation pulmonary toxicity. S. Krafft, et al. [120] calculated radiomic features from the lung volume of pre-treatment CT scans of 192 patients that had received RT for NSCLC. Of these 192 patients, 30 had presented a radiation pneumonitis of grade ≥ 3. These radiomic features were combined clinical and dosimetric features to produce a total of 6851 features. A least absolute shrinkage and selection operator (LASSO) based logistic regression model was trained on the patient features using 10-fold cross validation to split the data. This produced an average AUC of 0.68.

C. Puttanawarut et al. extracted radiomic and dosiomic features from the lung RoI of CT and Dose images respectively for 101 patients with esophageal cancer and 93 patients with NSCLC for the prediction of grade ≥ 2 radiation pneumonitis [121]. Multivariate logistic regression models were trained on different subsets of these features with the esophageal cancer dataset being used as the training dataset and the lung cancer cohort being used as the test dataset. The best performing model achieved an AUC of 0.77 though there were only 16 cases in the positive group. Z. Zhang et al. Follow a similar approach for the classification of radiation pneumonitis from pre-RT data by calculating radiomics, dosimetric and dosiomics and classifying these features using a multivariate logistic regression model. A training dataset with a size of 314 patients was used with a test dataset of 35 patients where only 9 patients developed grade ≥ 2 pneumonitis) [122]. The best performing model here produced an AUC of 0.85. Zhang et al. Then added to this by applying a deep learning model instead of a radiomic and dosiomic approach [123]. A 3D-ResNet model used for feature extraction and classification was trained on a dataset contaning 314 patients and tested on a dataset containing 352 patients from the RTOG-0617 trial dataset [124]. Their model achieved an AUCs between 0.55 and 0.83 for separately defined test sets which highlights the variability of results in this field due to the lack of large standardised datasets.

CT radiomic analysis of the lung volume has also been applied to predict the occurrence of pneumonitis for patients recieving immunotherapy. R. Colen et al. presented a pilot study where they predicted the pneumonitis outcome in 32 patients with advanced cancer, only two of which went on to develop pneumonitis, as a binary classification based on their pre-treatment CT scans to [125]. In this study they achieved a 100%

accuracy with the caveat that the dataset only contained two pneumonitis cases.

K. Tsujino, et al. [126] used the pulmonary fibrosis score and pulmonary emphysema score from a patients pre-RT CT scan combined with dose volume histogram metrics to produce a radiation pneumonitis prediction model. The pulmonary fibrosis score and pulmonary emphysema score were defined by a single experienced diagnostic radiologist independently. The dataset used contained 122 patients, 14 of which went on to develop radiation pneumonitis of grade $\geq$ 2. The maximum AUC ROC achieved with this dataset was 0.88.

3.1.4 Esophageal Radiation Toxicity Prediction

In this subsection, papers relating to the prediction of radiation induced esophageal toxicity are discussed. This directly relates to the work in chapter 5. A number of esophageal toxicities may occur in lung cancer patients treated with radiotherapy (RT) [127], for example, radiation esophagitis (RE), which is an inflammation of the esophagus, dysphagia and stenosism [19]. Most of these present as acute toxicities generally peaking in severity between 4 and 8 weeks from the start of treatment [128]. Late toxicity of the esophagus following RT is less common but can also be observed [129] and generally presents as stricture and associated dysphagia which would typically develop after 3 to 8 months following RT [130]. Of all these toxicities, RE is the most common acute toxicity [131, 132] and furthermore, as reported in the long-term follow-up result of the Radiation Therapy Oncology Group (RTOG) 0617 study [133], grade 3 or above RE is one of the factors with the highest overall predictive power for a patients overall survival.

As discussed previously, the prevalence of these toxicities is reduced by limiting certain dose based metrics which have been determined from studies such as the QUANTIC study [96]. The esophagus specific portion of the QUANTEC study determined that a mean esophagus dose of $<$ 34 Gy would result in a 5–20% chance of grade 3 or above esophagitis and keeping the constraints V_{35} $<$50%, V_{50} $<$40% and V_{70} $<$20% would result in a $<$30% chance of grade 2 or above esophagitis [134].

Dose volume histogram (DVH) and clinical factors, such as a patient's age, have been shown to be predictive of esophageal toxicities. Studies have found that the mean esophageal dose (MED) [135, 136, 137, 138, 139, 140], V_{20}, V_{30}, V_{40}, V_{50}, V_{60} [136, 137, 141, 142, 143] the maximum dose to 2cc [136] and concurrent chemotherapy [138, 139,

140, 143] are the strongest predictors of esophageal toxicity. Some of these studies are discussed below.

S.J. Ahn, et al. analysed DVH and clinical factors of 254 patients receiving RT for lung cancer by logistic regression analysis, contingency table analyses, and Fisher's exact tests to determine which factors were the most statistically significant for the occurrence of RE [130]. It was found that most dosimetric parameters were predictive with the maximal esophageal dose being the most predictive of acute toxicity. In a similar study, A. Ozgen, et al. applied the Kruskal-Wallis test of statistical significance to dose volume histogram parameters from a dataset of 72 patients to find which factors produced the highest predictive power for developing RE [135]. It was found that the risk of grade 2 esophagitis was significantly correlated with the mean esophageal dose (MED) with a MED of $\geq$ 28 Gy being the most statistically significant predictor of grade $\geq$ 2 esophagitis. No correlation was found between variables describing the volume of esophagus irradiated and esophagitis. Furthermore, the mean esophageal dose irradiated was the most statistically significant factor associated with acute esophagitis Grade 2 or worse. P. Paximadis, et al. performed a similar study on a dataset of 533 patients and showed using logistic regression that the DVH features with the most predictive power for grade $\geq$ 3 esophagitis are the V20, V30, V40, V50, V60, mean dose, the maximum dose to 2cc and the generalised equivalent uniform dose [136].

The literature regarding RE prediction published before 2009 has been reviewed in the systematic review paper [137]. This paper determined the following six factors as being highly predictive of RE; mean esophagus dose (MED), V_{20Gy}, V_{30Gy}, V_{40Gy}, V_{45Gy}, V_{50Gy} where V_{xGy} is the percentage of the total esophagus volume receiving over x Grays of dose.

More recent studies generally aim to produce a predictive model for esophageal toxicity based on both the DVH and additional features. These studies generally construct a binary classification model to group patients who developed esophageal toxicity less than or greater than or equal to certain grade threshold which is always either grade 2 or 3. In these studies the AUC ROC is the most commonly used metric for determining the predictive power of a model. These studies that have been identified from the literature are discussed below.

P. Hawkins et al. investigated the predictive power of pre-treatment cytokine levels as well as dose metrics on the occurrence of grade 3 or above esophagitis in patients for

NSCLC with external beam-RT [144]. The incidence of grade 3 esophagitis in the 126 analysed patients was 13 (13/126). A binary classification was performed using logistic regression with "elastic net" penalisation. Using only dose metrics produced an AUC of 0.75 and the best model used dose and clinical features which achieved an AUC of 0.78. This was not improved by the inclusion of the pre-treatment cytokine levels.

X Zheng et al. took a multi-omics approach to predict RE in patients with NSCLC receiving RT [145]. This approach combined radiomic and dosiomic features calculated from the esophagus VoI in the patients planning CT image and dose maps respectively as well as clinical features. The dataset used was from a single centre and contained RT sets from 162 patients, 51 of which developed grade 2 or over RE. Using a cross-validation approach so that the models could be tested on the full dataset, a logistic regression classifier was applied, this led to an average AUC of 0.75 on the test images. In this study it was observed that the radiomic features were producing most of the predictive power where the radiomics only model produced an AUC of 0.74 and the dosiomics only model produced an AUC of 0.60.

S. Wang et al. produced a study to retrospectively predict the occurrence of grade $\geq$ 2 RE by combining predictions from the esophagus generalized equivalent uniform dose, the patients IL-8 chemokine factor and their age [146]. Their dataset consisted of 129 patients with NSCLC treated by RT, 49 of which developed grade $\geq$2 RE. On its own, the esophagus dose model achieved an AUC of 0.70 and the combined model with the dose, IL-8 and age produced an AUC of 0.78.

A study by J. S. Niedzielski et al. [147] has shown that esophageal expansion as measured by changes in CT images can be used to predict the occurrence of RE with prediction, in this case, occurring after the RT treatment has been applied. In addition to the esophageal expansion features calculated, clinical and dose features were also used for classification. K-means clustering was applied to classify patients that would or would not develop grade 2 or above esophagitis with a training set of 94 patients and a test set of 32 patients. Cross-validation was applied so that the model could be tested on the full set of 126 patients. The maximum AUC achieved on the test set, averaged over all cross-validation repetitions, was 0.75.

Esophageal toxicity prediction with the RTOG-0617 dataset is has been applied in the study [148]. Here, different machine learning methods and data augmentation approaches to classify grade 3 esophagitis from dose and clinical features were applied,

achieving a maximum AUC of 0.706.

3.1.5 Radiation Toxicity Prediction in Other Anatomical Regions

As well as predicting pulmonary and esophageal toxicity, there have been studies aiming to predict radiation induced toxicities in other parts of the body such as the prostate [149, 150], the head and neck [151] and rectum [152]. These studies generally follow the same methodologies as the studies for pulmonary or esophageal toxicity prediction where dose features are calculated for a specific OaR which are then combined with clinical features and sometimes an additional feature source for the prediction. Deep learning methods such as CNNs are occasionally applied but this limited due to a lack of large datasets.

3.2 CNNs for Feature Extraction

In Chapter 4 a CNN is used for the extraction of image features from CT scans for the purpose of radiotherapy outcome prediction. A brief review of the use of CNNs for the extraction of features from medical images is presented here.

CNNs have been shown to generate features that are generalisable to tasks separate to those they were trained on [153]. This allows for their application in few or zero-shot approach for medical imaging even when they were trained on a dataset from a different imaging domain. The hierarchical feature representations of CNNs, where increasingly deep layers represent increasingly complex and abstract features, increase there generalisability across imaging domains.

Varshni, et al. applied several different CNN models including DenseNet-169, ResNet-50 and VGG-19 for the task of feature extraction from chest x-ray images for pneumonitis prediction [154]. The networks had all been pre-trained on ImageNet and features were extracted from all convolutional layers of the network with no additional model training. Separate classification models were applied for the binary pneumonitis classification including SVM [155], random forest [156] and Naive Bayes [157] models. The best performing model used the ResNet-500 CNN as a feature extractor with a SVM based classifier and achieved an AUC of 0.775. The segment anything model [158], which was trained on natural images, has been applied with model fine-tuning and in a zero-shot approach by Peilun, et al.[159] for the purpose of medical image

segmentation across several modalities. They found that when applied in a zero-shot manner, the results were satisfactory for most domains but worse than domain specific models. By fine-tuning they managed to improve the model performance across all modalities. Zero-shot learning has additionally been applied for histopathology image classification [160]. Outside of the medical field, zero-shot transfer learning has been applied for many tasks such as super-resolution imaging [161].

3.3 Registration of Pathology Slides to PET/CT images

Chapter 6 details the work developing a method for the automatic registration for the registration of PET/CT to pathology imaging. This follows on from the work by Reines March, et al. [2], as such, Reines March et al. is discussed here in detail.

PET-CT is currently the standard method for imaging most cancers as the tumour detection from the functional imaging of the PET scan combined with the anatomical imaging of the CT scan provides the information necessary for treatment and diagnosis decisions to be made. The accuracy of this information is limited by the spatial resolution of the images as well as errors in the imaging techniques such as the movement introduced by breathing. Additionally, it is not currently possible to confidently determine the microenvironment of the tumour from features in PET and CT images. To remove these uncertainties and allow more information regarding the cellular make-up of the tumour to be determined from PET-CT scans, the PET-CT scan needs to be validated against tumour pathology images which provide microscopic ground truth information [162]. There have been several previous studies aiming to relate the information in histopathology images to both PET/CT scans of the lung and other areas of anatomy as well as the inclusion of different imaging modalities such as MRI. These studies aim to do this by registering the tumour volume in pathology and PET/CT images so that the microscopic tumour details can be viewed in the same coordinate system as the PET/CT.

Studies that aim to register in vivo modalities with histopathology slides generally follow similar methodologies. The main aspects these methodologies are:

1. Pathological processing of the surgically removed specimen.

2. Reconstruction of the tumour from the pathology modality.

3. Registration of the pathology tumour volume to the PET/CT tumour volume.

Standard pathological processing of surgically resected specimens involves slicing the specimen free hand to expose the tumour for imaging. For registration tasks, this makes it impossible to reconstruct the geometric shape of the tumour. Therefore, the pathological processing methods in these registration studies have to involve some additional methodology to conserve the tumours geometrical information as accurately as possible. This usually involves inflating the lung with agar to match the geometry in-vivo before slicing.

The methods for reconstructing the tumour from the pathology modality depend on the specific aims and outcomes of any particular study. For studies aiming for volume-to-volume registration, it is necessary to reconstruct the 3D tumour volume. Other studies may only aim to compare the total tumour volumes in each modality in which case an accurate surface is not required and the tumour can be segmented in the 2D slices individually to infer the total volume. In both cases, tumour segmentations are acquired through manual segmentations by experienced clinicians.

The final stage in any pathology to in vivo modality registration study is the registration methods that are used. A 3D to 3D volume registration can be applied or 2D to 3D where the pathology images are matched to the closest slice in the CT volume. The registrations are applied in most studies manually by clinicians using rotations and translations only. While tumour volumes are comprised of soft tissue that would usually be registered by non-rigid methods, for registering PET-CT to histopathology slices, rigid registrations are usually used. This is because the aim of registering PET and CT images with histopathology slices is to understand what can be learned from the PET and CT images so distortions of the PET and CT images should be avoided.

In this section, the published works that are relevant to our work on the registration of PET/CT images to histopathology slides are reviewed. Section 3.3.1 details works on in-vivo to ex-vivo registrations for regions of anatomy other than the lung and in-vivo modalities other than PET/CT while section 3.3.1.1 reviews papers focused on PET/CT and the lung.

3.3.1 Registration of pathology and in-vivo imaging modalities

Puri et al. [163] describe their method of registering PET images and histopathology slices acquired from patients with laryngeal cancer. A PET-CT scan was performed

followed by a total laryngectomy to remove the tumour and surrounding tissue. The resected specimen was fixed in formalin, which is standard practice when processing histopathology samples, and sea urchin spines were inserted into the specimen to act as fiducials to aid the registration process. The resected specimen was then CT imaged again to produce an ex vivo CT image and the specimen was sliced and histopathology slides were produced. The ex-vivo CT scan was taken to understand what deformations have occurred in the tumour volume due to its surgical removal. This additional ex-vivo scan is also applied in other papers [164, 165, 166].The in-vivo PET and CT images were registered by a rigid mutual information based registration method then the in-vivo CT image was registered to the ex-vivo image by the same method. The transform applied to the in-vivo CT image was then applied to the PET image so that it was also registered to the ex-vivo CT image. The corresponding pathology slice for the 2D PET and CT images was found by finding the lowest RMS error between the fiducials in the pathology image and the ex-vivo CT image. This resulted in an average error in the registration between the PET images and histopathology slices of 3.0mm.

Garcia-Parra et al. [164] register 18F-FAZA PET/CT images with histopathology slices using an ex-vivo MRI scan as an intermediate step for prostate cancer patients. They aimed to test the potential of 18F-FAZA PET for the identification of areas of tumour hypoxia. The registration was achieved using mutual information based methods with a thin-plate spline deformation.

Meyer et al. [166] provide an overview of the challenges of 3D image registration of PET/CT images with pathology slices for prostate cancer. Here they describe how the 3D multi-modality registration problem has to be treated as fully 3D to avoid errors in the registration. What this means is that alignment of 3D organ or tumour surfaces will produce errors within the organ or tumour volume and to achieve the most accurate results the whole 3D volume has to be considered in the registration process. Due to the vastly different nature of the information contained within PET-CT and pathology images, this is a challenging problem to solve as there are generally no landmarks contained within a pathology tumour volume that can be seen in a PET-CT image. For this reason, the addition of an ex-vivo scan described above and in [163] is the only current solution to this problem. Another problem highlighted by Meyer et al. is that the mechanical cutting of a tumour volume into slices for 2D pathology imaging will introduce deformations to the tumour volume. There is again no standard solution to

this problem and the choice of registration model should take this into account. The prostate cancer registration problem shares many of the same challenges as those found in lung cancer so the problems highlighted are relevant to the research project.

Shao et al. present ProsRegNet [167], a deep learning based network for the registration of histopathology and MRI images of the prostate. This is a CNN based network trained using MRI and histopathology images of 99 patients. Testing of the network generated an average error of 2.7mm in the registration. The main advantage of this network compared to the state-of-the-art is the drastically reduced computational complexity required to perform a registration once the network has been trained.

3.3.1.1 Registration of Pathology and PET/CT Lung Images

The registration of in-vivo modalities to ex-vivo histopathology images of a surgically removed lung specimen presents additional challenges compared to other regions of anatomy due to the collapse of lung tissue that usually to occurs after it is extracted from a patient due to the mechanical properties of lung tissue. This introduces large deformations in the volume of a lung tumour between in-vivo modalities and histopathology slides. In registration studies, this deformation is generally counteracted by inflating the surgically removed specimen with formalin or agar so that the specimen more closely matches its in-vivo shape.

Stroom et al. [168] developed a method for pathology-correlated imaging for lung tumours with the aim to more accurately define the gross tumour volume and gross clinical volume for radiotherapy. PET-CT scans were taken of 5 patients with non–small-cell lung cancer (NSCLC) before a lobectomy was performed. The resected lung lobes were inflated with formalin before they were sliced, photographed and histopathologically imaged. The registration was performed by first matching the CT slices to the corresponding pathology slices then the PET/CT slices were deformed to match the corresponding block face photographs. A similar study was completed by Loon et al. [169] who investigated the extent of microscopic disease within the clinical target volume used in radiotherapy treatment planning. 34 patients with NSCLC were PET-CT imaged before undergoing a lobectomy. The resected lobes were inflated with formalin before they were sliced and histopathology images were taken. No registration was performed in this study, only the reconstructed tumour volumes were compared.

Yu et al. [170] aimed to determine the cut-off standard uptake (SUV) in 18F-FDG

PET/CT that creates the best volume match to pathological tumour volume. Fifteen patients with NSCLC were PET/CT imaged before undergoing a lobectomy to remove the tumour volume. The removed specimen was fixed in formalin, sliced and histopathology slices were produced. The tumour boundary in the slices was delineated by a pathologist. Reconstructed pathological tumour volumes were compared to reconstructed PET tumour volumes at different SUV cut-off values to determine the SUV cut-off value that results in the best match between the volumes of the two modalities.

Wanet et al. [171] aimed to validate a gradient-based segmentation method for gross tumour volume delineation in FDG-PET for NSCLC. Ten patients with NSCLC were PET/CT imaged before undergoing a lobectomy. The removed surgical specimen was inflated with agar, frozen and sliced. The sliced specimen was reconstructed to a 3D volume. The PET/CT images were manually registered to the histopathology slices using a rigid registration.

3.3.2 Previous EngD Work

The work in Chapter 6 of this thesis directly follows on from work detailed in the thesis titled "Registration of Pre-Operative Lung Cancer PET/CT Scans with Post-Operative Histopathology Images" by Gabriel Reines March completed in 2020 [2]. Chapter 6 of this thesis applies and advances many of these methods so the work by Reines March is covered in this subsection in detail. The goal of the work by Reines March was to create an imaging framework for registering in-vivo PET/CT scans with ex-vivo histopathology slices for patients with NSCLC to create a multi-modality map of the tumour environment.

3.3.2.1 Phantom Study, Simulations and Tumour Reconstruction

The first component of the work involved developing a registration algorithm and testing it on a tissue-mimicking phantom. A phantom was used as it provided a reproducible ground truth and avoided ethical constraints of using human participants in the early stages of the work. It also allowed computational experiments to be conducted on the phantom model which could then be verified experimentally. The phantom was designed to mimic anatomical features that may be seen in lung tumours in-vivo by having extruding lobes and depressions, these features were aimed to ensure the robustness of any registration tests.

A computer model of the phantom was virtually sliced at different thicknesses to test the effect that different slice thickness would have on volume reconstruction and to test different volume reconstruction techniques. Volume reconstruction involved an interpolation to generate pixel values for coordinates between the slices. Nearest-neighbour, linear and cubic-spline interpolation were all used to determine which one would provide the most accurate volume reconstruction. The reconstructed volume was compared to the known volume of the phantom using these different interpolation techniques for different slice thicknesses. It was found that larger slice thicknesses reduced the reconstructed volume as extrusions were omitted from the reconstructed volume, this decreased the alignment with the known volume of the phantom. It was also found that while nearest-neighbour interpolation provided the closest volume match to the known volume of the phantom, it also provided the largest shape distortion of the phantom so cubic-spline interpolation was chosen as the best interpolation method. An issue with these interpolation methods was that extrema region were poorly reconstructed due. This was overcome by applying methods for both linear and curvature-based extrapolation methods. The curvature-based method was found to produce more accurate reconstructions in the simulations. This curvature based method uses second-order differences from the previous two slices to extrapolate the end region of the tumour shape following the same curvature trend.

3.3.2.2 Pathology Slicing Process

A custom slicing rig was designed so that slices taken of both the phantom and later any surgical samples, would be of the same thickness and so that the thickness of a single slice would not change over that slice. This was not done in the literature as most similar studies performed a free hand slicing of the specimen, which is the standard pathology processing procedure. The slicing rig was comprised of a cylindrical tissue container mounted within an acrylic frame, featuring a threaded plunger and slot for the travel of a surgical blade. The plunger lifts the tissue, which was embedded in agar, to the desired slicing level, while the blade performs clean, flat cuts. The slice thickness is adjustable based on the plungers known pitch of 2.5 mm meaning two full turns of the plunger would raise the sample by 5mm. Each exposed cross-section was photographed by a camera in a fixed tripod mount before subsequent slicing iterations with care to remove any excess moisture from the pathology sample prior to imaging.

The pathology slicing workflow is summarised in Figure 3.2.

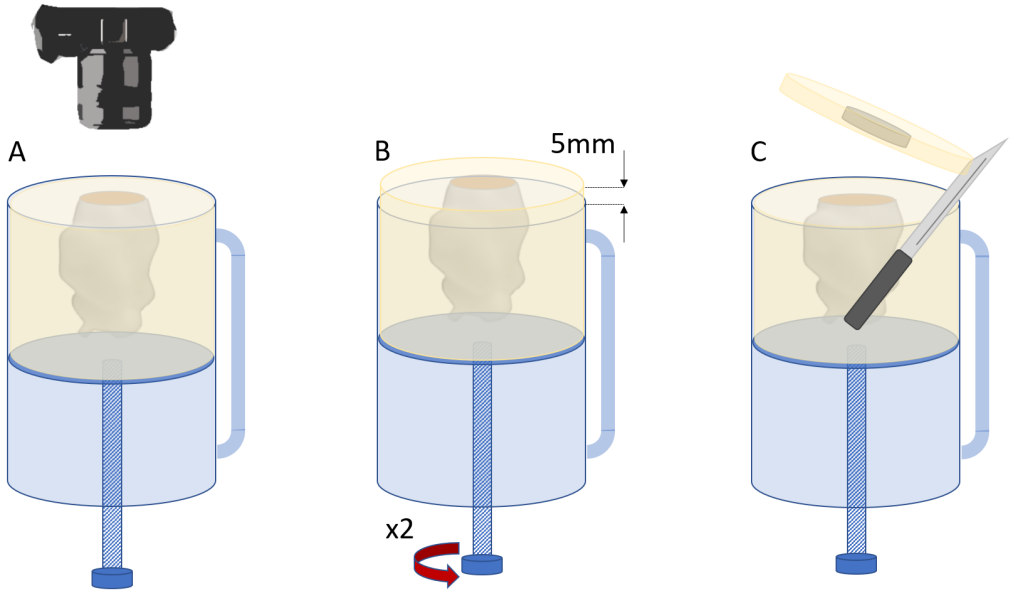


Figure 3.2: *The workflow used for the slicing rig developed and applied by Reines March et al. [2]. **A** the sample is imaged, **B** the sample is raised by 5mm, **C** the sample is sliced. This figure was taken from the thesis [2].*

3.3.2.3 Lung Tumour Image Registration

The registration methodology to align PET/CT images with histopathology slides can be broken down into individual steps. First, the microscopic histology slides were registered to gross photographs of the specimen. This was done using feature-based moving least squares with an affine deformation registration where anatomical landmarks such as blood vessels were manually selected by clinicians in both pathology images and used as control points for the algorithm. Next, the 3D volume reconstruction of the tumour volume in the CT and pathology modalities was performed by first manually segmenting the volumes and then performing a volume reconstruction as described in Section 3.3.2.1.

The third step was a registration of the reconstructed 3D volume from the PET/CT images to the reconstructed 3D volume from the pathology photographs. The registration was performed with a rigid transformation using the sum of squared differences of the surface points of the CT and pathology volumes as the similarity measure and a gradient descent algorithm as the optimiser. The final step was to apply this transfor-

mation to the PET and CT images so that they were aligned with the pathology slices. A flowchart of these steps can be seen in figure 3.3.

This registration algorithm was applied to full imaging datasets from 9 patients. It was found that while CT volumes closely matched pathology volumes, allowing for accurate registrations, the reconstructed PET volumes were on average around twice the size of the pathology volumes. This was because of the blurring of the PET images due to patient breathing though the only method applied for PET segmentation was a $0.5SUV_{max}$ threshold. Due to this, no results were produced for the PET images.

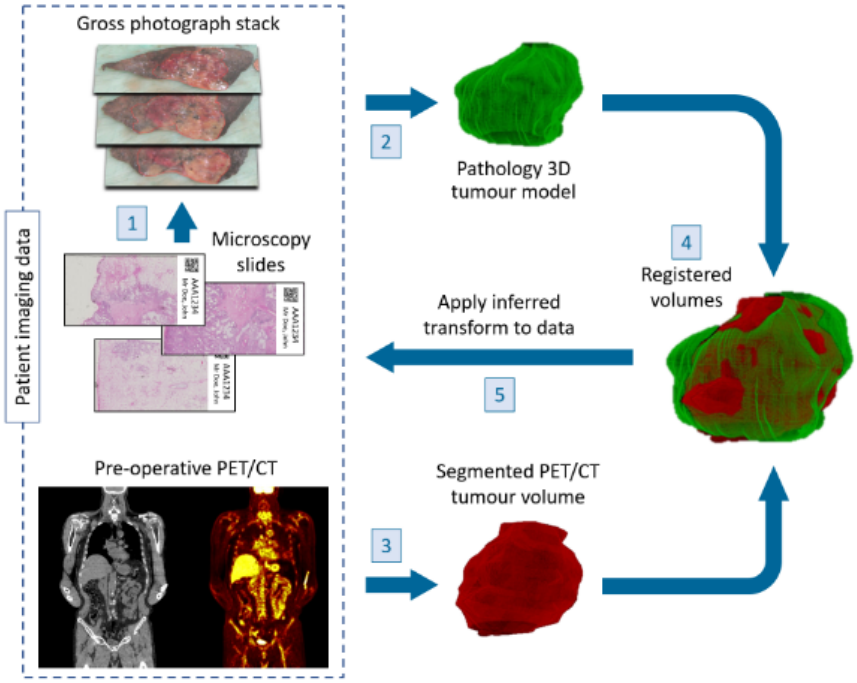


Figure 3.3: *Flowchart depicting the steps taken for registration of PET/CT images to histopathology slides in [2]. Figure taken from [2].*

The results were then compared to a manual registration performed by both a pathology consultant and a clinical oncology consultant individually to test the robustness of the proposed framework. This showed that for the CT data there was a good alignment with the manually registered data and in a qualitative analysis the developed automatic registration framework was equivalent to or out-performed the manual registrations for every assessed patient.

3.4 Gross Pathology Segmentation

The second half of Chapter 6 details the work on a study aiming to produce a deep learning based methodology for the automatic segmentation of gross pathology photographs of lung cancer specimens. The literature relevant to this is covered in this section.

Figure 3.4: *Example of a gross pathology photograph of a surgically resected lung lobe that has been sliced to reveal a lung tumour.*

Before any image processing of gross pathological specimens, the image capture process for gross pathology must be undertaken. The procedures and best practices generally used to capture gross pathology photographs have been described in [172]. Best practices include placing the pathology specimens on a background that provides a good contrast between the specimen and background. The specimen should be well-lit with lighting located to the sides of the specimen as overhead lighting is more likely to cause reflections that may obscure anatomy. Excess moisture should also be removed from the surface of the specimen as this may obscure the underlying anatomy through the liquids opacity or the increased reflections this may cause. The specimen should also be well framed, in focus and the imaging plane should be the same as the slicing plane. An example of a gross pathological photograph of a surgically resected lung specimen, sliced to reveal a lung tumour, is given in Figure 3.4. The International Association for

the Study of Lung Cancer (IASLC) recommends pathology photography as a standard part of pathology processing for NSCLC specimen processing after neoadjuvant therapy [173].

Gross pathology photography has been applied in some studies to provide the information necessary to transform WSI so that the geometry of the images more accurately represents what would have been observed in-vivo. This has often been for the application of registering images from the PET and pathology modalities [166, 165] as detailed in Section 3.3. Gross pathology photography has been used as an important feature in many studies where regions are generally segmented by experienced pathologists. These studies include investigations into the mechanical properties of tissues[174], ablation treatment monitoring [175] and histologically diagnosed cardiac sarcoidosis [176]. A semi-automatic vector quantisation based pathology segmentation approach has been applied to segment regions of fibrosis in gross photographs to determine the overall prevalence of fibrotic tissue in lymph nodes [4]. Hyperspectral image based tumour segmentation has also been applied for application in real-time tissue classification during laparoscopic surgery [177].

An area where gross pathology photography has been applied more extensively than the lung cancer domain is skin lesion photography. There are similarities between lung lesion photography and skin lesion photography that make the greater catalogue of previous work on skin lesion segmentation relevant here. One such example is the work by Y. Yuan et al. who produced a fully connected (FC) convolutional neural network (CNN) based approach for skin lesion segmentation with a Jaccard distance-based loss function with their highest performing method consisting of an ensemble of six separate FC CNNs [178]. Q. Ha et. al detail their work on skin lesion segmentation that achieved 1st place in the 2020 SIIM-ISIC melanoma classification challenge [179]. Their method involved using an ensemble-based model that averages the pixel prediction scores of multiple models using various versions of EfficientNet, SE-ResNeXt and ResNeSt as the network backbone. Also included was a thorough image augmentation pipeline. Additionally, the ISIC skin lesion segmentation challenge [180] has run every year from 2016 until 2020 so there is a large back-catalogue of skin lesion segmentation methods all trained and tested on a standardised dataset. A detailed review of the skin lesion segmentation literature can be found in [181] which summarises 356 publications on skin lesion segmentation and 238 on skin lesion classification published between 2011

and 2022.

Chapter 4

Predicting Lung Toxicity After Radiotherapy From Pre-Treatment CT scans and Dose Maps

This chapter is focused on improving methods for radiation pulmonary toxicity prediction using radiotherapy planning pre-treatment information with a focus on CT image information. Section 4.1 discusses the motivations for this work, Section 4.2 defines the datasets and methods, Section 4.3 contains the results and Sections 4.4 and 4.5 give the discussions and conclusions. The methods were applied to two separate datasets, one containing standard lung cancer patients receiving IMRT and one containing lung cancer patients with ILD receiving SABR.

4.1 Introduction

Radiation-induced pulmonary toxicity such as radiation pneumonitis, an inflammation of the lung tissue, is a common and potentially life-threatening group of toxicity experienced by patients receiving a lung dose during their treatment. The ability to predict if a patient will develop pulmonary toxicity after receiving radiation therapy would help the production of more personalised treatment plans and allow clinicians to monitor high-risk patients. As discussed in Section 3.1, prediction can be partially achieved

through the use of dose and clinical features which is the most advanced information that is currently used clinically. The main focus of this chapter is to include CT image information in the predictive models. Current standard clinical RT practice includes capturing a CT image to localise anatomy and calculate the attenuation maps necessary for RT planning. This planning CT image is currently not used beyond these purposes. The work in this chapter applies methods for the extraction of features from the lung region of RT planning CT scans using both standard radiomic approaches and the utilisation of a CNN pre-trained for lung segmentation. These CT features were also combined with dose and clinical features to gain a large improvement to pulmonary toxicity prediction performance when compared to the clinical baseline method. Additionally, the value of combining both radiomic and the pre-trained CNN method for feature extraction is investigated. Besides predicting pulmonary toxicity, the analysis of the SABR ILD dataset also explored additional outcomes for prediction, such as the FACT-L and EQ-5D-5L scores.

4.1.0.1 Contributions of this Chapter

In the work detailed in this chapter, methods for pulmonary toxicity prediction from RT dose maps are further developed using data from the ASPIRE-ILD clinical trial [182] and data from the Edinburgh Cancer Centre. The technical contributions are:

- Development of a radiomic workflow for radiation pulmonary toxicity prediction.
- The first use of a pre-trained UNet segmentation model for the extraction of CT Image features for the purpose of toxicity prediction.
- The first prediction models for ILD patients receiving SABR using dose, clinical and CT image features for the prediction of pulmonary toxicity, FACT-L and EQ-5D-5L.

4.2 Method

4.2.1 Datasets

Two datasets were used in this chapter for the prediction of pulmonary toxicity which were the Edinburgh Pneumonitis dataset and the ASPIRE-ILD dataset. Both of these datasets use separate RT methods and the ASPIRE-ILD dataset specifically contains

only patients with ILD. As a result, these two datasets cannot be merged into one larger dataset and must be studied separately, as was done in this chapter. The ASPIRE-ILD dataset also includes additional patient outcomes, such as their FACT-L scores, which were used to train further prediction models.

4.2.1.1 Edinburgh Pneumonitis Dataset

The Edinburgh Pneumonitis dataset consists of retrospectively collected data from patients with NSCLC that were treated with IMRT at the Edinburgh Cancer Centre, Western General Hospital, Edinburgh between 01/01/2009 and 01/10/2010. During data collection, the occurrence of pneumonitis in each patient was determined from that patient's medical records by experienced clinicians. Any patients with unclear medical notes regarding pneumonitis were not included. This generated a dataset of 66 patients, 12 of whom developed clinically validated grade 2 or above radiation pneumonitis after their RT treatment where grade 2 radiation pneumonitis corresponds to pneumonitis presenting symptoms where some form of clinical intervention has been performed as defined by the Common Criteria for Adverse Events (CTCAE) (v5.0) [183]. The RT data available included the patients' RT planning CT scans, the RT planning dose distribution maps and the anatomical segmentations produced during the RT planning. Examples of a planning CT and dose map segmented to the lung volume, defined during RT planning, are shown in Figure 4.1. The CT scans all had voxel dimensions of 1.0x1.0x3.0mm and slices in the transverse plane had a resolution of 512x512. Also available were some of the patients' clinical parameters such as their age. Of the 66 patients in the study, IMRT was applied to; 52 patients on a Varian Clinac 600C/D, 3 patients on a Varian Clinac iX and 9 patients on a Varian Clinac 21EX. All radiotherapy plans were calculated using the Varian Pencil Beam Convolution algorithm (v8.1.2).

4.2.1.2 ASPIRE-ILD Dataset

ASPIRE-ILD is a completed phase 2 clinical trial that investigated the use of SABR as a treatment for lung cancer in patients with ILD that are not able to undergo surgery [182]. The ASPIRE-ILD study recruited a total of 42 patients from 5 institutions in Canada and 1 in Scotland. Of these 42 patients, 39 underwent treatment with SABR and their progression was followed up for at least two years.

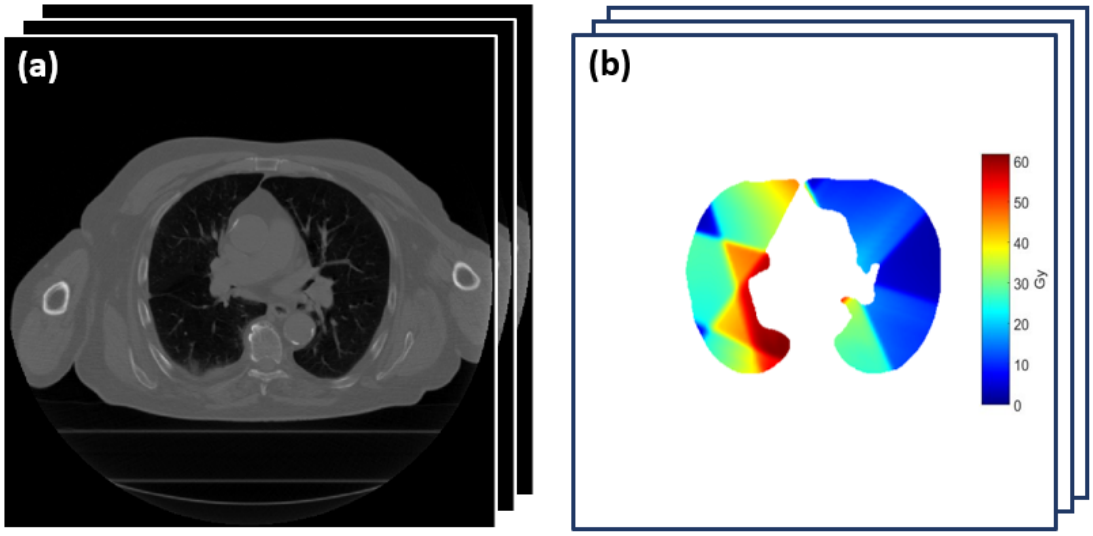


Figure 4.1: *Examples of* ***(a)*** *a 3D planning CT image stack and* ***(b)*** *a 3D dose map of only the lung volume.*

The data available from the ASPIRE-ILD study includes baseline CT scans, planning CT scans, RT dose maps, RT structure sets, patient clinical information and patient outcomes. The CT baseline scans are high quality diagnostic CT scans that were used to determine if a patient was appropriate for enrolment in the ASPIRE-ILD study. These are breath-hold CT scans with a high resolution making them ideal for radiomic analysis. The planning CT scans, however, are of lower quality and have inconsistent reconstructions; some are depicted as maximum intensity projections of multiple gated images, rendering them less suitable for radiomic analysis. For this reason, only the diagnostic CT scans from the ASPIRE-ILD dataset were investigated for radiomic analysis. The range of voxel resolutions of the baseline CT scans for the ASPIRE-ILD study are displayed in the box plot shown in Figure 4.2.

The ASPIRE-ILD study recorded multiple patient outcome metrics at intervals of 3, 6, 9, 12, 18 and 24 months post-RT to monitor the patients' progression. The outcomes relevant to this outcome prediction study are displayed in Table 4.1. The FACT-Lung B1 dyspnea outcome here is a sub outcome of the FACT-L index relating to only question B1 of the FACT-L questionnaire which quantifies patient dyspnea. The outcomes at the 3-month mark were utilised, as the number of patient outcomes available decreased with time post-RT.

For the CTCAE toxicity outcome, any toxicity, as defined by CTCAE v5 [183], that the patient developed during or after their radiotherapy treatment was recorded with

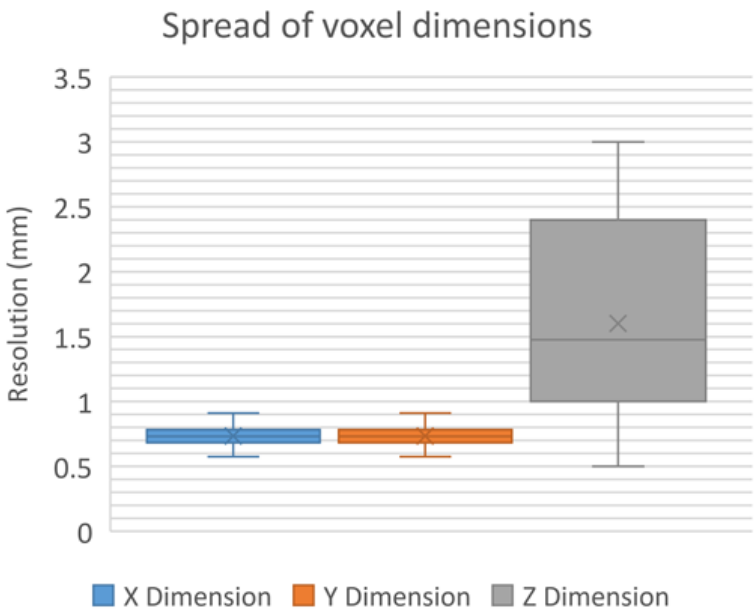


Figure 4.2: *The range of image dimensions of the baseline CT images for the ASPIRE-ILS dataset.*

Outcome	Number of Patients (at 3 months)
CTCAE Toxicity Grades	39
FACT-Lung	34
EQ-5D-5L	35
Overall Survival	39
Cough Severity	35
FACT-Lung B1 Dyspnea	34

Table 4.1: *The ASPIRE-ILD outcomes used in this prediction study.*

the exact CTCAE grade. It was also recorded if these toxicities were RT treatment-related or unrelated. For this outcome prediction study, only the RT-related pulmonary toxicities were considered. The RT-related pulmonary toxicities and the number of patients developing each toxicity are given in Table 4.2.

4.2.2 Dose and Clinical Features

Dose and clinical features have previously been shown to have predictive power for pulmonary toxicity prediction post-RT both on their own as well as to improve the performance of models based on additional features from separate sources [96, 103, 107, 108, 116, 117, 118, 119, 120, 121, 122]. For these reasons, as well as the fact that dose and clinical features will always be available pre-RT, they should always be

Adverse Event	Number of Patents with AE Grade ≥2
Bronchopulmonar Hemorage	1
Cough	0
Dyspnea	7
Lung Infection	1
Pleural Effussion	1
Pneumonitis	3
Pulmonary Edema	0
Pulmonary Fibrosis	0
Respiratory Failure	1
All Pulmonary (max grade per patient)	11

Table 4.2: *The number of RT related pulmonary adverse events recorded in the ASPIRE-ILD study.*

included in radiation toxicity prediction studies as has been done here.

4.2.2.1 Clinical Features

The clinical features available in both datasets differ and are outlined here. For the Edinburgh Pneumonitis dataset, clinical features were retrospectively extracted from the patient's medical information based on data that was deemed to be relevant to the classification task. The clinical features that were included were chosen based on their availability for all of the patients in the dataset as well as the ability to present these features numerically. For the ASPIRE-ILD dataset, clinical features were prospectively collected during the course of the original trial. The clinical features available for both datasets are displayed in Table 4.3.

4.2.2.2 Dose Features

Dose volume histogram features were calculated from the RT planning dose map, segmented to only include the whole lung volume, excluding the tumour region defined by the gross tumour volume (GTV). The lung volume and GTV were available from the RT planning data. An example of a dose map thresholded to only the lung region is given The calculated dose features were the same for both datasets. The dose volume histogram features calculated were; the mean lung dose (MLD), R_{50}, D_{2cm} and the V_x from V_5 to V_{70} in steps of 5 Gy presented as a percentage and an absolute volume. The features and methods here were the same for both datasets.

Edinburgh Pneumonitis	ASPIRE-ILD
Age	Age
Smoking status	Symptoms
COPD status	Smoking status
Chemotherapy status	Smoking pack years
T-stage	Forced vital capacity (FVC) test
N-stage	Lung diffusion test (DLCO)
Primary tumour volume	Clinically observed radiological lung pattern
RT fractions	Clinical consensus diagnosis
Prescribed maximum dose	ILD subtype
	Gender
	ILD-GAP index
	PET SUV max
	T-stage
	ECOG Performance Status
	Forced Expiratory Volume (FEV) score
	Prescribed maximum dose.

Table 4.3: *The clinical features available for the Edinburgh Pneumonitis dataset and the ASPIRE-ILD dataset.*

4.2.3 Radiomic CT Features

Prior to radiomic feature calculation, CT image pre-processing steps were applied and the volume of interest, which defined the voxels that radiomic features were calculated for, must be established. The lung volume, segmented using the contours available from the RT planning process, was used as the initial volume of interest. This was then thresholded to only include voxels with Hounsfield unit (HU) values ranging from -1000 HU to -400 HU to specifically target the soft lung tissue by removing any vessels and the tumour volume. Note that the CT image was not thresholded, the thresholding was only used to define the volume of interest. Before the radiomic features were calculated, the CT images were isotropically resampled to have voxel dimensions of 1.5x1.5x1.5mm. Quantisation was applied to the image to reduce the number of grey levels with a bin width of 25 HU being used producing 16 grey levels in the defined volume of interest. This is the standard quantisation bin width used in the PyRadiomics package [49] and has been successfully applied in other studies investigating prediction models from lung radiomic features [184, 185]. Radiomic features were computed from the defined volume of interest within CT images in Python using the PyRadiomics package (v3.1.0) [49]. All features from the following classes were included; first-order, GLCM, GLRLM, GLSZM, NGTDM and GLDM. These feature classes were discussed previ-

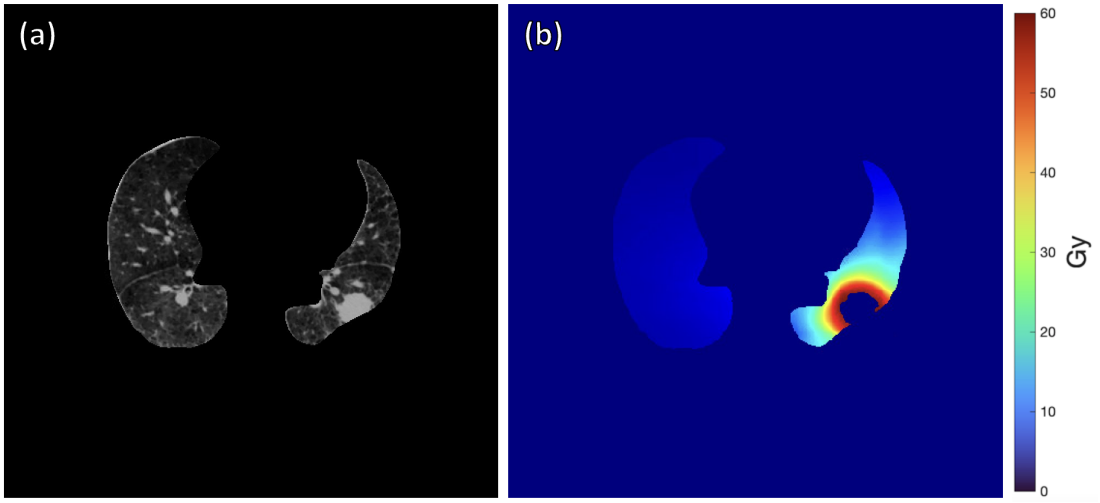


Figure 4.3: ***(a)*** *A dose image from the original dataset without any preprocessing steps,* ***(b)*** *the same dose slice as (a) that has been pre-processed so that it is aligned with its complementary CT image and all the non-lung anatomy has had its dose values set to 0. Dose values are given in Gy.*

ously in Section 2.3.2. The features calculated by PyRadiomics conform to the Image Biomarker Standardisation Initiative (IBSI) recommendations [54] and, where possible, the radiomic features were calculated in 3D. Further information on these metrics can be found in the PyRadiomics documentation (`https://pyradiomics.readthedocs.io/`). This process produced a feature table comprising 94 distinct radiomic features, each calculated from the CT lung region for every patient.

4.2.3.1 Dose Thresholding the CT Image

Up to this point, the image texture features have been calculated for the entire lung volume of each patient with no regard as to where the dose has been applied. As a patient who goes on to develop RP will generally develop the condition within the volume of the lung that has received a higher dose, it may be beneficial to calculate the image texture just within these regions of higher dose. This is investigated in this section only on the Edinburgh Pneumonitis dataset as the ASPIRE-ILD diagnostic CT images that are used for radiomic analysis are not registered to the dose maps.

The dose images that have been pre-processed as described in Section 4.2.2.2 were thresholded to only include pixels with an intensity value over a set threshold meaning only regions corresponding to a dose over a set level of Grays are kept. The threshold value can be changed depending on what Gray level is being investigated for the

thresholding. From this thresholded dose image, a binary mask is produced by setting all non-zero pixels in this thresholded dose image to a pixel intensity value of 1. This binary mask was used to define the region of interest in the 3D CT image to be used for the calculation of radiomic features. This produces radiomic features that are calculated only from regions of the lung above a set dose level. This process is shown in Figure 4.4. The same methods described at the start of Section 4.2.3 are used to calculate the image texture features. Radiomic features were calculated for the 3D CT images using dose thresholding levels of 5, 10, 15, 20 and 25 Gy producing 5 sets of radiomic features for the whole patient cohort.

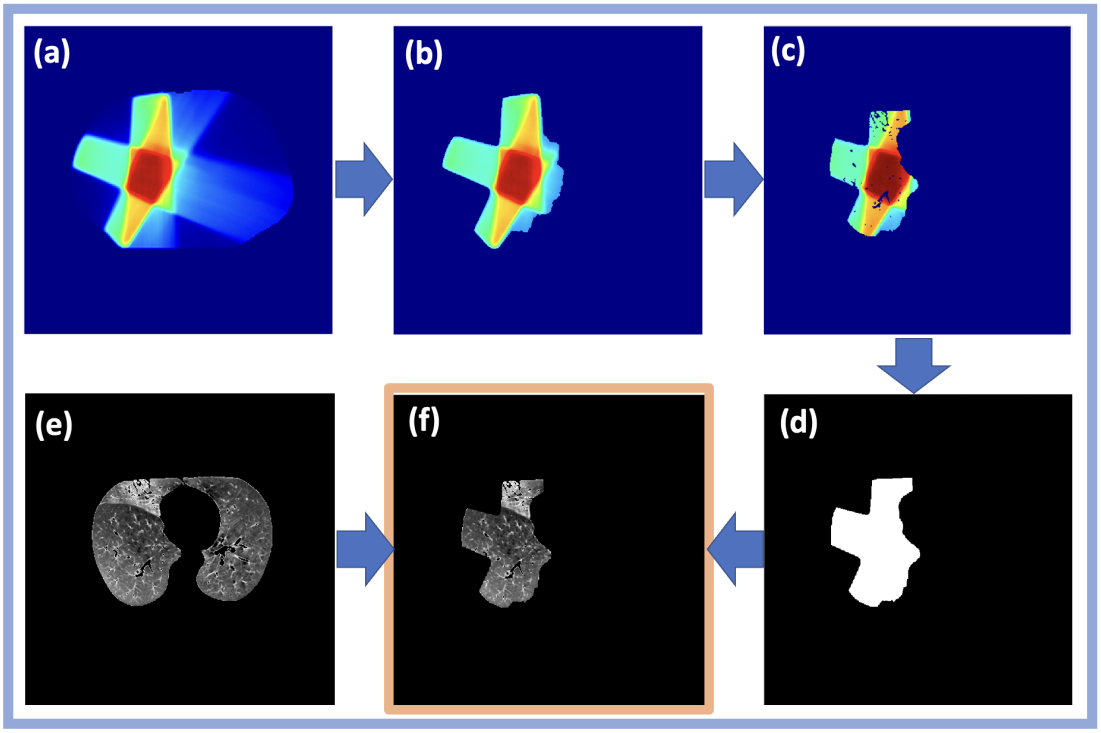


Figure 4.4: *The CT dose thresholding workflow.* ***(a)*** *An original dose image* ***(b)*** *the dose image thresholded to only include dose areas above 20Gy,* ***(c)*** *dose image further thresholded to only include dose areas that lie within the lung volume,* ***(d)*** *the binary mask calculated from the dose that is applied to threshold the CT image,* ***(e)*** *the CT slice corresponding to dose slice (a) that has been pre processed as described in Section 4.2.3,* ***(f)*** *the CT image where only the area within the lung receiving greater than 20Gy of dose are included.*

4.2.4 UNet CT Features

In addition to the PyRadiomics CT features, a publicly available pre-trained UNet segmentation model was leveraged as a feature extractor. The reason that a segmentation

model was chosen as opposed to a classification model is that the features learned by a segmentation model, trained on a dataset containing a variety of anatomies and medical conditions, are likely to be more generalisable than features from a classification model trained for a specific medical task. Additionally, there are more publicly available medical image segmentation models available than classification models so there is a larger pool of previous work that can be accessed.

The UNet model, used as our feature extractor, was originally trained for segmenting lungs in 2D CT images, this network is referred to as "UNet (R231)" [186] and is available from *https://github.com/JoHof/lungmask*. Details of the model's architecture and training can be found in the original paper [186]. Prior to extracting features from the CT image, no additional training of the model was performed. This approach of using a pre-trained model with no additional training was chosen due to the small size of our datasets, especially when considering the low number of positive cases, and the lack of public data available with radiation toxicity outcomes meaning training a new supervised CNN classification model was not possible. The activations from the convolutional layer before the final convolutional layer making the pixel-wise classifications were used as the features. This layer has a size of 256x256x64 meaning 64 features were extracted for every pixel. This output feature layer was chosen for feature extraction, as opposed to earlier feature layers, as our classification task is similar to the original segmentation task of the CNN as they share the same modality and anatomical focus. This means that the dimensionality reduction in the UNet model, taking place before the output feature layer, is likely to still be appropriate to our task. If using a model that had been trained on an unrelated task, then earlier, and likely more generalisable features would be a more appropriate choice. The pixel-wise features were converted to global features for the classification by averaging them over the spatial dimensions using a global average pooling layer added to the model so that, for each slice of the 3D CT image, 64 features were calculated. The location of this global average pooling layer in the model architecture and the convolutional layer used for feature extraction is shown in Figure 4.5. This feature averaging was done for every transverse slice of the 3D CT image containing any lung area. The features were then averaged in the z-direction producing 64 features from the 3D CT image for every patient.

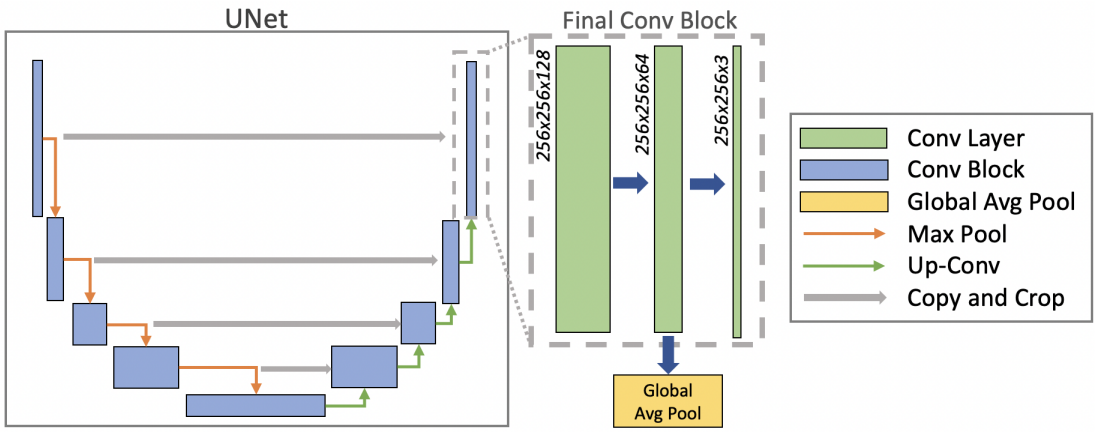


Figure 4.5: *Location of the global average pooling layer, used to extract features from the second last convolutional layer of the UNet model. Further details of this UNet model can be found in the paper that originally trained the model [186].*

4.2.4.1 Fine Tuning with MEDGIFT-ILD Dataset for Better Feature Extraction of ILD Patients

Although the UNet R231 model underwent initial training on a diverse dataset [186], there was no deliberate effort to encompass examples representing all ILD subtypes during this initial training phase. In an attempt to improve the performance of the model as a feature extractor when dealing with patients with ILD, the model was further trained for the task of segmenting the lung volume of CT scans of patients with ILD. To do this, the MEDGIFT-ILD dataset was used [8]. The MEDGIFT-ILD dataset includes CT scans with lung segmentations of 108 patients with ILD. These were used to train the UNet R231 model for lung segmentation. As the end goal is a feature extractor for the ASPIRE-ILD data, all of the MEDGIFT-ILD data can be used in the training set. The UNet R231 model was fine-tuned for ILD patient data by re-training it for 20 epochs with the MEDGIFT-ILD data using the adam optimiser, an initial learning rate of 1e-3, a learning rate drop period of 5 and a learning rate drop factor of 0.2. This re-trained version of the UNet R231 model was used for feature extraction from the ASPIRE-ILD CT scans.

4.2.4.2 UNet Feature Extraction With Masking

In addition to the global averaging method for converting the pixel-wise features into features for the full CT stack, region of interest (RoI) averaging can be applied to take the features from only a specific part of the anatomy. This involves simply averaging

the features in the spatial dimensions but only including voxels that are within a certain RoI. It should be noted that due to the down-sampling and up-sampling involved in the encoder-decoder based architecture of UNet, voxels further than just the adjacent few voxels will contribute to the features of a particular voxel. This means that isolating the voxels using the RoI in this way does not truly isolate the features to contain only information from the voxels within the RoI.

The RoIs that were used here were the Lung RoI, the RoI of the patient's body excluding the lungs and the background voxels, i.e. the voxels containing no biological tissue. The lung RoI used was the one defined during the RT treatment planning and available in the RT structure set, an example of this is shown in Figure 4.6 (a). To define the body excluding the lung RoI, a binary mask is defined by thresholding the CT image to only include voxels with an intensity over -500 HU, this excludes almost all of the background but the CT scanner bed remains. To remove the bed, a morphological opening is performed with a disk shaped structuring element with a radius of 5 pixels. The image is then filled to remove any holes and the lung mask is subtracted from the body mask leaving the final body mask with the lung excluded, this is shown in Figure 4.6 (b). The background mask is simply all of the voxels not included in the lung or body masks, this is shown in 4.6 (c).

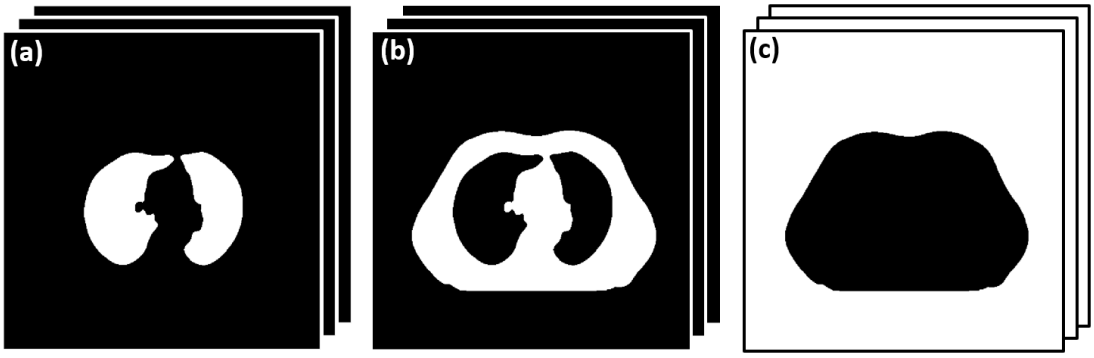


Figure 4.6: *Examples of the masks produced for* ***(a)*** *the lungs,* ***(b)*** *all of the body excluding the lungs and* ***(c)*** *the background.*

To show which regions of anatomy are being isolated for each mask, the masks have been applied to the original image and displayed in Figure 4.7. Note that this is just shown to clarify the anatomy within each RoI as the CT images were not masked before being applied to the UNet model for feature extraction. In Figure 4.7 (a) and (b) the lung and body regions are clearly visible and Figure 4.7 (c) shows the background where all that is visible is the CT scanner bed and a faint outline of the body.

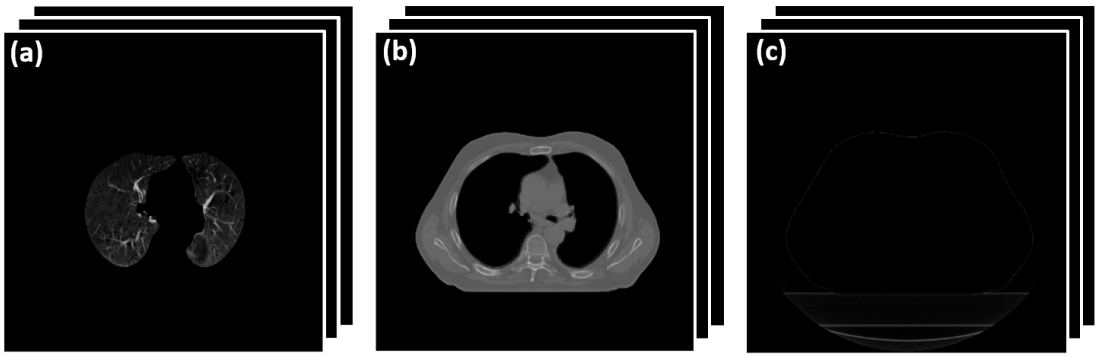


Figure 4.7: *Examples a CT image that has been masked to only include* ***(a)*** *the lungs,* ***(b)*** *all of the body excluding the lungs and* ***(c)*** *the background.*

4.2.5 Prediction Models

After processing and calculating the dose, clinical, CT PyRadiomic and CT UNet features, models can be fit to predict the outcomes of the Edinburgh pneumonitis and the ASPIRE-ILD dataset. The first step for all models is to normalise the by converting them to z-scores. For the prediction of all outcomes, boosted decision tree based algorithms were applied. For CTCAE grade predictions, which were available for both datasets, a binary classification was performed using the AdaBoost algorithm. For all other outcomes of the ASPIRE-ILD dataset, the LSBoost algorithm was used as it is more appropriate than the AddBoost algorithm due to the continuous nature of these outcomes. Details of the training of these specific algorithms are available in the following subsections. The details in this subsection apply to all models.

In order to determine the contributions of the different features to the overall model performance, the training and testing scheme was repeated on different combinations of the four feature subsets; dose, clinical, PyRadiomic and UNet.

It can often benefit model performance by reducing overfitting to remove non-important features from the input feature set. To determine the important features, an identical model to the final model was trained using an elevated learning rate (LR = 1.00) and a lower number of learning cycles (LC = 50). The feature importance was estimated from this model by summing the changes in the node error due to the tree splits for every predictor. Here the change in the node error is the difference between the error for the parent node and the total error for the two child nodes. The ten most important features were selected and the final models were trained, these features are given in appendix A for each model. This is summarised in Figure 4.8.

Figure 4.8: *The workflow for the feature selection aspect of the boosted decision tree model training.*

Training and testing were applied using a leave-one-out cross-validation (LOOCV) scheme. This involves training the model on all the data except one sample and then testing the model on the left-out sample to get a prediction. The trained model is then discarded and a different patient is selected as the one to be left out from training and the process repeats until testing has been completed on all of the patients. This is the same as k-fold cross-validation where k equals the dataset size. Figure 4.9 details this cross-validation scheme.

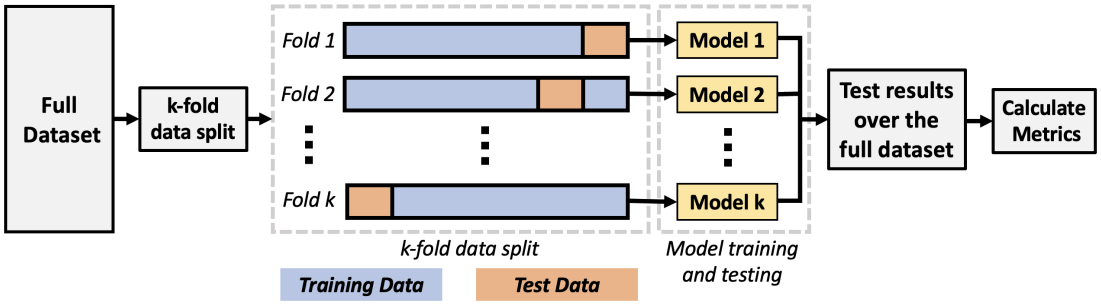


Figure 4.9: *Workflow for the LOOCV (k-fold cross-validation) training strategy.*

4.2.5.1 AdaBoost for CTCAE Toxicity Prediction

For the classification of CTCAE toxicity grades, boosted decision trees were used with the AdaBoost algorithm [187] applied as a binary classifier in MATLAB (R2023a). For the Edinburgh pneumonitis dataset, the ground truth available is a binary classification defined by the patient's clinical notes with the classes $<$ *grade 2* and $\geq$ *grade 2* pneumonitis. For the ASPIRE-ILD data, exact CTCAE grades were available for all pulmonary toxicities as given in Table 4.1. These exact grades were converted to binary classifications with the grade cutoff as $\geq$ *grade* 2 for the high-risk group. As both datasets are unbalanced, an increased misclassification cost was applied during training to the underrepresented class, in both cases$\geq$ *grade* 2 class, based on its proportional size in the dataset. Due to the small size of the dataset, it was not

possible to perform hyperparameter tuning of the model as this would require another hold-out set to validate the parameter tuning. The parameters used during the training of the AdaBoost model are displayed in Table 4.4.

Training Parameter	Value
Learning rate	0.1
Learning cycles	200
Maximal number of decision splits	10
Minimum observations per leaf	1
Minimum observations per branch node	2

Table 4.4: *The training parameters used for the AdaBoost prediction models.*

4.2.5.2 LSBoost for FACT-L and EQ-5D-5L Prediction

The FACT-L and EQ-5D-5L index, which were only available for the ASPIRE-ILD data, are continuous scores in the range of 0 to 136 for the FACT-L overall score and -0.224 to 1 for the EQ-5D-5L score. As these are continuous variables, a regression-based training scheme, as opposed to a binary classifier, can be applied. For this, the LSBoost algorithm for boosted decision trees was applied to predict the exact score for the FACT-L and EQ-5D-5L metrics. The parameters used during the training of the AdaBoost model are displayed in Table 4.5. The recorded outcome here when considering the prediction of these continuous variables is the mean average error (MAE) of grade predictions.

Training Parameter	Value
Learning rate	0.1
Learning cycles	200
Maximal number of decision splits	10
Minimum observations per leaf	5
Minimum observations per branch node	10

Table 4.5: *The training parameters used for the LSBoost prediction models.*

In addition to predicting the exact scores, a binary classification can be made by grouping the patient cohort into low and high-risk groups. The minimum clinically important difference of these two was used to define the high and low risk groups. A patient whose score changes by more than the minimum clinically important difference for each metric will be placed in the high risk group. The minimum clinically impor-

tant difference EQ-5D-5L has been reported to be 0.03-0.05 for diabetes [188], 0.051 for COPD [189], 0.078 to 0.095 in fibrotic ILD [190]. From this, the value 0.05 is used to define the minimum clinically important difference for the EQ-5D-5L binary classification. For the FACT-L scale, clinically relevant change scores have been estimated to be two to three points for the lung cancer subscale and five to seven points for the Trial Outcome Index aspects of the FACT-L score [191]. A minimum clinically important difference of 6 was taken for the FACT-L binary classification. For the binary classification, no new model was trained. The output of the LSBoost model for exact grade prediction was simply used and compared to the baseline scores. The ground truth for this binary classification was calculated by taking the change between the baseline score and the score at 3 months. The AUC, sensitivity, specificity and accuracy were calculated as the performance metrics.

4.3 Results

The metrics produced during the LOOCV testing scheme are the accuracy, sensitivity, specificity and the AUC ROC with the AUC ROC being the most descriptive of the overall predictive power of the models and most commonly applied metric in the recent radiation toxicity prediction literature [107, 108, 109, 110, 120, 121, 122, 123]. For the FACT-L and EQ-5D-5L scores, the MAE was also calculated to determine the performance of the models as exact score predictors. For all results, feature reduction was used so for every model the final number of features used is 10.

4.3.1 Edinburgh Pneumonitis Dataset Results

The only outcome available for the Edinburgh pneumonitis dataset was if the patients developed grade 2 or above pneumonitis. Results for predicting this are presented here.

4.3.1.1 Dose Thresholding the CT Image

The results of the dose thresholding to determine the CT volume of interest for the radiomic feature calculation are given in Table 4.6. The best performing model used no dose thresholding and the predictive performance decreased at higher dose thresholds, therefore a smaller lung sub-volume, was used. For this reason, no further dose thresholding experiments were applied and only radiomic features for the whole CT images

were considered.

Features	CT Dose Threshold	Acc (%)	Sens (%)	Spec (%)	AUC
PyRad	0 Gy	**72.7**	**66.7**	**74.1**	**0.760**
PyRad	5 Gy	71.2	**66.7**	72.2	0.713
PyRad	10 Gy	68.2	**66.7**	68.5	0.726
PyRad	15 Gy	69.7	58.3	72.2	0.662
PyRad	20 Gy	62.1	58.3	63.0	0.609
PyRad	25 Gy	62.1	50.0	64.8	0.617

Table 4.6: *Results from the LOOCV over the whole Edinburgh pneumonitis dataset (N=66) for the binary prediction of grade $\geq$ 2 pneumonitis with different levels of dose thresholding to determine the PyRadiomics RoI.*

4.3.1.2 UNet Masking Results

The results of applying masking to the UNet features are given in Figure 4.7. The best performing model here uses no masking and averages the features from the whole image. For this reason, in future results, when the UNet features are used, no masking was applied.

Features	RoI	Acc (%)	Sens (%)	Spec (%)	AUC
UNet	Lung only	62.1	**66.7**	61.1	0.627
UNet	Body only	63.6	**66.7**	63.0	0.652
UNet	Background only	63.6	50	66.7	0.607
UNet	Whole Image	**66.7**	58.3	**68.5**	**0.717**

Table 4.7: *LOOCV test results for the prediction of pneumonitis on the Edinburgh pneumonitis dataset with region of interest masking applied to the UNet feature pooling.*

4.3.1.3 Final Pneumonitis Prediction Results

The results for the binary prediction of grade ≥ 2 pneumonitis using LOOCV with the AdaBoost algorithm applied to different feature subsets are given here in Table 4.8. The best performing model used all of the feature subsets and achieved an AUC of 0.836 with a sensitivity and specificity of 75.0% and 81.5% respectively. This model correctly classified 9/19 patients as high-risk and 43/47 patients as low-risk giving a precision and negative predictive value of 47.4% and 93.6% respectively. ROC curves for some of the models are presented in Figure 4.10.

Features	Acc (%)	Sens (%)	Spec (%)	AUC
Dose	63.6	41.7	68.5	0.640
Clin	53.0	33.3	57.4	0.465
Dose, Clin	63.6	41.7	68.5	0.633
PyRad	72.7	66.7	74.1	0.760
UNet	66.7	58.3	68.5	0.717
Dose, Clin, PyRad	77.3	**75.0**	77.7	0.790
Dose, Clin, UNet	78.8	**75.0**	79.6	0.814
Dose, Clin, PyRad, UNet	**80.3**	**75.0**	**81.5**	**0.836**

Table 4.8: *Results from the LOOCV over the whole Edinburgh pneumonitis dataset (N=66) for the binary prediction of grade ≥ 2 pneumonitis.*

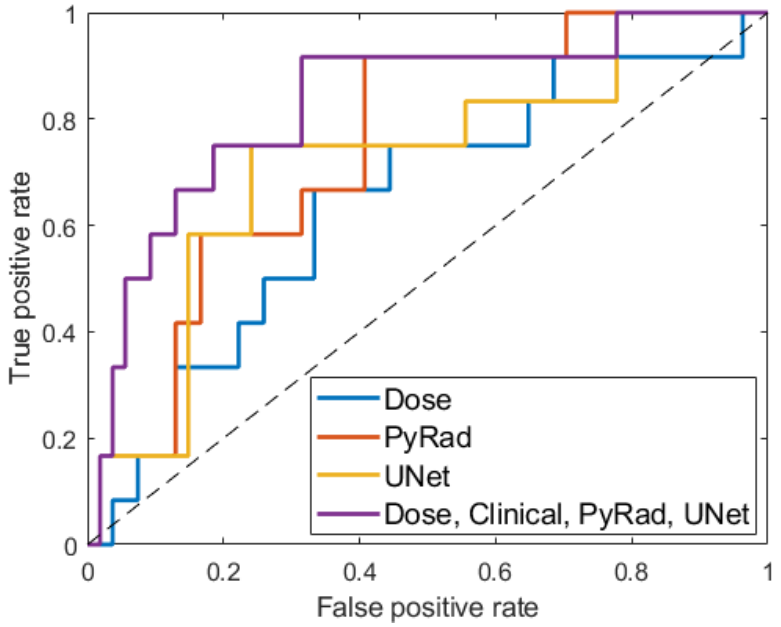


Figure 4.10: *ROC curves for the binary prediction of grade ≥ 2 pulmonary toxicity on the Edinburgh pneumonitis dataset using select feature subsets.*

4.3.2 ASPIRE-ILD Dataset Results

The ASPIRE-ILD dataset had several different prediction outcomes available. The results for training the prediction models using LOOCV for the CTCAE pulmonary toxicity, FACT-L index, EQ-5D-5L score, overall survival, cough index and FACT-L B1 dyspnea question are presented in the following subsections. Feature importance is presented in appendix A.

4.3.2.1 CTCAE Pulmonary Toxicity Prediction Results

The main prediction outcome of interest for the ASPIRE-ILD dataset was the occurrence of pulmonary toxicity. This differs partially from the Edinburgh pneumonitis dataset as with the ASPIRE-ILD dataset we are aiming to predict all pulmonary toxicities as listed in Table 4.2. Of the 39 patients, 11 developed grade ≥ 2 CTCAE v5 pulmonary toxicity with dyspnea (n=7) and pneumonitis (n=3) being the predominant toxicities. The best-performing model included dose, CT PyRadiomic and CT UNet features to achieve an AUC of 0.841 with a sensitivity of 81.8% and a specificity of 78.6%. This model correctly classified 9/15 patients as high-risk and 22/24 patients as low-risk giving a precision and negative predictive value of 60% and 92% respectively. Table 4.9 displays all model results and Figure 4.11 displays the ROC curve for some of the feature subsets.

Features	**Acc (%)**	**Sens (%)**	**Spec (%)**	**AUC**
Dose	56.4	81.8	46.4	0.687
Clinical	66.7	18.2	85.7	0.444
Dose, Clin	53.8	81.8	42.9	0.664
PyRad	51.3	45.5	53.6	0.515
UNet	51.3	36.4	64.3	0.479
Dose, PyRad	76.9	**81.8**	71.4	0.746
Dose, UNet	64.1	**81.8**	57.1	0.744
PyRad, UNet, Dose, Clin	76.9	72.7	**78.6**	0.834
Dose, PyRad, UNet	**79.5**	**81.8**	**78.6**	**0.841**

Table 4.9: *Results from the LOOCV over the whole ASPIRE-ILD dataset for the binary prediction of grade $\geq$ 2 pulmonary toxicity (N=39).*

4.3.2.2 FACT-L Prediction Results

The FACT-L prediction was analysed in terms of both a binary classification and regression where the prediction outcome was the exact FACT-L score. The metrics presented for the binary classification are the AUC, accuracy, sensitivity and specificity and for the exact score prediction, the MAE was used to analyse the performance. For the binary classification, a change in the FACT-L score of $\leq$ -6 from the baseline pre-RT score was used to define the high-risk group. The ground truth placed 22 patients into the high-risk group and 12 patients into the low-risk group out of the total N=34 patients who had 3-month follow-up information available. In terms of this binary

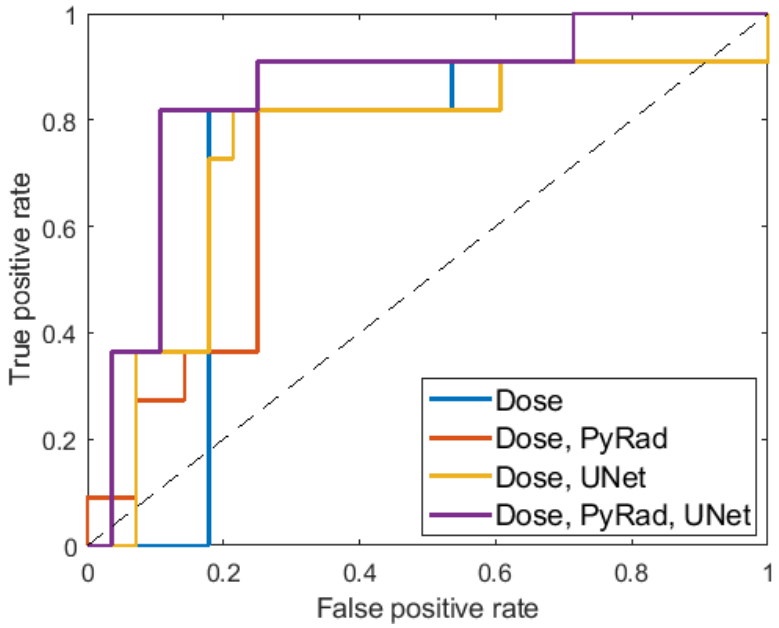


Figure 4.11: *ROC curves for the binary prediction of grade ≥ 2 pulmonary toxicity on the ASPIRE-ILD data using select feature subsets.*

classification, the best performing model used only dose based features and achieved an AUC of 0.769 with a sensitivity and specificity of 63.6% and 83.3% respectively. This model correctly classified 14/16 patients as high-risk and 10/18 patients as low-risk giving a precision and negative predictive value of 87.5% and 55.6% respectively. The best performing model in terms of predicting the exact FACT-L score was used the dose, clinical, PyRadiomic and UNet features which achieved a MAE of 13.6. Table 4.10 presents all model results and Figure 4.12 presents a plot of the true vs predicted FACT-L scores when using the dose, clin, PyRadiomic and UNet features.

Features	**Acc(%)**	**Sens(%)**	**Spec (%)**	**AUC**	**MAE**
Dose	**70.6**	**63.6**	**83.3**	**0.769**	18.7
Clinical	50.0	40.9	66.7	0.652	19.5
Dose, Clin	**70.6**	**63.6**	**83.3**	0.758	17.9
PyRad	50.0	36.4	75.0	0.688	16.8
UNet	55.9	45.5	75.0	0.652	20.3
Dose, Clin, PyRad	55.9	45.5	75.0	0.686	18.2
Dose, Clin, UNet	58.8	50.0	75.0	0.705	18.5
Dose, Clin, PyRad, UNet	61.8	54.5	75.0	0.667	**13.6**

Table 4.10: *Results from the LOOCV over the whole ASPIRE-ILD dataset for the binary prediction of a change of -6 or less to the FACT-L score as well as the exact FACT-L score (N=34).*

A possible clinical scenario could occur where a patients FACT-L score is known

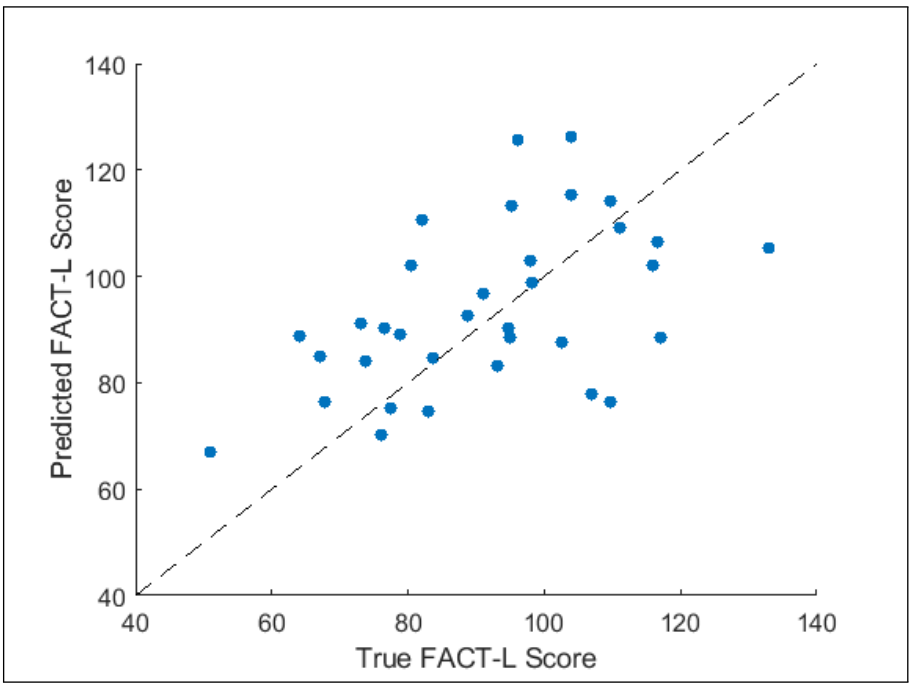


Figure 4.12: *Plot of the predicted versus true values for the FACT-L scores when using Dose, Clin, PyRad, UNet. The line of ideal prediction is displayed as a dashed line.* ***MAE = 13.6****.*

before applying RT. In this case, the baseline FACT-L score could be used as a feature of the model that is predicting the FACT-L score 3 months post-RT. The results for some of the feature subsets when including the baseline FACT-L as a predictive feature are given in Table 4.11. The best performing model for both binary classification and exact FACT-L score prediction is the model using only dose features which achieved an AUC of 0.821 with a sensitivity and specificity of 68.2% and 83.3% and a MAE of 10.9.

Features	**Acc (%)**	**Sens (%)**	**Spec (%)**	**AUC**	**MAE**
Dose	**73.5**	**68.2**	**83.3**	**0.821**	**10.9**
Clin	61.8	54.5	75.0	0.694	13.2
PyRad	64.7	59.1	75.0	0.679	13.0
UNet	61.8	54.5	75.0	0.671	13.2
Dose, Clin, PyRad, UNet	67.6	63.6	75.0	0.726	13.2

Table 4.11: *Results from the LOOCV over the whole ASPIRE-ILD dataset for the binary prediction of a change of -6 or more to the FACT-L score as well as the exact FACT-L score when the baseline FACT-L score is included as a predictive feature (N=34).*

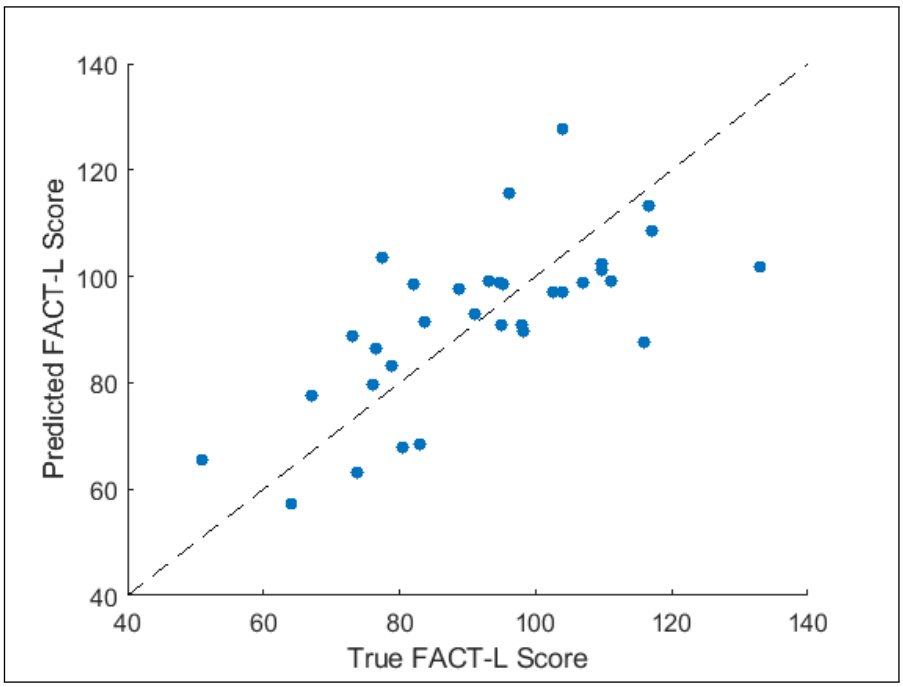


Figure 4.13: *Plot of the predicted versus true values for the FACT-L scores using only dose features and when the baseline FACT-L score is included as a predictive feature, the line of ideal prediction is displayed as a dashed line.* ***MAE = 10.9****.*

4.3.2.3 EQ-5D-5L Prediction Results

In the same manner as for the FACT-L prediction, the EQ-5D-5L prediction was analysed in terms of both a binary classification and a regression. For the binary classification, a change in the EQ-5D-5L score of $\leq$ -0.05 from the baseline pre-RT score was used to define the high risk group. For the ground truth, this placed 24 patients into the high-risk group and 11 patients into the low-risk group from the total N=35 patients who had 3-month follow-up information available. In terms of this binary classification, the best performing model used only PyRadiomic based features and achieved an AUC of 0.784 with a sensitivity and specificity of 53.3% and 85.0% respectively. This model correctly classified 13/15 patients as high-risk and 9/20 patients as low-risk giving a precision and negative predictive value of 86.7% and 45.0% respectively. The best performing model in terms of predicting the exact FACT-L score used the dose, clinical and PyRadiomic features which achieved a MAE of 0.139 but this only improved from PyRadiomics only model by 0.001. Table 4.12 presents all model results.

Again, in the same manner as for the FACT-L prediction, there exists a possible clinical scenario in which a patients EQ-5D-5L score is known before applying RT. The

Features	Acc (%)	Sens (%)	Spec (%)	AUC	MAE
Dose	51.4	41.7	72.7	0.633	0.228
Clin	57.1	45.8	**81.8**	0.663	0.260
Dose, Clin	51.4	41.7	72.7	0.644	0.259
PyRad	**62.9**	**54.2**	**81.8**	**0.784**	0.140
UNet	51.4	41.7	72.7	0.720	0.175
Dose, Clin, PyRad	57.1	45.8	**81.8**	0.765	**0.139**
Dose, Clin, UNet	48.6	37.5	72.7	0.629	0.193
Dose, Clin, PyRad, UNet	51.4	41.7	72.7	0.678	0.224

Table 4.12: *Results from the LOOCV over the whole ASPIRE-ILD dataset for the binary prediction of a change of -0.05 or more to the EQ-5D-5L score as well as the exact EQ-5D-5L score (N=35).*

baseline EQ-5D-5L score could then be used as a feature of the model that is predicting the EQ-5D-5L score 3 months post-RT. The results for some of the feature subsets when including the baseline EQ-5D-5L as a predictive feature are given in Table 4.13. For the EQ-5D-5L score prediction, all models perform worse when the baseline value is included.

Features	Acc (%)	Sens (%)	Spec (%)	AUC	MAE
Dose	42.9	33.3	63.6	0.572	0.214
Clin	54.3	45.8	72.7	0.686	0.197
PyRad	**60.0**	**54.2**	72.7	0.696	**0.171**
UNet	48.6	41.7	63.6	0.637	0.187
Dose, Clin, PyRad, UNet	**60.0**	50.0	**81.8**	**0.700**	0.204

Table 4.13: *Results from the LOOCV over the whole ASPIRE-ILD dataset for the binary prediction of a change of -6 or more to the EQ-5D-5L score as well as the exact EQ-5D-5L score when the baseline EQ-5D-5L score is included as a predictive feature (N=34).*

4.3.2.4 Overall Survival Prediction Results

Patients overall survival was available for all 39 patients of the ASPIRE-ILD trial. This was presented as months post-RT for survival (to one decimal place). For the binary classification, both 1-year and 2-year survival were taken as endpoints and the model was tested for both of these. This involved no retraining between models using the same features, the threshold that defined the binary classification was simply changed from 12 months to 24 months. For the regression, the MAE is presented in terms of months. The results for the overall survival prediction are presented in Table 4.14. A plot of the predicted overall survival against the true overall survival for the model

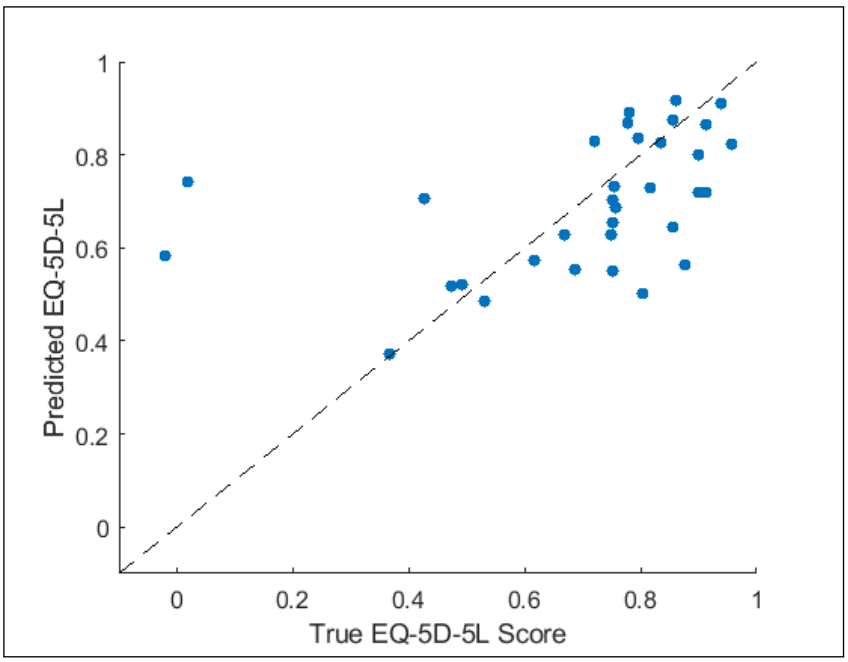


Figure 4.14: *Plot of the predicted versus true values for the EQ-5D-5L score when using Dose, Clinical and PyRadiomic features only. The line of ideal prediction is displayed as a dashed line.* ***MAE = 0.139****.*

using Dose, Clinical and UNet features is presented in Figure 4.15.

4.3.2.5 Cough Severity Prediction Results

The patient's cough severity was documented using the cough severity visual analogue scale (VAS) [192] prior to RT and after RT was applied at set time intervals. This severity scale ranges from 0 to 100 where 0 is no cough and 100 is the worst possible cough. The cough severity 3 months post-RT was predicted and the results are presented in Table 4.15. For the binary classification, any change > 5 in the cough severity scale was used as the classification threshold.

4.3.2.6 FACT-L B1 Dyspnea Prediction Results

Question B1 on the FACT-L questionnaire is directly related to the condition dyspnea and is a more specific outcome of the FACT-L scale. The question requires patients to respond to the phrase "I have been short of breath" with a score from 0 to 5 with 0 indicating "not at all" and 5 indicating "very much". The results for training the LSBoost model for the prediction of the FACT-L B1 question are given in Table 4.16.

1 Year Survival					
Features	**Acc(%)**	**Sens(%)**	**Spec (%)**	**AUC**	**MAE**
Dose	66.7	73.3	44.4	0.498	8.3
Clinical	**74.4**	76.7	**66.7**	**0.844**	8.1
Dose, Clin	66.7	73.3	44.4	0.633	8.3
PyRad	61.5	73.3	22.2	0.393	9.4
UNet	59.0	70.0	22.2	0.507	8.7
Dose, Clin, PyRad	71.8	76.7	55.6	0.659	9.7
Dose, Clin, UNet	71.8	**80.0**	44.4	0.774	**7.0**
Dose, Clin, PyRad, UNet	66.7	76.7	33.3	0.652	8.1
2 Year Survival					
Features	**Acc (%)**	**Sens (%)**	**Spec (%)**	**AUC**	**MAE**
Dose	64.1	0.0	80.6	0.498	8.3
Clinical	66.7	**25.0**	77.4	0.645	8.1
Dose, Clin	64.1	0.0	80.6	0.573	8.3
PyRad	61.5	0.0	77.4	0.407	9.4
UNet	64.1	0.0	80.6	0.407	8.7
PyRad, Dose, Clin	61.5	0.0	77.4	0.399	9.7
Dose, Clin, UNet	**74.4**	12.5	**90.3**	**0.706**	**7.0**
Dose, Clin, PyRad, UNet	**74.4**	**25.0**	87.1	0.520	8.1

Table 4.14: *Results from the LOOCV over the whole ASPIRE-ILD dataset for the overall survival prediction (N=39). Both one and two year survival were used as binary endpoints, the units of the MAE are months*

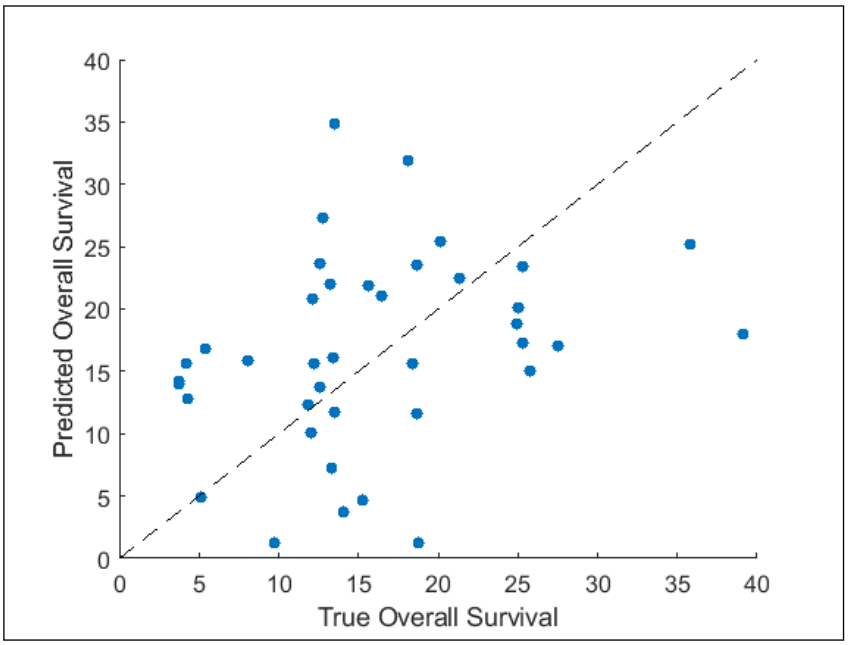


Figure 4.15: *Plot of the predicted versus true values for the overall survival score when using dose, clinical and UNet features only. The line of ideal prediction is displayed as a dashed line.* ***MAE = 7.0****.*

Features	Acc(%)	Sens(%)	Spec (%)	AUC	MAE
Dose	58.8	66.7	52.6	0.539	31.0
Clinical	58.8	63.2	53.3	**0.711**	**22.8**
Dose, Clin	58.8	63.2	53.3	0.575	27.7
PyRad	41.2	47.1	35.3	0.461	32.4
UNet	**67.6**	**72.2**	**62.5**	0.639	26.7
Dose, Clin, PyRad	61.8	68.8	55.6	0.564	33.3
Dose, Clin, UNet	55.9	62.5	50.0	0.579	31.9
Dose, Clin, PyRad, UNet	47.1	52.9	41.2	0.461	36.4

Table 4.15: *Results for the prediction of cough severity for the ASPIRE-ILD dataset (N=35). The binary classification was for any change > 5.*

Features	Acc (%)	Sens (%)	Spec (%)	AUC	MAE
Dose	42.9	57.1	21.4	0.446	1.61
Clinical	51.4	68.8	36.8	0.475	1.75
Dose, Clin	48.6	31.6	68.8	0.408	1.64
PyRad	62.9	75.0	46.7	0.656	1.22
UNet	57.1	75.0	42.1	0.667	1.26
Dose, Clin, PyRad	54.3	38.1	**78.6**	0.591	1.33
Dose, Clin, UNet	42.9	28.6	64.3	0.498	1.43
Dose, Clin, PyRad, UNet	**65.7**	76.2	50.0	0.612	**1.18**
Dose, PyRad, UNet	**65.7**	**78.9**	50.0	**0.670**	1.30

Table 4.16: *Results for the prediction of FACT-L B1 question relating to dyspnea for the ASPIRE-ILD dataset (N=34). The binary classification was for any change > 0.*

4.4 Discussion

4.4.1 Edinburgh Pneumonitis Prediction

4.4.1.1 Dose thresholding the CT image

The best-performing radiomic feature-based prediction model when applying dose thresholding to the CT image was the model that used no dose thresholding. The performance of the model generally decreased as the thresholding value in terms of Grays increased. The decreasing performance may be due to the thresholding causing important features of the lung to be missed or due to a decrease in the signal-to-noise ratio of the features due to a smaller volume being used for the radiomic calculations. This suggests that the CT radiomic features calculated from the lung are global features indicative of a patients individual susceptibility to radiation toxicity effects or overall health and can not be used to highlight regions of the lung where radiation dose should be avoided.

4.4.1.2 UNet Masking

The best-performing model used no masking and simply took a global average pooling of the spatial features of the UNet model with an AUC of 0.717. The performance decreased with any form of masking. Here, masking everything but the body, lungs and background regions achieved AUC scores of 0.652, 0.627 and 0.607 respectively. This suggests that features from the patient's body in addition to the lungs, which was shown by the PyRadiomics approach, may provide predictive information about the patient's susceptibility to RT toxicity. The architecture of the UNet model, where convolution kernels are applied to downsampled versions of the image, creates an increasing receptive field size meaning that features of one pixel can be calculated based on pixels from a distant region. This means that the masking applied in this way can only partially separate the different regions of anatomy and the background. This may explain why the background pixels still contain useful information as they may still contain information about the features of the body or lungs. These methods for separating the features of the image from the extracted UNet features require further investigation as it may still be beneficial to keep the spatial information of the UNet features.

4.4.1.3 Final Pneumonitis Predictions

For the prediction of pneumonitis on the Edinburgh dataset, the model using only the dose and clinical features achieved an AUC of 0.633. The dose feature only model achieved an AUC of 0.640 and the clinical feature only model achieved an AUC of 0.465. This shows that for this dataset the clinical features provide no predictive power and may reduce model performance by increasing the noise in the features input to the model. The models using only PyRadiomics and only UNet-based CT features produced an AUC of 0.760 and 0.717 respectively, both an improvement on the model using only dose and clinical features. While it performed lower than the radiomic method, the technique of extracting features from the UNet model has achieved comparable results. Combining the CT features with the dose and clinical features achieved an AUC of 0.790 and 0.814 when using the PyRadiomics and UNet features respectively. This shows a large improvement in the AUC of over 0.15 for both CT feature sources when combined with dose and clinical features compared to only using dose and clinical features. There is a large clinical relevance here as this shows that CT image features could improve the treatment planning process by allowing for the production of more personalised treatment plans. Finally, the model achieving the highest AUC used the dose, clinical, PyRadiomic and UNet features. This shows that combining CNN-based features with radiomic features can further improve radiation pneumonitis prediction model performance.

4.4.2 ASPIRE Discussion

As the use of SABR for the treatment of lung cancer in patients with ILD is still an emerging treatment option, this is the first study investigating the prediction of toxicity for ILD patients receiving SABR. This means that the predictive performance of any of the features has not been previously investigated adding to the clinical relevance of the results presented.

4.4.2.1 Pulmonary toxicity prediction

For the pulmonary toxicity prediction on the ASPIRE-ILD dataset, the model using dose features only achieved an AUC of 0.687 and the model using only clinical features achieved an AUC of 0.444. The dose and clinical feature model achieved an AUC of 0.664. This follows the same pattern as the Edinburgh dataset where the clinical fea-

tures provide no predictive power and can reduce the performance of the dose features by adding noise to the feature set. Using the PyRadiomic and UNet features on their own achieved an AUC of only 0.515 and 0.479 indicating no predictive power from these features here. When including dose features with the CT features, the AUC improved to 0.746 and 0.744 for the PyRadiomic and UNet features respectively. Combining dose, PyRadiomic and UNet features achieved the highest performance in terms of AUC with an AUC of 0.841. This shows, that while the model failed to gain any predictive power from the CT features on their own, combining them with the dose features leads to a substantial improvement in performance.

4.4.2.2 FACT-L Prediction

For the FACT-L prediction, the dose, clinical, PyRadiomic and UNet features all showed some predictive power achieving AUC values of 0.769, 0.652, 0.688 and 0.652 respectively. In terms of a binary classification the best performing model was the dose feature only model with an AUC of 0.769 and the best performing model in terms of the exact FACT-L score prediction is the model using dose, clinical, PyRadiomic and UNet which achieved a MAE of 13.6. This indicates that, potentially due to a lack of information regarding patient health, the dose only model makes more extreme guesses of the FACT-L score based which would allow the AUC to be higher. Including additional features then allows the model to make more exact predictions as it can somewhat predict the patients' baseline health state before the RT is applied.

Predicting question B1 on the FACT-L questionnaire relating to dyspnea produced a maximum AUC of 0.670 which was achieved when using dose, PyRadiomic and UNet features. This is lower than when predicting the outcome of the FACT-L index as a whole suggesting that it may not be beneficial to look at individual questions of the subscale and that dypsnea is not the only outcome involved in the full FACT-L score prediction.

When the baseline pre-RT FACT-L value is included the performance of the dose only model increases with an AUC of 0.821 and an MAE of 10.9. The inclusion of any other features in this case degrades the model performance. This may indicate that the CT features act as predictors for the baseline FACT-L score meaning that when the baseline FACT-L score is included the CT features provide no benefit.

4.4.2.3 EQ-5D-5L Prediction

For the EQ-5D-5L prediction, the dose, clinical, PyRadiomic and UNet features all showed some predictive power achieving AUC values of 0.633, 0.663, 0.784 and 0.720 respectively. The best performing model in terms of the binary classification was the PyRadiomic model with an AUC of 0.784 and the best performing model in terms of an exact EQ-5D-5L score prediction was the dose, clinical and PyRadiomic feature model which achieved a MAE of 0.139. Here the PyRadiomic features seem to provide most of the predictive power indicating that these radiomic lung features may be a good indicator of a patient's overall health state as defined by the EQ-5D-5L scale.

When the baseline EQ-5D-5L value was included as a predictive feature, the performance of the model when using all of the feature sets individually decreased, the performance only increased when using all of the features together but this still does not beat the best performing model with no baseline features. It is not clear why this is occurring but may be a sign that the models without the baseline are overfitting though there is no way to evaluate if this is occurring without access to a larger dataset.

4.4.2.4 Overall Survival Prediction

For the prediction of the patients overall survival the best MAE for the is 7 months which was observed when using the dose clinical and UNet features. For 1 year survival the best performing model achieved an AUC of 0.844 using clinical features only and for 2 year survival the best performing model achieved an AUC of 0.706 using the dose, clinical and UNet features. Figure 4.15 does not show any clear correlation between the predicted and true overall survival times. This implies that there is unlikely to be much predictive power here though the reasonably high AUC scores suggest that further investigation would be worthwhile.

4.4.2.5 Cough Severity Prediction

When predicting cough severity the best performing model in terms of both AUC and MAE was the model using only clinical information. This achieved an AUC and MAE of 0.711 and 22.8 respectively. Including any other features failed to improve on this prediction.

4.4.3 General Discussion

On both the Edinburgh pneumonitis and ASPIRE-ILD datasets, CT image features calculated from both radiomic and UNet based approaches greatly improved the performance of the models for the prediction of CTCAE pulmonary toxicity. This shows that CT images are an important source of information for pulmonary toxicity post-radiotherapy and that current clinical practice is under-utilising the CT images, taken as standard practice but not used for this type of analysis. This is true of both a standard cohort of lung cancer patients receiving IMRT and patients with ILD receiving SABR.

While these results are significant, it is important to highlight where future improvements can be made. The main limitation of this work is the dataset size. With a larger dataset, a validation dataset could be used for model hyperparameter tuning. Additionally, expanding this work to data from more centres would improve the robustness of the methods and the confidence in the results. Establishing a large benchmark dataset for radiotherapy outcome prediction would also allow for different methods to be directly compared. An additional caveat for the Edinburgh pneumonitis dataset is that due to the outcome collection methods where clinical notes were analysed retrospectively, the less extreme cases such as grade 1 cases may be underrepresented as the clinical notes may not have a clear diagnosis in these cases. It is likely that these cases are harder for a model to accurately predict as they will likely fall closer to the decision boundary. This means that the results on this dataset may be overoptimistic but again this would require more data to confirm or disprove.

4.5 Conclusion

In current clinical practice, only the dose and clinical features are considered during radiotherapy planning for the prediction of toxicity to healthy tissue. The work detailed in this chapter has shown that CT images, which are collected during standard RT planning, can be a source of valuable information when predicting pulmonary toxicity in lung cancer patients. This was shown to be true for both standard patients receiving IMRT and patients with ILD receiving SABR. Additionally, it was shown that a pre-trained CNN for lung segmentation can be used as a feature extractor for this application and can improve the performance of radiomic based prediction models

when the two CT image feature sources are combined. For the ILD SABR cohort, it was also shown that combinations of dose, clinical and CT image features can be used for the prediction of changes in the FACT-L and EQ-5D-5L scores.

Chapter 5

Predicting Esophageal Toxicity After Radiotherapy From Pre-Treatment Dose Maps

This chapter is focused on improving methods for radiation esophageal toxicity prediction using radiotherapy planning dose maps. Section 5.1 discusses the motivations for this work, section 5.2 defines the dataset used, sections 5.3 to 5.5 detail the first set of methods, results and discussion, sections 5.6 to 5.8 detail the second set of methods, results and discussion, finally the conclusions for the whole chapter are presented in section 5.9.

5.1 Introduction

Damage to the esophagus is one of the most common sources of toxicity experienced by lung cancer patients treated with radiotherapy. The ability to predict esophageal toxicity prior to radiotherapy being delivered would allow the adaptation of treatment plans to significantly reduce toxicity.

While there has been much previous work on the prediction of esophageal toxicity from dose information in the literature as detailed in section 3.1.4 of this thesis, this previous work has mostly been focused on including additional feature sources, such as patients pre-treatment cytokine levels [144] or changes in a post-treatment CT scan [147], to improve the predictive performance of dose based models. There has been little focus on the technical details of the application of machine learning models to

the prediction of esophageal toxicities. This is likely due to the lack of any large open-access benchmark datasets with the necessary data required to perform this analysis. The work presented in this chapter aims to advance the field of esophageal toxicity prediction by applying technical adaptations to machine learning models and using the largest currently publicly available dataset, the RTOG-0617 study dataset [193], to validate these methods. Additionally, the techniques applied in this study are likely to be beneficial in toxicity prediction in any region of anatomy.

5.1.0.1 Contributions of this Chapter

In the work detailed in this chapter, methods for esophageal toxicity prediction from RT dose maps are further developed using data available from the RTOG-0617 study [193]. The technical contributions are:

- The first use and development of convolutional neural networks (CNNs) for the prediction of esophageal toxicity from RT using dose images.
- The inclusion of additional esophageal adverse events in the prediction model.
- The development of a novel regression based training approach.
- Application of AI on a validated multi-centre data set (RTOG-0617) with fine-grained follow-up information.
- The testing of the robustness of ANN and LSBoost models to random noise and methods to improve this robustness.

5.1.0.2 Format of Chapter

The methods, results and discussions in this chapter are split into two separate sections. 5.3 to 5.5 detail the first set of methods, results and discussion, sections 5.6 to 5.8 detail the second set of methods, results and discussion. This has been done as different training schemes were applied during each methods section and the end goals were different.

Methods 1: CNN and Decision Tree Prediction Models This section develops the methods for applying 3D CNNs, decision trees and normal tissue complication models to the task of predicting esophageal toxicity from external beam RT. The aim

here is to determine the methods that produce the best performance for predicting esophageal toxicity.

Methods 2: ANN and LSBoost Hyperparameter Tuning and Robustness Tests This section builds on the previous methods to test the robustness of decision tree and ANN based methods for esophageal toxicity prediction and develop methods to improve this robustness. Additionally, the effects of hyperparametrer tuning are investigated.

5.2 Dataset

All patients in this study (N=397) were previously recruited to the RTOG-0617 clinical trial. The RTOG-0617 trial was a multi-centre study set up to investigate the effects of using standard (60 Gy) versus high (74 Gy) doses of radiation to treat lung cancer patients [193]. The RTOG-0617 study recruited 544 patients with NSCLC who received either IMRT or 3DCRT as well as either the chemotherapy drug cetuximab or no concurrent chemotherapy. The patients were recruited from 185 institutions in the USA and Canada between November 2007 and November 2011. The data from the RTOG-0617 clinical trial is currently the largest open dataset available that contains the necessary information for esophageal toxicity prediction from pre-RT dose maps. For the work presented in this thesis chapter, patients with an overall survival of less than 6 months were excluded unless they had developed grade 3 or above esophagitis. This was done to remove patients that may not have had time to develop esophageal toxicity before their death which removed 57 patients from the main cohort. Patients who withdrew consent or were not eligible (N=49) and patients who had issues in their radiotherapy data (N=41), such as missing esophagus contours or missing information required to register the dose and CT images, were also removed. This resulted in a final cohort of N=397. The data from the RTOG-0617 study was accessed through The Cancer Imaging Archive (TCIA) [194], [195] after data access was provided by the National Cancer Institute.

5.2.0.1 Esophageal Toxicity Prevalence

During the RTOG-0617 clinical trial, all radiotherapy induced toxicities, as defined in the Common Terminology Criteria for Adverse Events (CTCAE) v3 [183], were

recorded. While esophagitis is the most common acute toxicity from RT delivered for the treatment of lung cancer [132], there are several esophageal toxicities that can occur after RT. As defined by CTCAE v3, the following esophageal toxicities were identified:

1. Esophagitis
2. Acquired tracheo-esophageal fistula
3. Dysphagia
4. Dyspepsia
5. Esophageal ulcer
6. Esophageal stenosis
7. Esophageal perforation

The grade definitions for esophagitis, as defined by CTCAE v3, are shown in table 5.2. This chapter investigated models to predict if a patient would develop grade $\geq$3 esophagitis and grade $\geq$3 of any esophageal toxicity. The number of patients that developed each grade of esophagitis and the maximum grade of any esophageal toxicity are shown in table 5.1. The maximum grade of any esophageal toxicity was used when looking at all toxicities as it is common for more than one toxicity to occur for a single patient. When looking at splitting the grades into the categories $<$ grade 3 and $\geq$ grade 3, an extra four grade $\geq$ 3 cases were available when looking at all esophageal toxicities.

	Grade 0	**Grade 1**	**Grade 2**	**Grade 3**	**Grade 4**	**Grade 5**
Esophagitis Only	179	51	105	60	1	1
All Esophageal Toxicities	84	99	148	62	2	2

Table 5.1: *The number of patients developing each grade of esophagitis and all esophagus toxicities. The maximum grade of any esophagus toxicity is used in the "all esophageal toxicities" row.*

Grade	Description
1	Asymptomatic pathologic, radiographic, or endoscopic findings only
2	Symptomatic; altered eating/swallowing (e.g., altered dietary habits, oral supplements); IV fluids indicated <24 hrs
3	Symptomatic and severely altered eating/swallowing (e.g., inadequate oral caloric or fluid intake); IV fluids, tube feedings, or TPN indicated $\geq$ 24 hrs
4	Life-threatening consequences
5	Death

Table 5.2: *Radiation esophagitis grade definitions from the Common Terminology Criteria for Adverse Events (CTCAE) ver 3.0 [183].*

5.3 Methods 1: CNN and Decision Tree Prediction Models

5.3.1 Dose Image Pre-Processing

The RT dose maps were initially segmented using the esophagus contours that were available from the RT planning process. To calculate dose volume features the voxel sizes of the dose map, available in the DICOM information, were used. The V_{xGy}, the proportion of the total esophagus volume receiving a dose greater than xGy and the Vol_{xGy}, the total esophagus volume receiving a dose greater than xGy given in units of cm^3, were calculated. These features were calculated over a range of x from 5 to 80 in steps of 5. Additionally, the mean and maximum dose to the esophagus were calculated.

For use with the 3D-CNN, the dose image was isotropically resampled to produce voxel dimensions of 2cm in the x,y and z directions. The images were then cropped to reduce the area of the image outside of the esophagus that had been segmented and set to a voxel intensity value of 0 Gy. The final image resolution was 100x100x120 voxels (200x200x240 cm), which was larger than necessary to include the esophagus in the x and y directions to allow for data augmentation by translation to be applied in the x and y directions without losing any of the esophagus volume. An example of the final 3D esophagus dose map used with the CNN is shown in figure 5.1 (b). The full dose image pre-processing workflow is summarised in figure 5.1.

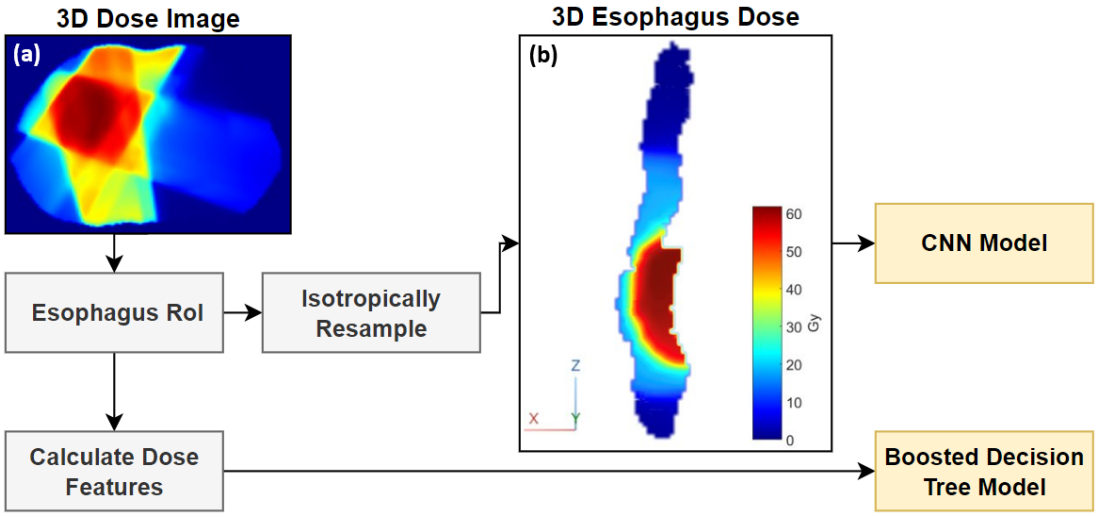


Figure 5.1: *The dose image pre-processing workflow. The esophagus RoI is used to segment the image, features are calculated for feature based prediction models and the segmented image is used as the input to CNN based models.* ***(a)*** *shows a slice of the original 3D dose map* ***(b)*** *is the dose in the esophagus only displayed as a maximum intensity projection in the median plane.*

5.3.2 3D-CNN Based Classification Model

To generate and classify features of the 3D Dose image, a 3D implementation of ResNet-18 [196] was applied as both a binary classification model and as a regression model where the grades were treated as a continuous variable. The same architecture as described in [196] was applied here with the following changes: image input normalisation was replaced by rescaling, setting maximum and minimum dose values, 80 Gy and 0 Gy, to preserve the absolute dose values, dropout with a 0.25 probability was applied after the final global average pooling layer to reduce overfitting and the number of convolutional filters was halved in all convolutional layers to reduce overfitting.

All methods were applied to classify both esophagitis on its own and all esophageal toxicities with the number of each grade defined in table 5.1. The main outcome of all models was the binary classification of grade ≥ 3 esophageal toxicity. The training scheme was the same in all cases which is described as follows.

5.3.2.1 ResNet18 as a Binary Classifier

The first CNN based method applied was to treat the prediction task as a binary classification, a standard approach reported in the literature [144, 145, 146, 147, 197]. To do this, the patients were grouped by toxicity grade into the classes $<$ grade 3 and

$\geq$ Grade 3. For the classification problem, the 3D-ResNet18 model was trained with a weighted cross entropy loss function.

$$WCE = -\frac{1}{N}\sum_{n=1}^{N} w_i \ln \hat{y}_{ni} \tag{5.1}$$

Here, N is the mini-batch size, w_i is the weighting applied to cases of class i and is defined in equation 5.2 and $\hat{y}_{ni}$ is the probability that the network associates the n^{th} input sample with class i.

$$w_i = \begin{cases} w_{<grade3} & \text{if } Grade < 3 \\ w_{\geq grade3} & \text{if } Grade \geq 3 \end{cases} \tag{5.2}$$

The class weights, $w_{<grade3}$ and $w_{\geq grade3}$, were defined by setting the weight for the class with the most samples, in this case $w_{<grade3}$, to 1 and setting the other weight, $w_{\geq grade3}$, such that when multiplied by the number of samples in the under-sampled class, it equals the number of samples in the over-sampled class as described in [66]. Using the number of cases of each grade in the whole dataset gives weightings $w_{\geq grade3}$ = 5.40 and $w_{\geq grade3}$ = 5.02 when looking at just esophagitis and all esophageal toxicities respectively, note that a cross-validation scheme is used for training and testing so these values will change for separate folds, this cross-validation is described later in Section 5.3.2.3.

5.3.2.2 ResNet18 with a Continuous Grade Output

When treating the problem as a binary classification with grade 3 as the class boundary, all of the grade 0, 1 and 2 cases are grouped together in one class. This may not be the ideal training scheme for the CNN to learn features relating to toxicity as the grade 2 cases likely have dose features that are more similar to grade 3 cases than grade 0 cases. To leverage more of the information available in the training of the deep learning model, the toxicity grade classifications were treated as a continuous variable so that they do not have to be grouped into two classes. Additionally, once discarded prior to model training, this information can not be recovered. This means that a model trained for binary classification can only perform binary classification at the grade boundary it was trained for. A model trained with all the grades available can be used for binary prediction at any grade boundary or for exact grade prediction.

The CNN was applied as a regression model where the aim was to predict the exact grade instead of a binary classification. Also, the exact grade is not limited to being an integer value so the toxicity grade was treated as a continuous variable. A similar approach has been applied for severity prediction in different medical imaging applications [198, 199]. Once exact grade predictions are determined, a binary classification can be performed on these continuous grades by defining a cutoff value where any predictions above will be classified as grade ≥ 3. In addition to the potential for an increased accuracy of the binary classification, the continuous grade output approach may be beneficial for estimating the expected severity of the toxicity.

To apply 3D-ResNet18 as a regression model that can treat the toxicity grades as a continuous variable, it was adapted from use as a binary classifier by removing the softmax layer and replacing the final fully connected layer with one with a fully connected layer that has a single output, which acts as the grade prediction. Additionally, the loss function when treating the grades as a continuous variable has to change. The loss function used was the weighted mean square error (WMSE) [200], this is defined in equation 5.3.

$$\text{WMSE} = \frac{1}{N}\sum_{n=1}^{N} w_i \left(y_n - \hat{y}_n\right)^2 \tag{5.3}$$

Here, N is the number of cases in the mini-batch, y_n is the true grade and $\hat{y}_n$ is the predicted grade of the n-th sample in the mini-batch. A weighting, w_i, is again applied to the contribution to the loss from each sample that is dependant on whether the grade is $\geq$ grade 3 $<$ grade 3 as defined in equation 5.2.

Due to the grade distribution of the training set and the discreet nature of the ground truth toxicity grades, the grades predicted by the regression model will be biased to be close to the mean value of the training dataset, i.e. the variance of the model output will be lower than reality. This can be corrected for by comparing the distribution of the training data ground truth to the distribution of the model output with the training data input. To do this, the variance, $var(Y_{train})$, and mean, $mean(Y_{train})$, of the ground truth toxicity grade distribution of the training dataset were calculated. Then the training data was input to the model after it had been trained and the grade predictions were calculated. The variance, $var(\hat{Y}_{train})$, and mean, $mean(\hat{Y}_{train})$, were then calculated for these training set grade predictions. Equation

5.4 [201] was then used to reduce the bias of the model output, $\hat{y}$, by setting the output of the model to $\hat{y}_C$ which, for the training dataset, will make the mean and variance of the adjusted model output equal to the mean and variance of the ground truth toxicity grades.

$$\hat{y}_C = (\hat{y} - mean(\hat{Y}_{train})) * \sqrt{\frac{var(Y_{train})}{var(\hat{Y}_{train})}} + mean(Y_{train}) \tag{5.4}$$

On its own, the adjusted model output, $\hat{y}_C$, could be clinically used as a risk score as discussed in [202], but to compare the results to the literature and our other methods, this needs to be converted into a binary classification. To do this, a cutoff value has to be defined for the model output where any value above the cutoff will be classed into the grade ≥ 3 class and any value below the cutoff will be classed into the grade < 3 class. The AUC ROC metric is calculated by varying this cutoff value and calculating the true positive rate and the false positive rate across the range that causes these values to vary from 0 to 1. To calculate the sensitivity and specificity of the binary classification, a single cutoff value has to be selected. A softmax function was used to create a binary classifier based on the continuous grade output of the model. This was trained by splitting the data into the classes $<$ grade 3 and $\geq$ grade 3. The input to the softmax function was the grade prediction from the CNN model. The training data was used by applying it to the trained CNN model which produced floating-point grade predictions for each case. The softmax layer was then trained to convert these grades into a probability. The training was completed for 1000 epochs using the Adam optimiser and an initial learning rate of 0.01. This was done using the MATLAB deep learning toolbox. The full workflow for the training of the regression based CNN is shown in figure 5.2.

5.3.2.3 CNN Training and Cross Validation

For both the binary classification and continuous grade regression based applications of the 3D-ResNet18 model, the same training protocol was used. The dataset was randomly split into four subsets with an equal split while keeping an equal proportion of grades in all of the sets, note that this was not possible for grade four and five cases as there were not enough patients receiving those grades in the dataset. A four-fold cross-validation approach was taken so the networks were trained on three of the

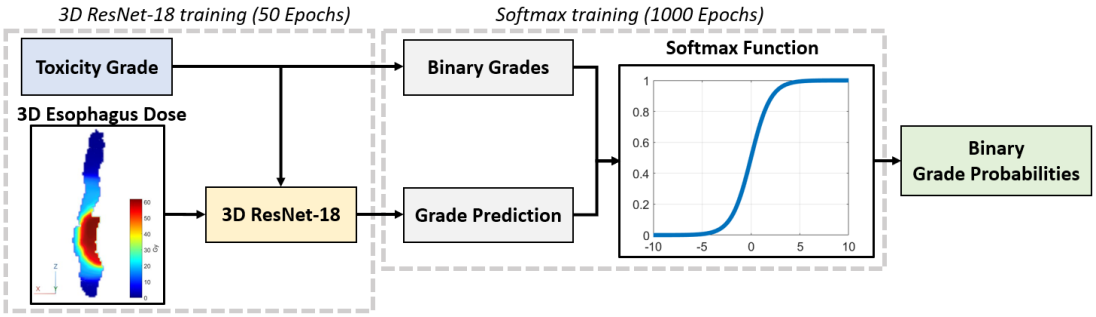


Figure 5.2: *The workflow for training the regression based model for esophagus toxicity prediction using ResNet-18 with a continuous grade output and separate training of a softmax layer to produce a binary classification.*

subsets and tested on the remaining subset of the data. This was repeated three times using a different subset as the test set so that the models could be tested on the full dataset. The cross-validation strategy is shown in figure 5.3. Note that a new model was trained for each new fold so that each model was not trained on any of the data it was being used to test as this would bias the model. The training was performed using the parameters defined in table 5.3.

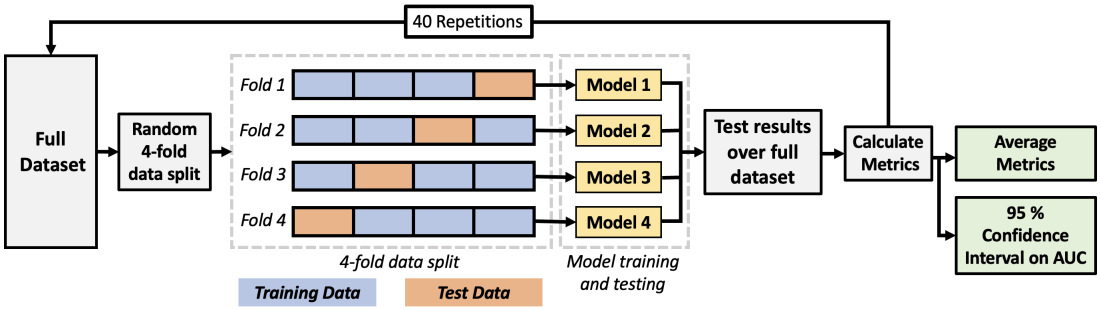


Figure 5.3: *The workflow for the 4-fold cross-validation training strategy and the 95% confidence interval on the AUC calculation.*

Training Parameter	Value
Optimisation Algorithm	adam
Epochs	50
Mini-Batch Size	12
Initial Learning Rate	0.01
Learning Rate Drop Period (epochs)	10
Learning Rate Drop Factor	0.2

Table 5.3: *The training parameters for the training of the 3D ResNet18 network as both a binary classifier and continuous output regression model.*

During the training of the network, data augmentation steps were applied after

every epoch of training to reduce overfitting and increase the robustness of the features learned by the network. These augmentation steps were a random rotation between 0 and 360 degrees around the z-axis and a random translation between -10 and 10 pixels in both the x and y directions. The training was run on two NVIDIA GeForce GTX 1080 Ti graphics cards in parallel.

5.3.3 Dose Feature and Decision Tree Based Classification Model

As well as the CNN based approach to the classification, a dose volume histogram feature based approach was taken with boosted decision trees used as the classifier. Again, all methods were applied to classify both esophagitis and all esophageal toxicities with the number of each grade defined in table 5.1. The features used here were the V_{xGy} and Vol_{xGy} over the range of x from 5Gy to 80Gy in steps of 5 as well as the mean and maximum esophagus dose. The calculation of these was described in section 5.3.1.

In the same manner as with the 3D CNN approach, both a binary and continuous grade definition approach was taken for the training of the models. For the binary grade approach, the grades were again grouped into the classes $<$ grade 3 and $\geq$ grade 3. For the binary classification approach, the AdaBoost algorithm [187] was applied. For the regression approach, decision trees with the LSBoost algorithm [201] was applied. Just like with the CNN approach, the output from the LSBoost model was a continuous grade from which an AUC ROC, sensitivity and specificity metrics could be calculated following the same methods described in section 5.3.2. The parameters used for the training of both the AdaBoost and LSBoost methods are shown in table 5.4. For the AdaBoost method, class weights were again applied to the loss function to account for the unbalanced grades. All training was completed using a 4-fold cross-validation approach in the same manner as the 3D-CNN training which is detailed in figure 5.3.

Training Parameter	Value
Learning rate	0.01
Learning cycles	200
Maximal number of decision splits	10
Minimum observations per leaf	1 (AdaBoost), 5 (LSBoost)
Minimum observations per branch node	2 (AdaBoost), 10 (LSBoost)

Table 5.4: *The training parameters for the classification of the dose features using the boosted decision tree classification methods AdaBoost and LSBoost.*

5.3.4 LKB NTCP Model

The final model applied was the LKB NTCP model which is a mathematical model for predicting the probability of normal tissue toxicity due to dose volume effects. The LKB NTCP model is defined by equation 5.5, 5.6 and 5.7 [97, 98, 99, 100]. These equations are described in more detail in section 3.1.1.

$$\text{NTCP} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{x^2}{2}} dx \tag{5.5}$$

$$t = \frac{\text{EUD} - TD_{50}}{m^* TD_{50}} \tag{5.6}$$

$$\text{EUD} = \left(\sum_i v_i D_i^{\frac{1}{n}} \right)^n \tag{5.7}$$

The LKB NTCP model was applied by fitting a non-linear function to the data to solve for TD_{50}, n and m through the application of the Levenberg-Marquardt nonlinear least squares algorithm [203, 204]. The V5 to V80 in steps of 5 were used to fit the model. In the same manner as for the previous models, this was done with a 4-fold cross-validation approach where the model was fit using data from three folds and tested on data from a single fold, repeating this until the full dataset had been tested over.

5.3.5 Ensemble Model

Ensemble models combine the predictions of two or more models to make a new prediction based on the predictions from all of the models and have been shown to consistently improve performance when applied to other classification problems [205, 206]. A simple ensemble model was produced by averaging the classification probabilities of the LKB NTCP model, dose feature decision tree based model and the CNN based model. This was done separately for the binary classification and regression approaches as well as the models looking at only esophagitis and all esophageal toxicities producing four separate ensemble models that could be compared.

5.3.6 Classification on only the more extreme cases

It is likely that the model performs better on the more extreme cases than on the marginal cases as the extreme cases are more likely to have distinct dose features. Marginal cases here are grade 1 and 2 cases where some form of esophageal toxicity has been noted but it is below grade 3 so it would be grouped in the grade < 3 group. To determine if there is a difference in the predictive power on the more extreme cases, the testing of the model with the highest performance was repeated on only the grade 0, 3, 4 and 5 cases. This was applied with the same 4-fold cross-validation testing scheme that was previously applied.

5.3.7 Grade $\geq$2 Classification

While in the literature, grade ≥ 3 is usually considered the most clinically relevant boundary to define for a binary classification of esophageal toxicity, there is still a benefit in reporting model performance at other grade boundaries. The next most relevant grade boundary is to predict if a patient develops grade ≥ 2 esophageal toxicity. The methods described in sections 5.3.1 to 5.3.5 were repeated using grade ≥ 2 as the boundary for positive cases.

5.3.8 Exact Grade Classification and Risk Score

Instead of creating a binary classification model, which is the standard approach in the literature, it may be more clinically relevant to create a model that outputs a continuous risk score or predicts an exact toxicity grade. With the regression based models, the output is an estimate of the patient's toxicity grade. This was converted to an integer grade by simply rounding to the nearest integer. Here grades 3, 4 and 5 are grouped together as there are not enough grade 4 and 5 cases to draw any meaningful conclusions on their own. Additionally, the model's direct output was recorded which represents a risk score. This was all done for a single iteration of the CNN-LSBoost ensemble.

5.4 Results 1: CNN and Decision Tree Prediction Models

5.4.1 Grade ≥3 Binary Classification

For all model implementations, the performance metrics calculated were the AUC ROC, sensitivity and specificity for grade ≥3 classification these are shown in table 5.5. 95% confidence intervals of the AUC were calculated by repeating the 4-fold cross-validation training and testing of each model on new random subsets of the data 40 times. The model achieving the highest AUC is model 12 which is the CNN-LSBoost ensemble model using regression models applied to all of the esophageal toxicities.

	Model	Model Type	Adverse Events	Sens (%)	Spec (%)	AUC	95% CI (AUC)
1.	LKB NTCP	Classification	Esophagitis	59.9	65.7	0.646	0.639, 0.652
2.	AdaBoost	Classification	Esophagitis	46.3	76.2	0.652	0.640, 0.662
3.	3D ResNet-18	Classification	Esophagitis	70.0	53.8	0.662	0.653, 0.675
4.	Ensemble	Classification	Esophagitis	64.1	62.5	0.671	0.663, 0.678
5.	LKB NTCP	Classification	All Esophageal	62.4	64.8	0.653	0.640, 0.664
6.	AdaBoost	Classification	All Esophageal	56.4	66.4	0.652	0.647, 0.658
7.	3D ResNet-18	Classification	All Esophageal	69.1	56.4	0.678	0.669, 0.687
8.	Ensemble	Classification	All Esophageal	65.2	60.4	0.682	0.675, 0.689
9.	LSBoost	Regression	Esophagitis	50.6	78.6	0.682	0.677, 0.687
10.	3D ResNet-18	Regression	Esophagitis	72.4	51.3	0.676	0.669, 0.681
11.	Ensemble	Regression	Esophagitis	61.3	66.8	0.680	0.673, 0.686
12.	LSBoost	Regression	All Esophageal	58.1	71.1	0.686	0.683, 0.690
13.	3D ResNet-18	Regression	All Esophageal	72.4	54.6	0.702	0.694, 0.707
14.	Ensemble	Regression	All Esophageal	65.2	65.3	**0.705**	0.699, 0.711

Table 5.5: *The sensitivity, specificity and AUC metrics for all models tested over the whole dataset using 4-fold cross-validation. 95% confidence intervals are also included for the AUC metric. "Ensemble" is an ensemble of the LKB NTCP, decision tree and CNN models.*

5.4.1.1 Testing on only the more extreme cases

The results of testing the ensemble models, which are models 3, 6, 9 and 12 in table 5.5, on only the grade 0 and grade ≥ 3 cases are shown in table 5.6. It should be noted here that when looking at all esophageal toxicities rather than just esophagitis produces a different number of test cases as there are more grade 1 and 2 cases when considering all esophageal toxicities. When considering only esophagitis there are 179 grade 0 and 62 grade ≥ 3 cases, when considering all esophageal toxicities there are 84 grade 0 and 66 grade ≥ 3 cases.

	Model	Model Type	Adverse Events	Sens (%)	Spec (%)	AUC	95% CI (AUC)
4.	Ensemble	Classification	Esophagitis	64.1	65.3	0.697	0.692, 0.702
8.	Ensemble	Classification	All Esophageal	65.3	71.8	0.743	0.738, 0.748
11.	Ensemble	Regression	Esophagitis	61.3	70.7	0.710	0.706, 0.714
14.	Ensemble	Regression	All Esophageal	65.2	75.3	**0.768**	0.765, 0.772

Table 5.6: *The accuracy, sensitivity, specificity and AUC metrics for all models tested on only grade 0 and grade ≥ 3 cases. 95% confidence intervals are also included for the AUC metric.*

	Model	Model Type	Adverse Events	Sens (%)	Spec (%)	AUC	95% CI (AUC)
1.	LKB NTCP	Classification	Esophagitis	62.2	55	0.628	0.619, 0.637
5.	LKB NTCP	Classification	All Esophageal	61.6	59.6	0.651	0.641, 0.660
9.	LSBoost	Regression	Esophagitis	49.0	65.2	0.628	0.621, 0.635
10.	3D-ResNet	Regression	Esophagitis	72.1	48.4	0.646	0.637, 0.653
11.	Ensemble	Regression	Esophagitis	60.8	58.5	0.651	0.645, 0.657
12.	LSBoost	Regression	All Esophageal	65.6	61.1	0.670	0.662, 0.677
13.	3D-ResNet	Regression	All Esophageal	73.4	53.6	0.673	0.665, 0.679
14.	Ensemble	Regression	All Esophageal	68.6	59.3	**0.676**	0.669, 0.682

Table 5.7: *The accuracy, sensitivity, specificity and AUC metrics for all regression models when predicting grade ≥ 2 cases. 95% confidence intervals are also included for the AUC metric.*

5.4.2 Grade $\geq$ 2 Binary Classification

The results for the grade ≥ 2 classification are displayed in table 5.7. Here only the regression models are presented as the models do not need to be retrained, only the grade cutoff for binary classification needs to be recalculated. The LKB NTCP models were also retrained for grade ≥ 2 classification for the purpose of comparison and use in the ensemble models. The best performing model here is again the ensemble regression model for classifying all esophageal toxicities which achieved an AUC of 0.676. A similar performance increase is observed here for grade ≥ 3 classification when classifying all esophageal toxicities as opposed to only using the esophagitis grades.

5.4.3 Exact Grade Classification and Risk Score

The exact grade prediction is presented in the confusion matrix in Figure 5.4. While the accuracy is low when predicting exact grades, this is expected due to both the higher level of granularity when looking at exact CTCAE toxicity grades and the large variation in dose response of patients due to currently poorly understood factors. Analysing the model's performance with this level of granularity can help shed light on the limits of toxicity prediction models. It can be seen in Figure 5.4 that the predicted grade for

each true grade appears to follow a distribution that peaks on the correct grade and reduces further away from that grade. This distribution can be highlighted further by displaying a box plot of the predicted grades for each ground truth grade, this box plot is displayed in figure 5.5.

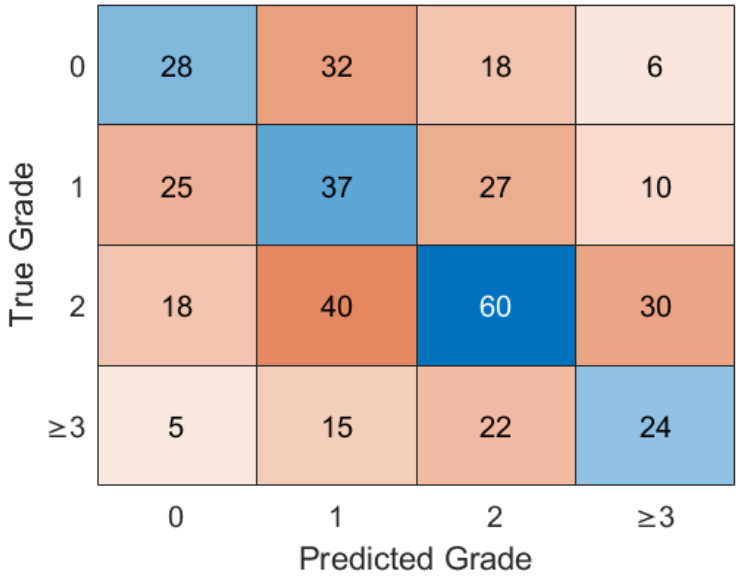


Figure 5.4: *Confusion matrix produced by rounding the output of the regression based model to the nearest integer.*

In figure 5.5 it can be seen that the model has a poor differentiation between grade 0 and grade 1 cases. This aligns with the understanding of CTCAE grades as grade 1 cases are mild, potentially asymptomatic cases, with clinical observations only which causes challenges for clinicians when differentiating between grade 1 and grade 0 cases which introduces a large interobserver error when reporting these cases. Additionally, if a patient's symptoms are mild they may not report these to a clinician. The clearest split in the model predictions presented in figure 5.5 is between the grade 1 and grade 2 cases. This again can be explained through the CTCAE grade definitions as grade 2 cases are symptomatic cases that generally have some, although minimal, clinical intervention. This definition is more clear and allows clinicians to more accurately define cases as grade 2. There is again a poor prediction split between grade 2 and grade 3 cases. This is potentially again due to an unclear point at which a grade 2 case becomes a grade 3 case again creating the potential for a large interobserver error.

It can also be seen that while the variability of the model output has been corrected using equation 5.4, there is still some bias towards the average predicted grade which

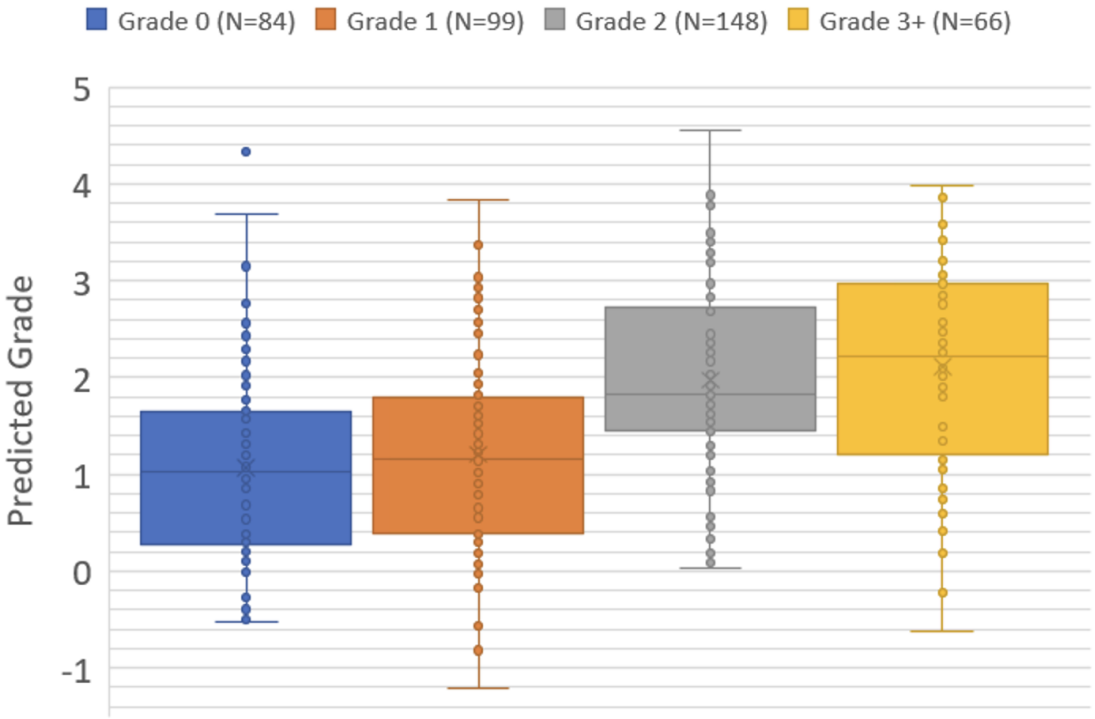


Figure 5.5: *Box plot showing the predicted toxicity grades for all ground truth grade sets.*

is hard to remove due to dataset size limitations and the difficulty converting the quantised grades into a continuous output. To highlight the importance of applying equation 5.4, the same box plot is shown without including this correction in figure 5.6.

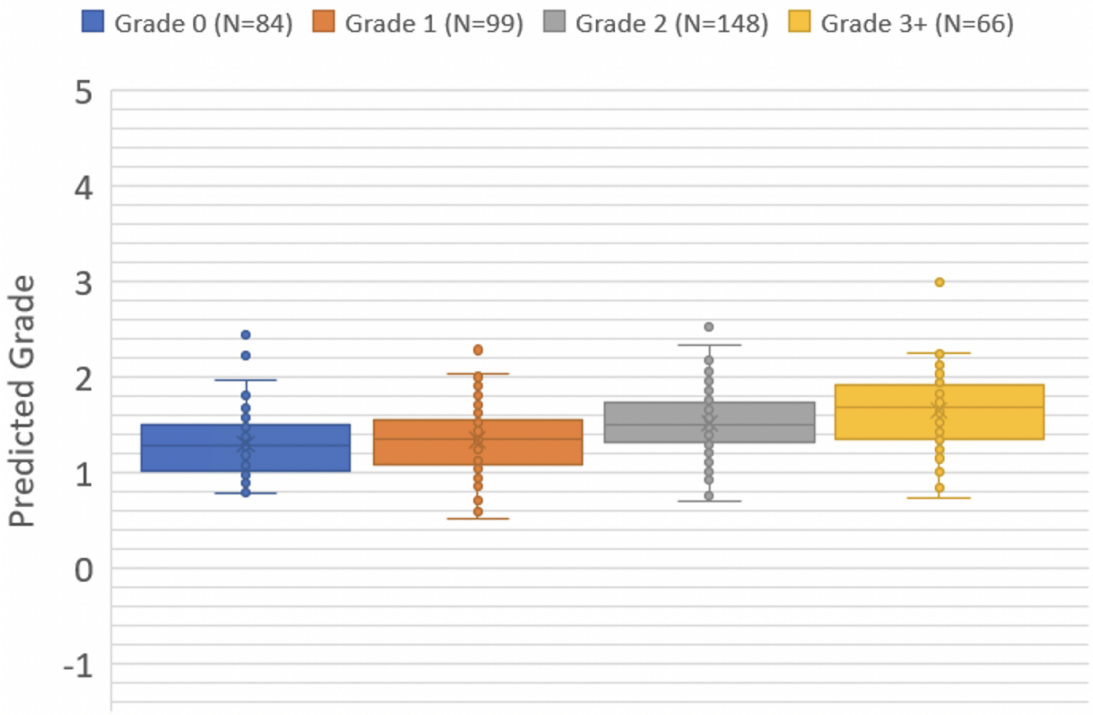


Figure 5.6: *Box plot showing the predicted toxicity grades for all ground truth grade sets without applying equation 5.4 to adjust the models variance.*

5.5 Discussion 1: CNN and Decision Tree Prediction Models

In the literature, when predicting grade ≥ 3 esophageal toxicity using only dose data, AUC values of 0.75 (N=15/147) [144] and 0.70 (N=21/177) [148] have previously been reported with Luna et al. [207] reporting no statistically significant predictive power (N=23/202). Here N=23/202 means 23 positive cases out of 202 total cases. For grade ≥ 2 esophageal toxicity, AUC values of 0.60 (N=51/161) [145], 0.70 (N=49/129) [146] and 0.76 (N=36/79) [197] have been reported. As all of these studies use separate datasets, little information can be gained from comparing these AUC metrics directly to the work in this chapter work. The small size of the datasets used in this field, with the largest test set containing only 23 grade ≥ 3 cases [207], creates much uncertainty when comparing separate works. This is highlighted by Luna et al. [207] reporting no predictive power from dose and clinical features that were included in other works. For this reason, only direct comparisons between the separate methods trained in this study can be used to draw conclusions.

The standard method in the literature for the prediction of toxicity to the esophagus using dose information is to calculate dose features from the dose map and use machine learning techniques such as decision trees as a binary classifier to create a prediction model for esophagitis only [144, 145, 146, 147]. The model applied here that is most similar to this is model 2 in table 5.5 which achieved an AUC of 0.652. An older but more established method in the literature is to use NTCP models such as the LKB model [99, 96]. The implementation of the LKB NTCP model here achieved a similar performance to the decision tree based methods with an AUC of 0.646 when predicting esophagitis. The best performing model was model 14 in table 5.5 which achieved an AUC of 0.705. This gives an increase of the AUC of around 0.05 from applying the methods discussed in this study. By comparing the different models in table 5.5, it can be seen that this increase in predictive power arises from a combination of the regression based training scheme, including additional esophageal toxicities and combining the 3D-CNN, boosted decision tree and LKB NTCP models as an ensemble. Combining the models in an ensemble is also likely to improve the models robustness as if one method fails due to errors in data then there is a chance the other methods will not fail and maintain a more accurate prediction.

While many studies have achieved a higher AUC by using additional features that were not available in the RTOG-0617 dataset, such as patients pre-treatment cytokine levels [144] or changes in a post-treatment CT scan [147], these studies all include dose features in their final models. Therefore, the advancement in the predictive power of dose features reported here may be of benefit to all of these models. Further, the methods described in this study are not exclusive to either dose based prediction or to the esophagus, therefore, they may be of benefit to all toxicity prediction studies and AI-guided RT.

5.6 Methods 2: ANN and LSBoost Hyperparameter Tuning and Robustness Tests

The use of the CNN model in the previous training scheme made it impractical to employ validation for hyperparameter tuning in addition to the cross-validation approach for testing due to the training times of the CNN. It was observed that the use of a CNN on its own provided little benefit to the AUC performance metric compared to

the LSBoost, these results are given in section 5.4.1. For these reasons, an ANN (or MLP) using the same dose features as the boosted decision tree model was used in place of the CNN for investigating how the performance metrics can be improved through hyperparameter tuning in addition to some robustness tests which also required many iterations of the training cycle. Only the regression models were tested here with all esophageal toxicities used.

5.6.1 Hyperparameter Tuning of the ANN and LSBoost Regression Model

The models applied here were a simple ANN and the LSBoost algorithm for boosted decision trees, both using the dose features calculated in section 5.3.3 as the model input. The ANN consists of an input layer, a fully connected layer, an activation layer, a fully connected layer and an output layer connected in order. The models were trained as regression models with a continuous output. This output was adjusted for variance bias using equation 5.4 as described in section 5.3.2.2. The models were evaluated as exact grade prediction models and as binary classifiers by defining a cutoff value, 1.5 for grade ≥ 2 and 2.5 for grade ≥ 3 classification. The models were evaluated using the AUC for both grade ≥ 2 and grade ≥ 3 binary classification and the mean average error (MAE) for the exact grade output. Additionally, the ANN and LSBoost models were combined as an ensemble model by taking a weighted average of the two model outputs where the weighting was a tunable parameter. A nested cross-validation approach was taken for hyperparameter tuning and testing.

Test Data Split A 4-fold data split was taken where one fold was held back for testing and the other three folds were used for training and hyperparameter tuning. The training and hyperparameter tuning were repeated four times, leaving out a separate fold for testing each time so that testing could be repeated for the full dataset.

Training and Hyperparameter Tuning The data in the three folds left out for training and hyperparameter tuning were again split into four folds. One of these folds was used for hyperparameter tuning and the others were used for training. Cross-validation was performed so the hyperparameter tuning was repeated on each fold and the average hyperparameters were used to train the final model on all of the folds (not including the test data). Hyperparameter tuning was applied by Bayesian optimisation to minimise the MSE on the validation set with 40 iterations of Bayesian

optimisation run for every fold. The loss function used during training for both models was the mean square error (MSE). The hyperparameters tuned and their average values across the training repetitions are: **ANN Hyperparameters:** L2 regularisation term λ, fully connected layer size, activation function, loss gradient tolerance. **LSBoost Hyperparameters:** Maximum number of splits, minimum leaf size, minimum parent size, number of learning cycles, learning rate. **Ensemble Hyperparameters:** The weighting proportion for the LSBoost and ANN models.

5.6.2 Testing the Model Robustness

An often overlooked aspect of model performance in RT toxicity prediction is the model robustness. A robust model will perform well on the dataset it is trained and tested on as well as data from other datasets which may have dataset specific biases. The best way test the robustness of a model is to apply it to a large variety of test data from different sources. This would allow the performance of the model to be observed under the presence of additional random and systematic variance in the features. In the RT toxicity prediction domain, sources of this variance include OaR contouring differences, scanner and treatment linac differences, scanner and linac calibration errors, patient movement and unrelated patient health changes. Additional large sources of systematic variation include using a different RT method such as VMAT, treatment of a different type of cancer, different fractionation and the use of concurrent treatment options such as chemotherapy. The RTOG-0617 study was a multi-centre study using both IMRT and 3D-CRT as well as a mix of concurrent chemotherapy or no chemotherapy so there will is already reasonable variance in the data meaning the previous test results have evidence for having a good level of robustness. A negative of using this data is that the patient recruitment and RT application will have followed the study protocol [193] which contains constraints on patient eligibility depending on their tumour stage and overall health as well strict definitions for the OaR contouring and RT application meaning the variability in terms of patient health and RT application will be reduced compared to a general clinical setting.

With this all in mind, this section aims to test the robustness of the ANN and LSBoost models for both grade ≥ 3 toxicity prediction and exact toxicity grade prediction. As there is a lack of public data in the toxicity prediction domain, these robustness tests are performed by adding random noise to the test data features and testing the

performance of the models as the level of noise increases. Before training and testing, the dose features were converted to z-score values by setting the mean of each feature to zero and the feature values such that a value of 1.0 would indicate a value that is one standard deviation from the mean. This was done simply so that the level of noise can be interpreted easily in terms of feature standard deviations. Noise can then be added to feature F converting it to feature F_N through equation 5.8.

$$F_N = F + X \tag{5.8}$$

Where $X \sim U(-\alpha, \alpha)$ and $U(-\alpha, \alpha)$ is a uniform distribution with maximum and minimum values α and $-\alpha$. This distribution is represented by figure 5.7.

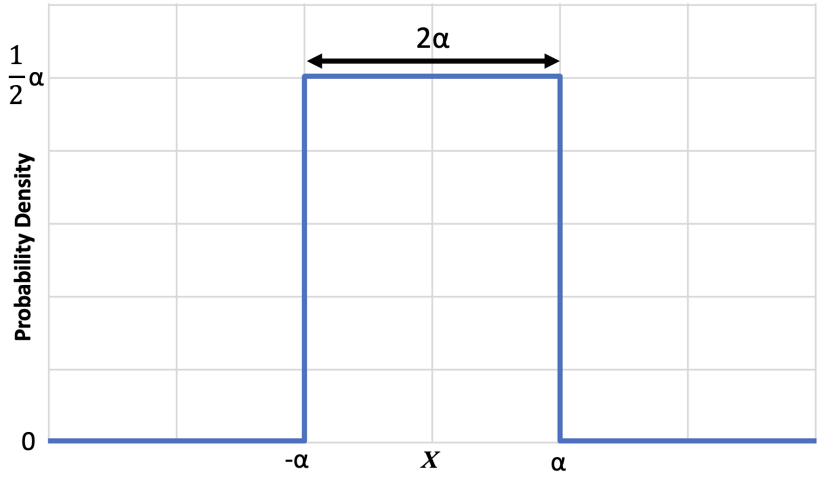


Figure 5.7: *Probability distribution for the noise added to the dose features.*

A training and testing scheme was set up to add noise to the test set with increasing values of α and record the effect on the AUC ROC and MAE test metrics.

To do this, the 4-fold cross-validation training and testing scheme was repeated for the ANN and LSBoost models without the added hyperparameter tuning aspect of cross-validation. Noise was added to the test set prior to model predictions. This was repeated 5 times to get an average for the AUC and MAE. This was all repeated for noise with values of α in the range of 0 to 5 increasing by 0.1 each iteration. Plots of the MAE and AUC ROC were produced for the ANN and LSBoost models for the increasing values of noise.

5.6.3 Improving Model Robustness

After the baseline results for model robustness under increasing noise added to the features were produced, methods to improve the model robustness were applied.

5.6.3.1 Ensemble Model

The first method to improve model robustness investigated was combining the ANN and the LSBoost model as a weighted ensemble. The most commonly selected weighting from the hyperparameter tuning was used.

5.6.3.2 ANN L2 Regularisation

L2 regularisation is known to increase model robustness by forcing the model to rely on features more evenly which reduces scale of overfitting [208]. The hyperparameter tuning showed that a high level of L2 regularisation is required for optimum performance of the ANN model. The robustness test was repeated for the ANN with different values of λ, the L2 regularisation term. L2 regularisation was not available for the LSBoost implementation used.

5.6.3.3 SMOTE

Synthetic minority oversampling technique (SMOTE) [209] is commonly applied in radiotherapy prediction models [210] to improve model performance by generating synthetic training data. SMOTE was applied to the training data to oversample the underrepresented grades, grade 2 was the most common grade for esophageal toxicity so SMOTE was applied to oversample grade 0, 1 and 3 cases. SMOTE was not applied to the grade 4 and 5 cases as there were only two of each case in the full dataset so it is not possible to oversample these cases. The performance of the ANN and LSBoost model was recorded as noise was applied to the model with SMOTE applied.

5.6.3.4 Adding Noise to the Training Data

The training dataset was expanded by adding random noise to the training data features. This was applied following equation 5.8 using a α value of 0.2. The training dataset was expanded to 5 times its original size using this technique. This was done for both the ANN and LSBoost models both on its own and with SMOTE.

5.7 Results 2:

5.7.1 ANN and LSBoost Hyperparameter Tuning

The results of the model training with hyperparameter tuning are displayed in figure 5.8.

Model	AUC (Grade $\geq$ 3)	AUC (Grade $\geq$ 2)	MAE
ANN	**0.709**	0.690	0.894
LSBoost	0.682	0.669	0.925
Ensemble	0.707	**0.691**	**0.884**

Table 5.8: *Results for the hyperparameter tuning of the ANN and LSBoost models.*

The hyperparameter values most commonly selected by the Bayesian hyperparameter tuning for both the ANN and LSBoost models are given here. **ANN Hyperparameters:** L2 regularisation $\lambda = 0.1$, fully connected layer size = 20, activation function = sigmoid, loss gradient tolerance = $1e^{-4}$. **LSBoost Hyperparameters:** Maximum number of splits = 2, minimum leaf size = 2, minimum parent size = 5, number of leaning cycles = 400, learning rate = 0.01. The average weighted average ensemble weighted the ANN 4 times higher than the LSBoost model on average. Further details regarding the hyperparameters can be found in the MATLAB R2022b files for the functions *'fitensemble()'* and *'fitrnet()'* for the LSBoost and ANN models respectively.

5.7.2 Baseline and Ensemble Model Robustness

The noise response of the AUC for the ANN, LSBoost and ensemble models are shown in figure 5.8. From this, it can clearly be seen that the ANN is much more robust to noise than the LSBoost model when considering the AUC as the AUC drops off much more slowly for the ANN than the LSBoost model. The AUC for the LSBoost model drops off sharpy as noise is added to the test data suggesting that it is a less robust model than the ANN. The ensemble model here achieves almost identical performance to the ANN.

The noise response of the MAE for the ANN, LSBoost and ensemble models are shown in figure 5.9. This follows a similar pattern to the AUC where the LSBoost model's performance sharply drops off and the performance of the ANN remains close

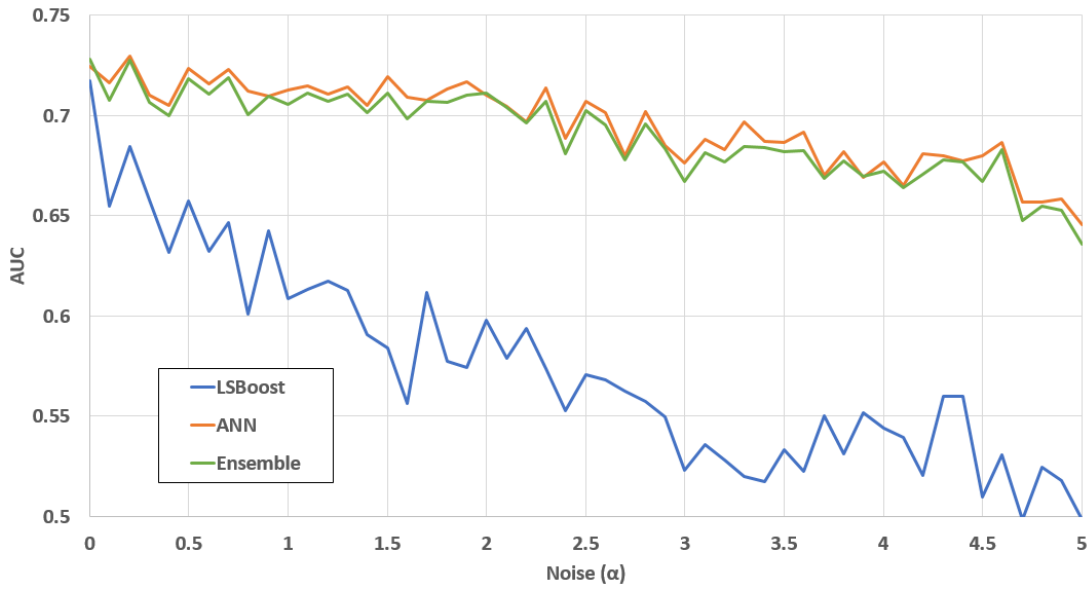


Figure 5.8: *AUC for the ANN, LSBoost and Ensemble baseline models with increasing noise.*

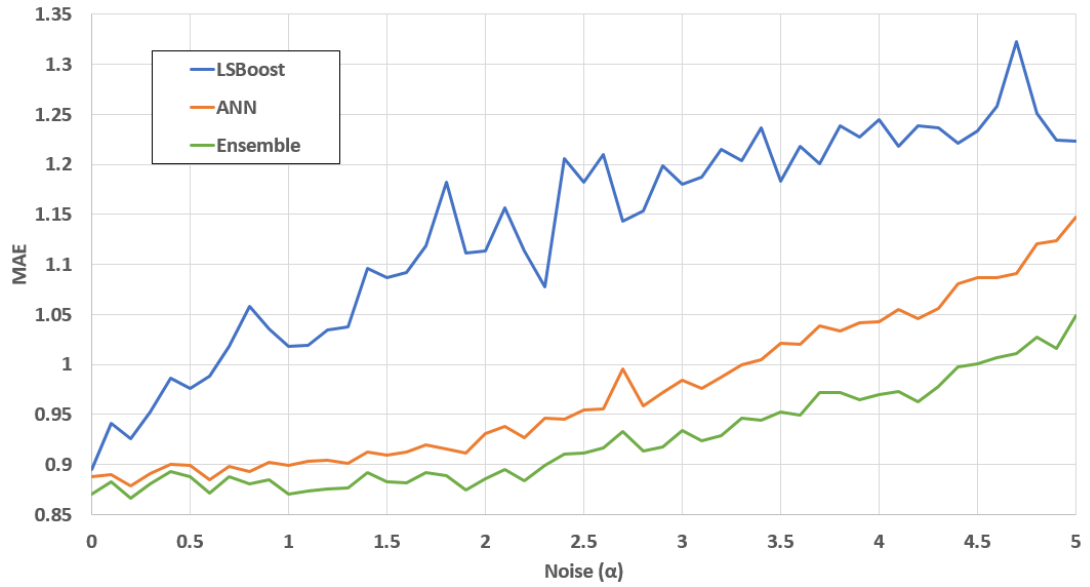


Figure 5.9: *MAE for the ANN, LSBoost and Ensemble baseline models with increasing noise.*

to optimum under the presence of a large level of noise. In this case, the ensemble model has a slightly improved performance compared to the ANN showing that there is still a benefit to choosing a weighted ensemble model over a single model even when one model is outperforming the other.

5.7.3 L2 Regularisation

The results for testing the response of the ANN to noise added to the test set for different L2 regularisation strengths is presented here. Figure 5.10 shows the ANN response in terms of the AUC where it can be clearly seen that the lower λ values produce models that are less robust to noise in terms of the AUC. There is no discernible difference between the models with λ from 0.25 to 0.95.

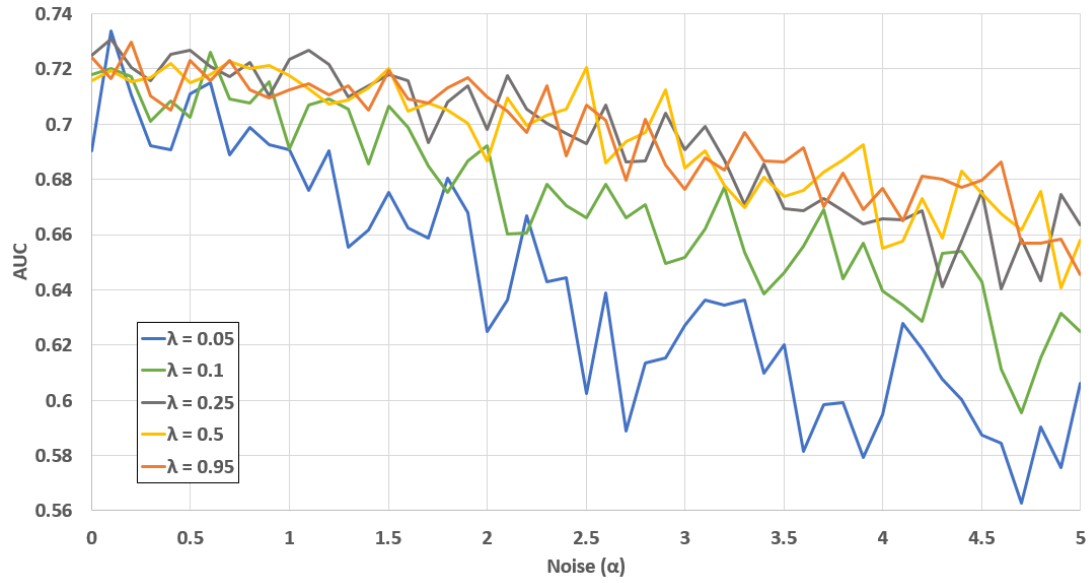


Figure 5.10: ***ANN AUC*** *with increasing noise with different L2 regularisation terms, λ.*

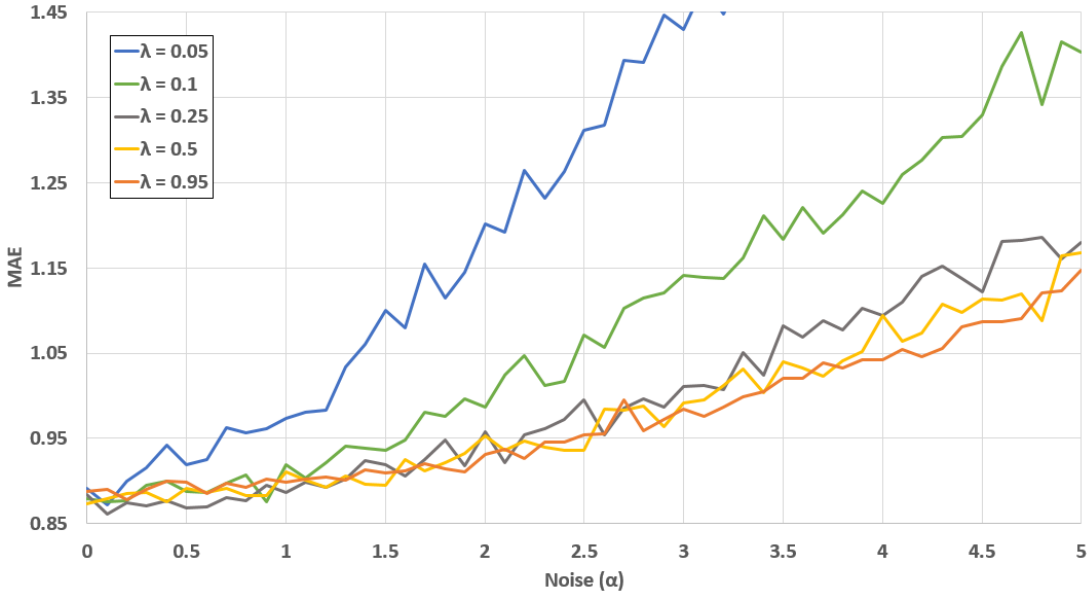


Figure 5.11: ***ANN MAE*** *with increasing noise with different L2 regularisation terms, λ.*

Figure 5.11 shows the ANN response to added noise in terms of the MAE. It is again observed that the lower values of λ have a sharper increase in the MAE as the

5.7.4 SMOTE and Adding Noise

The results of testing the response of the LSBoost and ANN models to increasing noise in the test set with both SMOTE and random noise used to augment the training data is presented here.

5.7.4.1 LSBoost

The AUC response of the LSBoost model is displayed in figure 5.12. There is no observed difference in the robustness of the LSBoost model in terms of the AUC when applying SMOTE, random noise or both during the training of the model.

The MAE response of the LSBoost model is displayed in figure 5.13. Here it can be seen that at higher levels of random noise added to the test set, all the methods of data augmentation improve the MAE of the model. This effect is stronger with SMOTE than augmentation by random noise. The SMOTE method changes the grade distribution of the training dataset which may bias the model to predict grades more uniformly, altering the MAE response. The lack of divergence at low levels of noise suggests that this is not the case so most of the improved response should be from a reduction of overfitting. The added noise method of data augmentation does not alter the grade distribution so the improvements seen here are again likely entirely due to a reduction in overfitting.

5.7.4.2 ANN

The AUC response of the ANN is displayed in figure 5.14. It can be seen from this that there is again no discernible change to the robustness in terms of the AUC when applying SMOTE or noise during training.

The MAE response of the ANN is displayed in figure 5.15. From this, it can be seen that the model trained on only the original data scores a better MAE at every level of added noise.

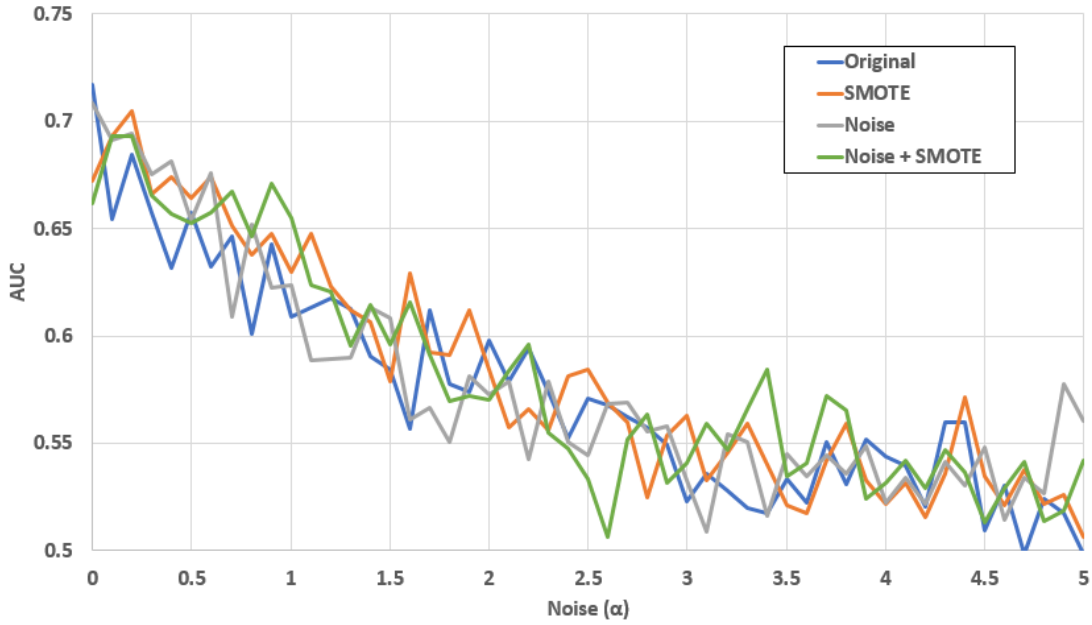


Figure 5.12: ***LSBoost AUC*** *response to added noise with the original data only and data augmentation with SMOTE, added noise and SMOTE with added noise.*

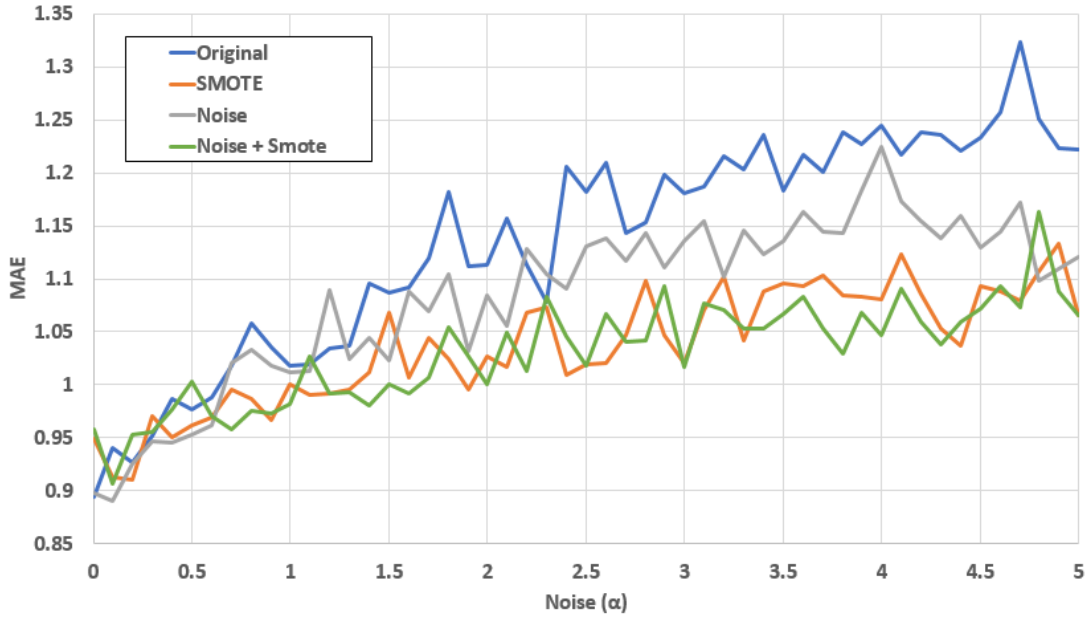


Figure 5.13: ***LSBoost MAE*** *response to added noise with the original data only and data augmentation with SMOTE, added noise and SMOTE with added noise.*

5.7.5 Final Models

As the LSBoost model and the ANN did not respond in the same manner to the added data augmentation steps, separate training schemes should be applied. The final

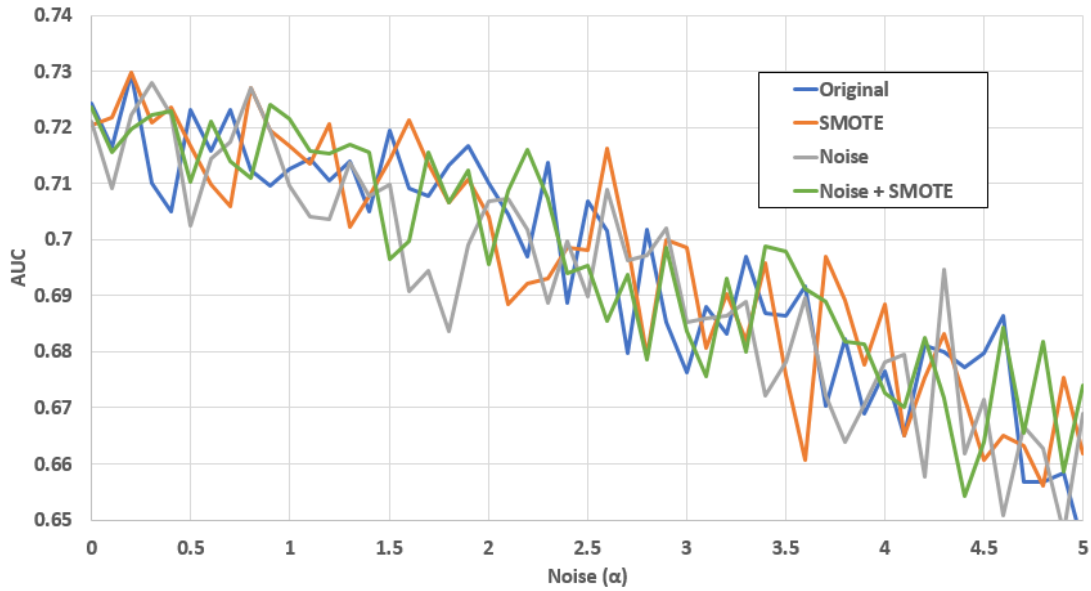


Figure 5.14: ***ANN AUC*** *response to added noise with the original data only and data augmentation with SMOTE, added noise and SMOTE with added noise.*

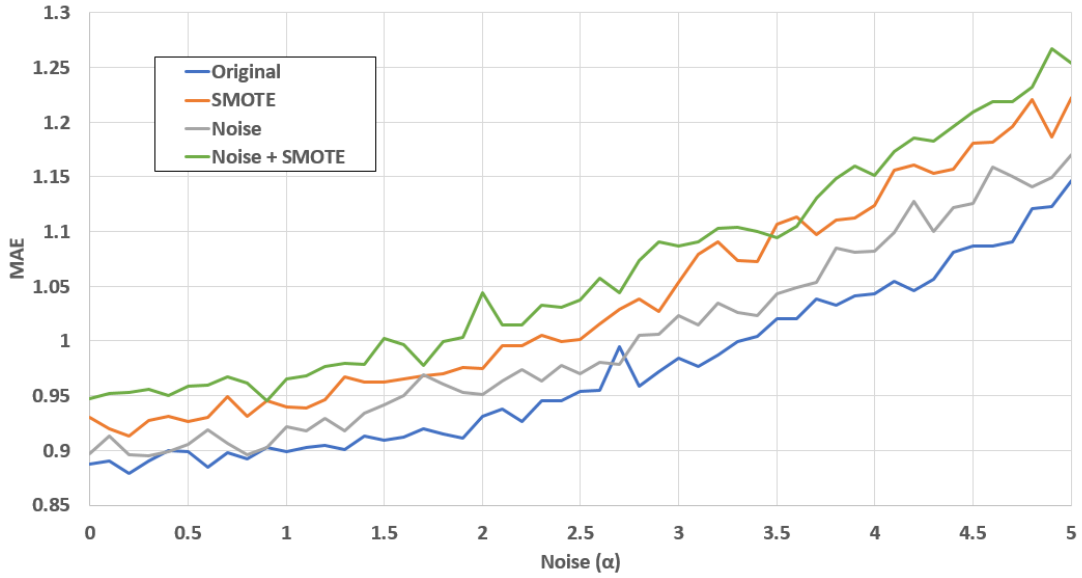


Figure 5.15: ***ANN MAE*** *response to added noise with the original data only and data augmentation with SMOTE, added noise and SMOTE with added noise.*

LSBoost model used SMOTE and additional noise added to the training dataset. The final ANN model used only the original data and a high L2 regularisation of $\lambda = 0.95$. An ensemble model was again produced by taking a weighted average of the LSBoost and ANN models. The AUC and MAE response of these final models is shown in figures 5.16 and 5.17 respectively.

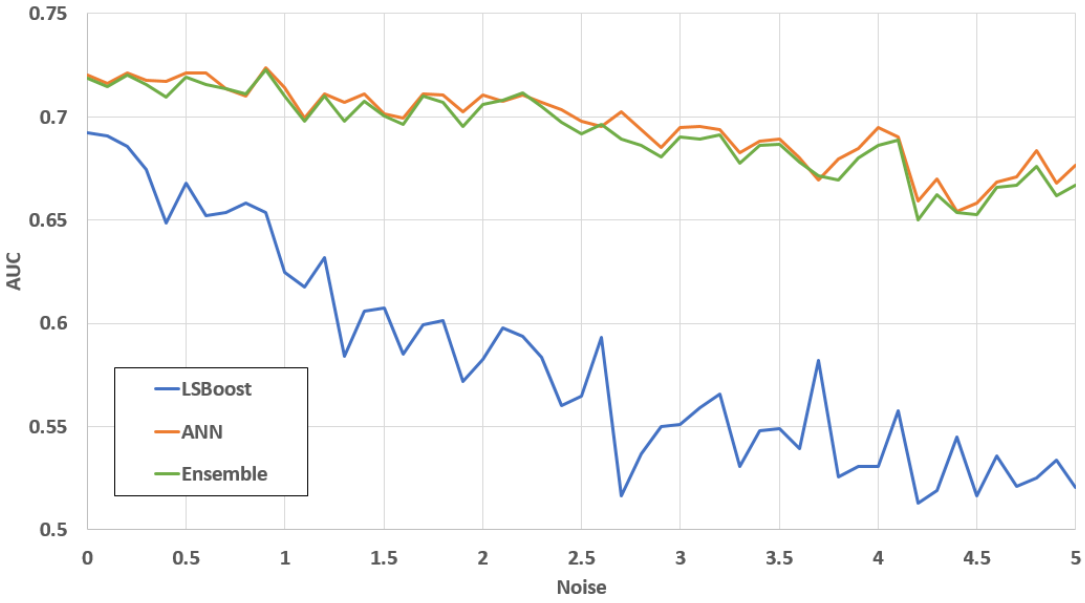


Figure 5.16: ***Final Model AUC****, the AUC for the LSBoost with smote and noise, the ANN with high L2 regularisation and the ensemble combination of the two models.*

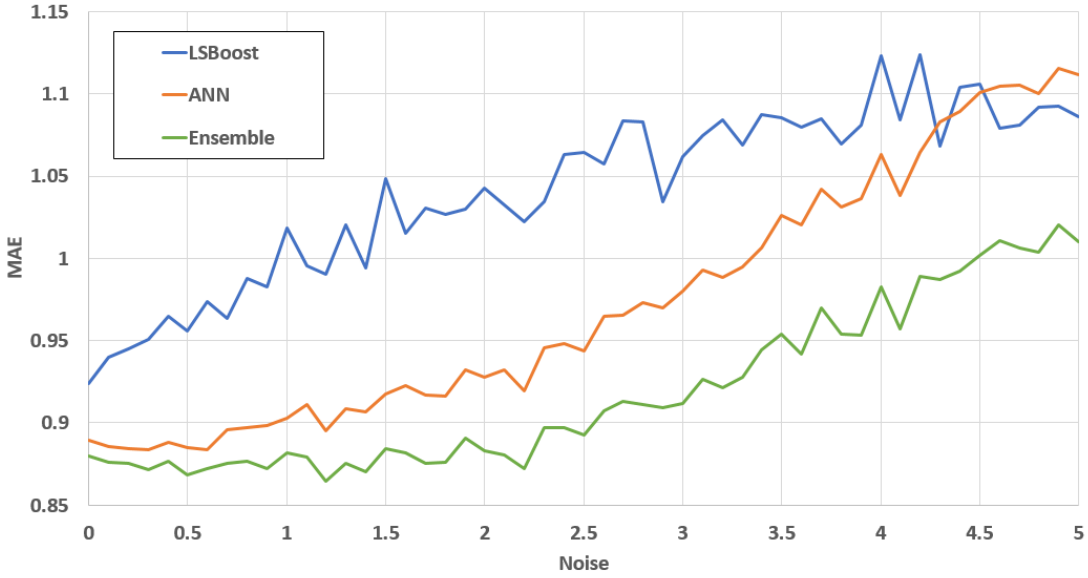


Figure 5.17: ***Final Model MAE****, the MAE for the LSBoost with SMOTE and noise, the ANN with high L2 regularisation and the ensemble combination of the two models.*

5.8 Discussion 2

The work in this section has shown that when using hyperparameter tuning, a model for the prediction of esophageal toxicity from dose map features can be produced to achieve an AUC of 0.707 and 0.691 for the classification of grade ≥ 3 and grade ≥ 2 re-

spectively with a MAE of 0.884 when trying to predict exact toxicity grade. The ANN outperformed the LSBoost model on all metrics when no test set noise was added. Additionally, the ANN outperformed the LSBoost model for all robustness tests provided there is a high level of L2 regularisation present. However, using a weighted ensemble of the two models provides benefit from the LSBoost model by increasing the model robustness to noise in terms of the MAE while not impacting the AUC. To improve the robustness of the models it was observed that a high level of L2 regularisation was beneficial for the ANN and using SMOTE and adding noise during training was beneficial for the LSBoost model. L2 regularisation was not available for the LSBoost model but may also be beneficial if implemented.

Many of the improvements to the robustness of the models would not be apparent if the tests to add noise were not applied. For example, in figure 5.11, when $\alpha = 0$ and no noise is added, all of the models perform equally. A sharp divergence is then observed for lower values of λ with added noise. This highlights the benefit of these robustness tests for reducing overfitting and building robust models.

Future work may investigate the application of systematic errors to the test data to study robustness by altering the esophagus volume to simulate contouring errors or by randomly translating the esophagus volume to simulate positioning errors. Additional data augmentation techniques, such as the use of GANs to generate data, and testing their effect on robustness may also be an interesting avenue of further research.

5.9 Conclusions

While there has been much work on the prediction of adverse events from pre and post RT information, most of this has been focused on the specific imaging modalities and features used for the classification as opposed to focusing on advancing the technical details of the machine learning methods. This is likely due to the lack of large benchmark datasets applicable to the toxicity prediction domain that are necessary to determine if an increase in a prediction metric is due to a change in method or due to uncertainty due to the dataset size. In this chapter, the RTOG-0617 dataset has been applied as the largest publicly available dataset with the necessary data to investigate technical improvements in prediction models for RT esophageal toxicity from dose features.

The initial tests investigated the use of 3D-CNNs, boosted decision trees and the

LKB NTCP model. Applying a 3D-CNN to the dose map produced a marginal improvement when compared to classifying DVH features using decision trees or the LKB NTCP model, the ensemble model using the predictions from both methods did produce a further marginal improvement to the AUC. Treating the grade definitions as a continuous variable for training purposes and using regression based techniques again produced a further marginal increase to the AUC. Finally, grouping all esophageal based adverse events instead of trying to isolate esophagitis provided another increase to the AUC. Combining these three increases lead to a larger increase in the predictive power of the final model. The best performing model achieved an AUC of 0.705.

Further experiments focused on an ANN and LSBoost model, both with feature inputs, to improve model training time and allow for hyperparameter tuning and model robustness tests to be applied. The hyperparameter tuning allowed for the final model to achieve an AUC of 0.707 and 0.691 for grade $\geq$ 3 and grade $\geq$ 2 prediction respectively with a MAE of 0.884 for exact grade prediction. The AUC here is a small improvement on the initial tests without hyperparameter tuning using the CNN and LSBoost model. The robustness of the ANN and LSBoost models was then evaluated by applying random noise to the test sets and observing the degradation of performance. Methods to improve the robustness were tested which found that L2 regularisation was beneficial for the ANNs robustness and SMOTE and random noise-based data augmentation were beneficial for the LSBoost models robustness. Combining both models as an ensemble model was also shown to be effective.

Future work and Direction of the Field

The major limitation currently in the field of radiotherapy outcome prediction is the lack of large benchmark datasets that can be used to test methods with a high level of certainty regarding model performance and to directly compare separate methods easily. This data scarcity is mainly due to data privacy and ethical concerns as, in the UK, over 100,000 patients are treated with RT annually [211] but only a small proportion of these patient's data will be accessible to researchers, usually for specific clinical trials. Additionally, there is an increased workload for clinicians to report RT toxicity outcomes using the CTCAE grading scale, necessary to conduct these prediction studies as these are not recorded during standard care.

Future studies will be focused on applying machine learning methods to larger

datasets in a manner that would be clinically implementable. An example of such a study is the PROSECCA trial which is currently in progress and aims to analyse the data from over 10,000 prostate cancer patients [212]. Additionally, recent advancement in natural language processing means it is now becoming feasible to automatically generate CTCAE, or other, toxicity metrics from clinical notes that are collected during standard care [213, 214] which has been shown to be possible for esophageal toxicity [215].

Chapter Acknowledgements

This chapter was prepared using data from Datasets (RTOG-0617; NCT00533949-D1, D2, D3) from the NCTN/NCORP Data Archive of the National Cancer Institute's (NCI's) National Clinical Trials Network (NCTN). Data was originally collected from clinical trial NCT number NCT00533949, titled "A Randomized Phase III Comparison of Standard-Dose (60 Gy) Versus High-Dose (74 Gy) Conformal Radiotherapy With Concurrent and Consolidation Carboplatin/Paclitaxel +/- Cetuximab (IND #103444) in Patients With Stage IIIA/IIIB Non-Small Cell Lung Cancer". All analyses and conclusions in this thesis are the sole responsibility of the author and do not necessarily reflect the opinions or views of the clinical trial investigators, the NCTN, or the NCI. All clinical and imaging data, including dose maps and RT planning structures, are available from the NCTN/NCORP data archive upon request.

Chapter 6

Registering 4D-PET/CT to Pathology Images and the Automatic Segmentation of Gross Pathology Images

This chapter is focused on methods for the registration of PET/CT and pathology imaging modalities as well as the automatic segmentation of tumour volumes in gross pathology images for NSCLC. Section 6.1 discusses the motivations for this work, sections 6.2 to 6.4 detail the first set of methods, results and discussion relating to the registration task and sections 6.5 to 6.7 detail the second set of methods, results and discussion related to the segmentation task. Finally, the conclusions for the whole chapter are presented in section 6.8.

6.1 Introduction

The work in this chapter is split into two distinct tasks which are; the registration of PET/CT to pathology images and the segmentation of gross pathology image. The motivations for these tasks are discussed in the following subsections.

6.1.1 PET/CT to Pathology Registration

18F-FDG PET/CT is a specialised method of non-invasive in-vivo imaging used in the diagnosis and treatment of most cancers including lung cancer. Currently, PET and CT imaging provide little information about the cellular makeup of a tumour and its environment. Increasing the knowledge that can be gained from these imaging modalities would aid clinicians in making treatment decisions which would improve the treatment and survival rates. To do this, PET/CT images would need to be compared to pathological images of a tumour after it has been surgically removed which would provide cellular information. This would involve a registration of the different imaging modalities so that they are as well aligned as possible. The research undertaken in the first half of this chapter follows on from a previous EngD project completed in 2020 by G. R. March [2]. March produced a registration framework for registering PET/CT images to histopathological slices of a surgically removed lung cancer tumour. While these methods worked well for CT imaging, the motion induced by patients breathing during the scans caused the PET tumour volumes to not correlate with the pathological volumes meaning the images could not be registered. The work here introduces respiratory motion reduction techniques into the PET/CT imaging protocol to improve the PET image quality and allow for its inclusion in the registration to pathology. Additionally, a separate method for the registration of PET to pathology images was developed based on SUV thresholding the PET image.

6.1.2 Gross Pathology Segmentation

Pathology photography can be a useful tool for documenting ground truth anatomy before it has been distorted by the slicing processes that are used for whole slide imaging (WSI). Segmentation of regions in pathology photographs can therefore provide ground truth for the shape of an area, or volume if three dimensions are considered, of a particular anatomical region. Additionally, the current pathological assessment of tumour size, which is a strong predictor of patient outcomes [216], is generally made by measuring the gross length of the tumour across its largest dimension by hand with a ruler which often has to be reevaluated at the time of microscopic assessment [217]. Automatic segmentation of gross tumour area would provide a more reliable method of estimating the tumour volume and cellular load which are the metrics that are being estimated by gross measurements with a ruler. Additionally, if a method of automatic

segmentation of singular tumours is successful it could then be expanded to identify other more subtle nodules that could be easily missed by the naked eye but may have been seen in radiology images and, if used in real-time, this would allow the pathologist to sample these nodules at the time of dissection. An automatic segmentation method for non-small cell lung cancer (NSCLC) tumours in gross pathology photographs, therefore, has both clinical and research applications. The work in this study aims to produce and test a methodology for the automatic semantic segmentation of lung tumours in pathology photographs of specimens that have been surgically removed from patients with NSCLC.

6.1.3 Contributions of This Chapter

In the work detailed in this chapter, methods for image registration between histopathology and PET/CT imaging as well as a method for automatically segmenting gross pathology images are produced. The technical contributions are:

- The application of PET respiratory gating techniques to the registration of histopathology and PET/CT images.
- The development of a method for improving histopathology to PET/CT image registration using a PET volume based registration.
- The first application of deep learning and CNNs to the task of automatic segmentation of tumour regions in gross pathology photographs.

6.2 PET/CT to Pathology Registration: Methods

6.2.1 Patient Recruitment

In order to include respiratory gating for the further advancement of the work by Reines March et al. [2], additional patients had to be recruited to the study to receive a 4D PET/CT scan before the surgical resection of their lung tumour. These patients were to receive an additional PET/CT scan at the West of Scotland PET Centre that included deviceless respiratory gating. They would then undergo a lobectomy operation to remove the lung lobe containing their tumour, which was part of their standard care, at the Golden Jubilee University National Hospital. The specialised pathology processing for the trial was then applied at the Queen Elizabeth University Hospital.

Nine patients were initially enrolled in the trial, although two did not contribute complete datasets. One trial patient was excluded upon reevaluation of their previous CT scan, which revealed that their tumour did not meet the inclusion criteria. The other trial patient with incomplete data experienced a progression of their cancer between the initial PET-CT scan and the trial 4D PET-CT scan which resulted in a cancellation of their scheduled surgery. A table of all the recruited patients' trial IDs is given in Table 6.1. Patients PETPATH-006 and PETPATH-007 were not included in any of the analyses as full datasets were not collected for these patients. The following sections contain the details of the patient recruitment.

Patient ID	Full Dataset?	Notes
PETPATH-001	Yes	
PETPATH-002	Yes	
PETPATH-003	Yes	
PETPATH-004	Yes	
PETPATH-005	Yes	
PETPATH-006	No	Cancer advanced between standard and trial PET scans
PETPATH-007	No	Tumour was not large enough for inclusion
PETPATH-008	Yes	
PETPATH-009	Yes	

Table 6.1: *Trial IDs of all the patients recruited to the trial.*

6.2.1.1 Patient Eligibility Criteria

To be recruited to the study, the patients had to have NSCLC, be age 18 or over and be booked to undergo a curative surgical lobectomy for removal of the tumour. Additionally, the patient eligibility criteria for inclusion in the study in regard to their tumour were:

1. The tumour is identified primarily as a single mass lesion.
2. The main volume of the tumour is located within the lung tissue (i.e. not concentrated in the pulmonary pleura).
3. At least one of the tumour's major axes is larger than 30mm.

6.2.1.2 Patient Recruitment Workflow

Eligible patients were initially identified through a multi-disciplinary team (MDT) which included the pathologist and surgeon who were involved in the trial. The surgeon would then inform eligible patients about the trial in their one-on-one pre-surgery meeting with the patients where they were given the participant information sheet (PIS), the patient would confirm if they are happy to be contacted by phone at a later date by staff at the PET centre to confirm their inclusion in the trial. This phone call would occur at least 24 hours after the meeting with the surgeon to give the patient enough time to make an informed decision.

If patients agreed to participate in the trial, their informed consent was obtained and they underwent an additional 4D PET-CT scan at the West of Scotland PET Centre. Subsequently, patients proceeded with their scheduled surgery, which remained unchanged from their standard care regimen, regardless of their participation in the trial. Following surgery, pathology samples were transported to the Pathology Department at the Queen Elizabeth University Hospital and the trial-specific specialised processing and imaging techniques were employed. These pathology samples were then processed, analysed and reported using the standard procedures of the NHS. This whole workflow is summarised in Figure 6.1.

6.2.1.3 Research Ethics

In order to proceed with patient recruitment, ethical approval was required. An application was made to the Integrated Research Application System (IRAS) with an IRAS ID *287316* and short title "Correlation of pre- and post-operative cancer imaging techniques". This included developing a study protocol and a participant information sheet (PIS), which were adapted from the work by Reines March, et al. [2], among the other required IRAS documents. All clinical trials involving trial patients must be approved by a research ethics committee (REC). The trial was approved by the REC "North West - Preston" on 09/06/2021 and given the REC reference *21/NW/0088*. As the trial involves the administration of a radioactive substance to the trial patients beyond their standard care procedures, approval was required from the Administration of Radioactive Substances Advisory Committee (ARSAC). ARSAC approval was given on 07/01/2022 with the ARSAC reference number *AA-3260*. The trial was registered on clinicaltrials.gov with the identifier *NCT04776291*.

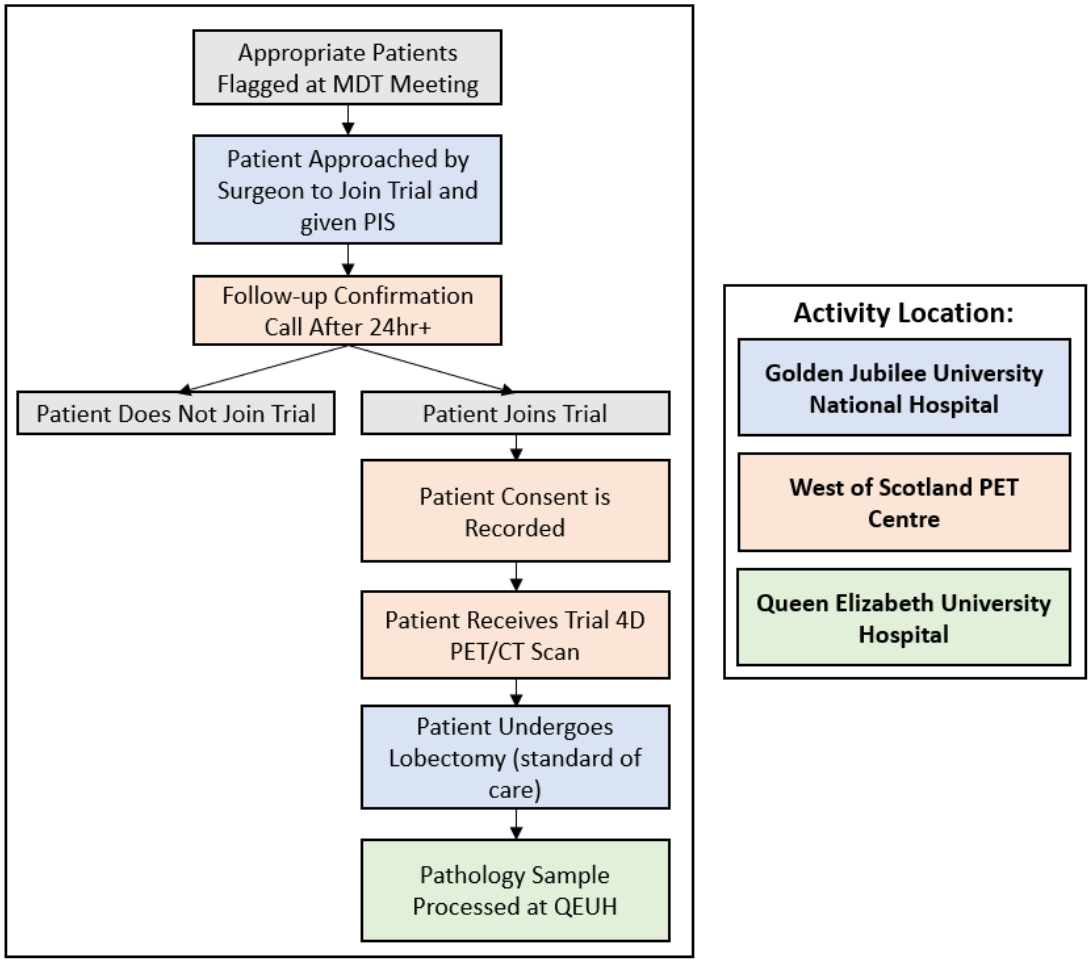


Figure 6.1: *Flowchart depicting the different stages of the patient recruitment, imaging and pathology processing. Locations for each activity is given by the colour scheme.*

6.2.2 4D PET-CT Scan

The trial patients underwent an additional PET-CT scan at the West of Scotland PET Centre. The Siemens Biograph Vision* PET-CT scanner [218] was used for the 4D-PET-CT scans of all trial patients. The 4D-PET-CT scan was taken at a slower speed of bed motion to allow for a higher SNR in the resulting PET images, this allowed for the PET image to be gated without reducing the SNR of the image below a reasonable level. The standard speed of bed motion used for a chest scan of a lung cancer patient at the West of Scotland PET Centre is 1mm/s, for the trial 4D-PET-CT scan, a bed speed of 0.3mm/s was used. The scanned region of anatomy was different in the standard and trial PET-CT scans meaning the increase in the length of time of the scan

*Siemens Healthineers, Siemens AG, Erlangen (Germany)

was not uniform but in general, the standard PET scan took around 8 minutes and the trial scan took around 20 minutes. The Siemens scanner uses deviceless respiratory tracking where the respiratory motion reduction is applied during image reconstruction. This means that multiple methods of respiratory motion reduction can be investigated by applying several image reconstructions with different reconstruction settings. The methods investigated were the Siemens* OncoFreeze and time-based gating with 4 gates algorithms. Additionally, all analysis was also applied to a reconstructed version of the image with no respiratory motion reduction i.e. the standard PET image.

6.2.3 Pathology Specimen Processing

The pathology specimen processing and imaging methods were the same as described in the work by Reines March, et al. [2] which was detailed in the literature review Section 3.3.2. This involved suspending the lung specimen in agar prior to slicing and imaging with a digital camera† in a specialised slicing rig. This process produced a gross pathology photograph stack with a resolution of 5mm in the z-axis. After imaging, the gross pathology photographs were manually segmented using the ImageJ software [219] by a consultant pathologist to provide the pathology tumour volume.

6.2.4 Pathology and CT Image Interpolation and Registration

To align the pathology and CT images in the same coordinate system, the pixel dimensions of the tumour segmentation in each modality had to be matched. The CT images had an original image resolution of 0.97×0.97×1.5mm. The original resolution of the pathology image stack was 5mm in the z axis and around 0.075mm in the x-y plane which depended on the camera set up on each specific sample. Following the results of the work by Reines March et al.[2], the interpolation was applied with a cubic spline based interpolation with a curvature based extrapolation for extrema region estimation. The CT to pathology registration developed by Reines March et al. was applied, again none of the methodology was altered here so it has already been discussed in Section 3.3.2 and will not be discussed in detail here. This registration process involved an initial alignment of the minimum bounding box of the tumour and main airway of the lung lobe in both modalities followed by an iterative closest points registration of the

*Siemens Healthineers, Siemens AG, Erlangen (Germany)
†Canon EOS M3 digital camera, Canon Inc., Tokyo (Japan)

surface points of the tumour volume in both modalities.

6.2.5 PET Image Analysis

To quantify the effects of the respiratory gating, several approaches to image analysis were taken. For all analysis, the tumour volume in the PET images was segmented by initially isolating the tumour region manually with a bounding box and then thresholding the tumour based on a percentage of the tumours SUV_{Max}. To make sure any necrotic tumour regions, which have low metabolic activity and therefore a low 18F-FDG uptake, were included in the segmentation, morphological operations were applied to this thresholded segmentation. A morphological dilation operation was applied with a 3D spherical structuring element of radius 3mm. This was followed by filling any holes in the 3D volume and finally an erosion operation by the same structuring element.

Comparison of the Tumour Volume in Different Modalities The absolute tumour volume, expressed in mm^3, was calculated for the gross pathology, CT and standard PET tumour segmentations. The PET tumour segmentations investigated were 0.5 SUV_{Max} and 0.3 SUV_{Max}. Tumour volumes were calculated by summing the number of pixels in the masks and multiplying by the voxel volume for each modality. For the PET modality, this was repeated for the separate respiratory motion reduction methods with 0.3 SUV_{Max} thresholding.

Matching the PET volume to Pathology Segmenting the PET tumour based on an SUV threshold as opposed to using a manual segmentation allows for this threshold to be chosen such that the total PET volume matches the total pathology volume. This was applied by initially segmenting the PET tumour using 0.3 SUV_{Max} and then calculating the difference between the total PET volume produced by this and the pathology volume. The SUV threshold was multiplied by 0.99 if the PET volume was larger than the pathology volume and 1.01 if it was smaller. The process iteratively repeated until the PET and pathology volumes were equal at which point the optimum SUV threshold was recorded.

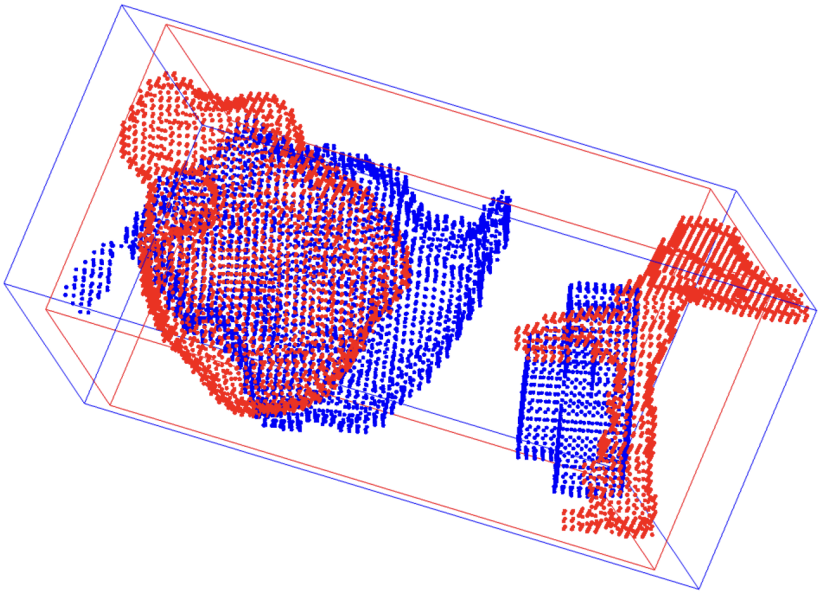


Figure 6.2: *Example minimum bounding box pre-registration registration between the pathology (blue) and PET (red) tumour and airway joint volumes.*

6.2.6 PET to Pathology Registration

As the absolute volume of the tumours in the CT images poorly matches the pathology tumour volumes, presented in the results in section 6.3.3, it may be beneficial to replace the CT volume in the rigid registration between the CT and pathology with the PET volume. As the PET segmentation is based on a threshold, the PET volume can be set to match the pathology volume by taking the optimum threshold value to match the absolute pathology volume defined from the manual segmentations. An issue with this compared to manual segmentations by clinicians is that this will be a segmentation of regions with a high 18F-FDG uptake which may not exactly correspond to a tumour regions. This may very closely match a tumour segmentation for some patients depending on the level of inflammation surrounding the tumour which may also have a high uptake of 18F-FDG. Considering the mismatch between the volumes of the CT and pathology segmentations, the manual CT segmentation cannot be considered to consist purely of tumour volume either.

To align the PET volume with the pathology volume the same registration process as for the CT to pathology registration was used. The pre-registration step involving the alignment of the minimum bounding box of the tumour and main airway in the lobe presents an additional challenge for the PET modality as the airways are not visible

in the PET image. This is easily overcome by simply taking the airway segmentation from the CT image and placing it in the PET coordinate space as the PET and CT images are registered by the co-imaging of these two modalities. An example of the registered pathology and PET minimum bounding boxes is given in Figure 6.2 where the PET surface points and bounding box are presented in red and the pathology is presented in blue. After this pre-registration step, the registration process is identical to the CT to pathology registration.

6.3 PET/CT to Pathology Registration: Results

6.3.1 Qualitative Comments on Individual Patients

In this section, any notable details for specific trial patients' imaging are given.

6.3.1.1 PETPATH-001

Trial patient PETPATH-001 displayed two features visible in their imaging that are worth noting, these are; a highly necrotic tumour causing a completely hollow tumour core and a high level of respiratory motion visible in the static. These can both be seen in Figure 6.3 which displays the tumour region in the frontal plane from the standard PET and gated PET images. A high level of blurring is visible in the standard image which is greatly reduced in the gated image. The necrotic core of the tumour is most easily seen in Figure 6.3 (b) as a region of low image intensity surrounded by a ring of high intensity. The necrotic tumour core for this patient presents challenges for the registration as the assumption that the tumour volume will remain rigid between all separate modalities is less likely to be true. The high level of respiratory motion present in the standard PET image allows for a purely qualitative justification for the use of the respiratory motion reduction methods as there is clearly an improvement in Figure 6.3 (b) compared to Figure 6.3 (a).

6.3.1.2 All Other Patients

For all other patients aside from PETPATH-001, there were no notable features of their tumours or large levels of respiratory motion observed. Even without a large level of respiratory motion, there can still be some qualitatively observable improvements to the PET image quality by using respiratory motion reduction techniques. As an example,

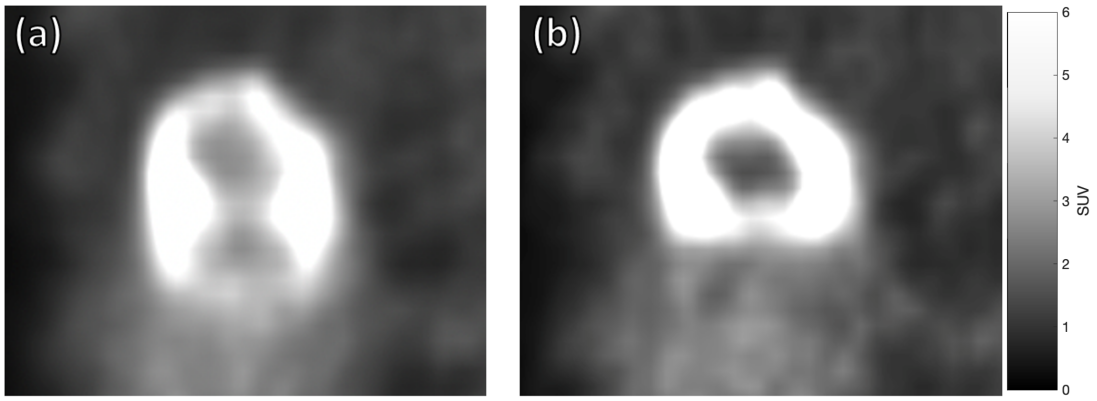


Figure 6.3: *PET images using* ***(a)*** *standard PET and* ***(b)*** *gated PET of the lung tumour of patient PETPATH-001 displayed in the frontal plane.*

the same slice of the standard and gated PET images for patient PETPATH-002 are displayed in Figure 6.4 (a) and (b) respectively. Subtle differences can be observed between these images, in the gated image it appears that there is more texture within the tumour. This can be highlighted by taking the difference of the two images which is displayed in Figure 6.4 (c).

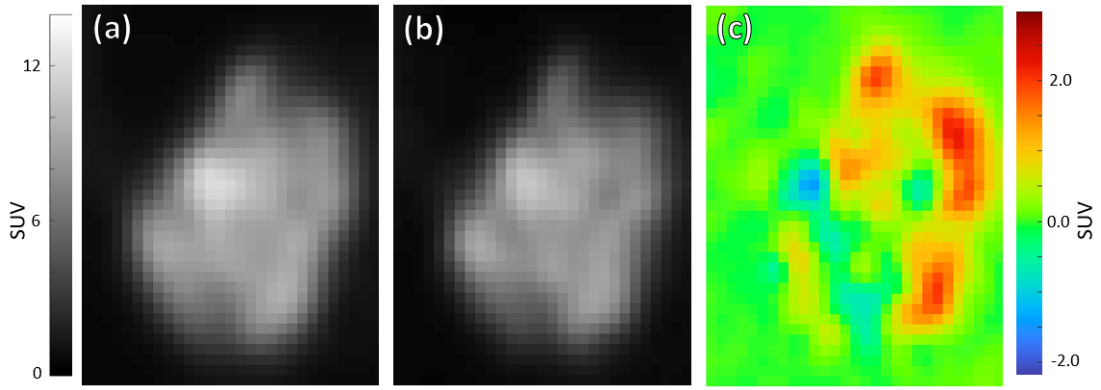


Figure 6.4: ***(a)*** *The standard PET image,* ***(b)*** *gated PET image and* ***(c)*** *the difference between the standard and gated PET images for patient PETPATH-002.*

6.3.2 PET Image Metrics

6.3.2.1 PET SUV Max

The SUV_{max} values for each patient's tumour in the PET image for each different respiratory motion reduction method are presented in Table 6.2. For both the OncoFreeze and gated methods of thresholding, the tumour SUV_{max} is observed to be higher than

the standard PET tumour SUV_{max} for all patients except patient PETPATH-004 for the OncoFreeze reconstruction. This is an expected outcome as, when there is no motion reduction included in the PET reconstruction, regions within the tumour will be blurred which reduces the maximum intensity of the PET image by reducing the sharpness of the highest intensity points.

	SUV_{max}		
Patient ID	**Standard**	**OncoFreeze**	**Gated**
001	6.07	7.35 (21.18%)	6.58 (8.52%)
002	13.98	14.45 (3.36%)	14.94 (6.89%)
003	17.35	17.8 (2.73%)	18.34 (5.78%)
004	10.22	10.16 (-0.51%)	10.63 (4.09%)
005	5.76	6.25 (8.49%)	6.44 (11.85 %)
008	4.68	4.92 (5.20%)	4.98 (6.49%)
009	18.38	18.78 (2.16%)	19.25 (4.71%)

Table 6.2: *SUV_{max} of the tumour for each method of PET respiratory motion reduction for all trial patients.*

6.3.2.2 PET and CT Tumour Center of Mass Differences

One benefit of using respiratory motion reduction in the PET image is the potential to improve the alignment between the PET and CT images by reconstructing the PET image at the point in the breathing cycle that best matches the CT image. For the gated reconstructions, the gate that most closely matched the CT image in terms of the tumour location was manually selected. For the OncoFreeze algorithm, there is no way to specify which point in the breathing cycle is chosen as the fixed point to align the rest of the PET counts to as this is automatically selected as part of the algorithm and as it is proprietary technology there is no way to alter this algorithm. To quantify the alignment of the PET and CT tumours, the centre of mass of the tumour segmentation in both modalities was found and the difference between the two was calculated. For the PET modality, the 0.3 SUV_{Max} thresholded segmentation was used. The difference between the centre of mass of the two modalities is presented in Table 6.3. Here it can be seen that for every patient the gated method produces a better alignment in terms of the centre of mass than the standard PET image. For the OncoFreeze method, as the point in the breathing cycle cannot be chosen to match the CT image, the alignment of the centre of mass is often worse than the standard PET reconstruction.

	Center of Mass Difference (mm)		
Patient ID	**Standard**	**OncoFreeze**	**Gated**
001	9.9	12.6	3.7
002	1.4	3.1	1.4
003	1.8	1.5	1.2
004	8.4	8.1	7.9
005	12.2	11.2	7.7
008	6.5	6.1	5.9
009	4.4	5.5	3.0

Table 6.3: *The difference between the PET and CT tumour segmentation center of mass in mm for all PET reconstructions. A 0.3 SUV_{Max} threshold was used for all PET segmentations.*

6.3.3 Segmentations

The absolute volume of tumour in the pathology, CT and standard PET modalities are given in Table 6.4 where the percentage of the pathology tumour size is given for the CT and PET modalities. It can be seen here that, aside from patient 001, which is a special case due to their necrotic tumour, the manual segmentation of the tumour in the CT modality overestimates the tumour size compared to the pathology, this was also observed by Reines March et al. [2]. This may partially be due to the way in which radiation oncologists are trained for manual segmentation where it is generally preferable to encompass healthy tissue within the delineated tumour boundaries rather than risk omitting actual tumour tissue. As a consequence, this often leads to an overestimation of tumour volume during segmentation procedures. The PET thresholding based on a percentage of SUV_{Max} also generally does not line up well with the pathology volumes which was also observed by Reines March et al. [2].

The PET tumour volumes when thresholding based on 0.3 SUV_{Max} for all the different methods of respiratory motion reduction are presented in Table 6.5. The percentage change from the standard PET image is also presented for the OncoFreeze and gated PET images. Here it can be seen that for both the OncoFreeze and gated PET images, the estimated tumour volume is reduced. This is what would be expected from a reduction in the respiratory motion as the larger size of the standard PET tumour volume is likely due to respiratory blurring of the image.

Patient ID	Volume (mm^3)			
	Pathology	CT	PET (0.5 SUV_{Max})	PET (0.3 SUV_{Max})
001	25258	16806 (66.5%)	7873 (31.1%)	29296 (116.0%)
002	37454	47925 (128.0%	20181 (53.9%)	35520 (94.8%)
003	19101	28873 (151.2%)	13607 (71.2%)	22240 (116.4%)
004	5428	18371 (338.4%)	3879 (71.4%)	10544 (194.3%)
005	8599	13276 (154.4%)	12814 (149.0%)	39661 (461.2%)
008	14273	21444 (150.3%)	6975 (48.9%)	18507 (129.7%)
009	32110	61948 (192.9%)	26715 (83.2%)	37382 (116.4%)

Table 6.4: *The volume of the tumour for each patient in each of the different modalities. The pathology and CT modalities are manual segmentations and PET volumes are based on thresholding as a percentage of SUV_{Max}. For the CT and PET modalities the volume is also expressed as a percentage of the pathology volume.*

Patient	0.3 SUV_{Max} Thresholded Volume (mm^3)		
	Standard	OncoFreeze	Gated
001	29297	22155 (-24.4%)	26862 (-8.3%)
002	35520	33895 (-4.6%)	34221 (-3.7%)
003	22240	21521 (-3.2%)	21251 (-4.4%)
004	10544	10340 (-1.9%)	10013 (-5.0%)
005	39661	32522 (-18.0%)	34001 (-14.3%)
008	18508	17200 (-7.1%)	17372 (-6.1%)
009	37383	36173 (-3.2%)	36492 (-2.4%)

Table 6.5: *The tumour volume for each patient in both the standard, OncoFreeze and gated PET images based on a 0.3 SUV_{max} threshold. The percentage change from standard when using motion reduction is also presented.*

6.3.4 Pathology to PET and CT Image Registration

The results of both the pathology to CT and pathology to PET registrations are presented in this section. Dice scores for the registrations are presented in Table 6.6 using both the CT volume and the PET volume from all different reconstructions. It should be noted that the PET image was thresholded to match the pathology volume which would allow for higher Dice scores to be achieved in general. For all patients, it is observed that matching the pathology to the PET volume achieves a better registration in terms of the dice score than matching to the CT volume. There is no way to know if the PET volumes truly match the pathology in terms of the anatomy, all that can be relied upon is the fact that regions of high uptake in the PET image are more likely to be tumour tissue. This is likely to provide a more accurate registration

than the CT registration due to the large mismatch between the CT and pathology volumes as shown in Table 6.4. Interestingly, comparing the SUV_{max} values in Table 6.2 to the DICE scores of the pathology to PET registration in Table 6.6 it is observed that the DICE score increases as the tumour SUV_{max} increases. This observation implies that tumours with high uptake levels are more likely to yield segmented volumes that closely align with the ground truth pathology volume when segmented utilising a threshold derived from the SUV in PET images. This correlation can be explained by the higher tumour-to-background uptake ratio present in such tumours, reducing the probability of non-cancerous tissue regions being erroneously included within the PET tumour threshold.

Patient ID	**Dice Score**			
	CT	**PET**		
		Standard	**OncoFreeze**	**4 Gates**
001	0.678	0.736	0.724	0.743
002	0.770	0.798	0.800	0.796
003	0.717	0.814	0.821	0.824
004	0.453	0.761	0.754	0.759
005	0.608	0.756	0.736	0.768
008	0.758	0.748	0.757	0.779
009	0.667	0.830	0.807	0.829

Table 6.6: *Dice scores for the pathology to CT and pathology to PET tumour volume registration. The CT volume was based on the manual segmentation and The PET volume was thresholded such that it matched the pathology volume.*

While it is challenging to display the registered volumes in a meaningful way in the 2D medium of this thesis, Figures 6.6 and 6.5 display images of the registered point clouds when using the PET and CT volumes respectively for all patients. Only the points of the volume surfaces are displayed with the pathology points displayed in blue and the PET or CT points displayed in red. In Figure 6.5 the effects of the mismatched CT and pathology volumes can clearly be seen, for trial patients PETPATH-004 and PETPATH-005 in particular the mismatch is most apparent giving little confidence in the alignment beyond the initial alignment of the minimum bounding box. When looking at Figure 6.6 the PET tumour shape does appear to more closely align with the pathology in terms of shape.

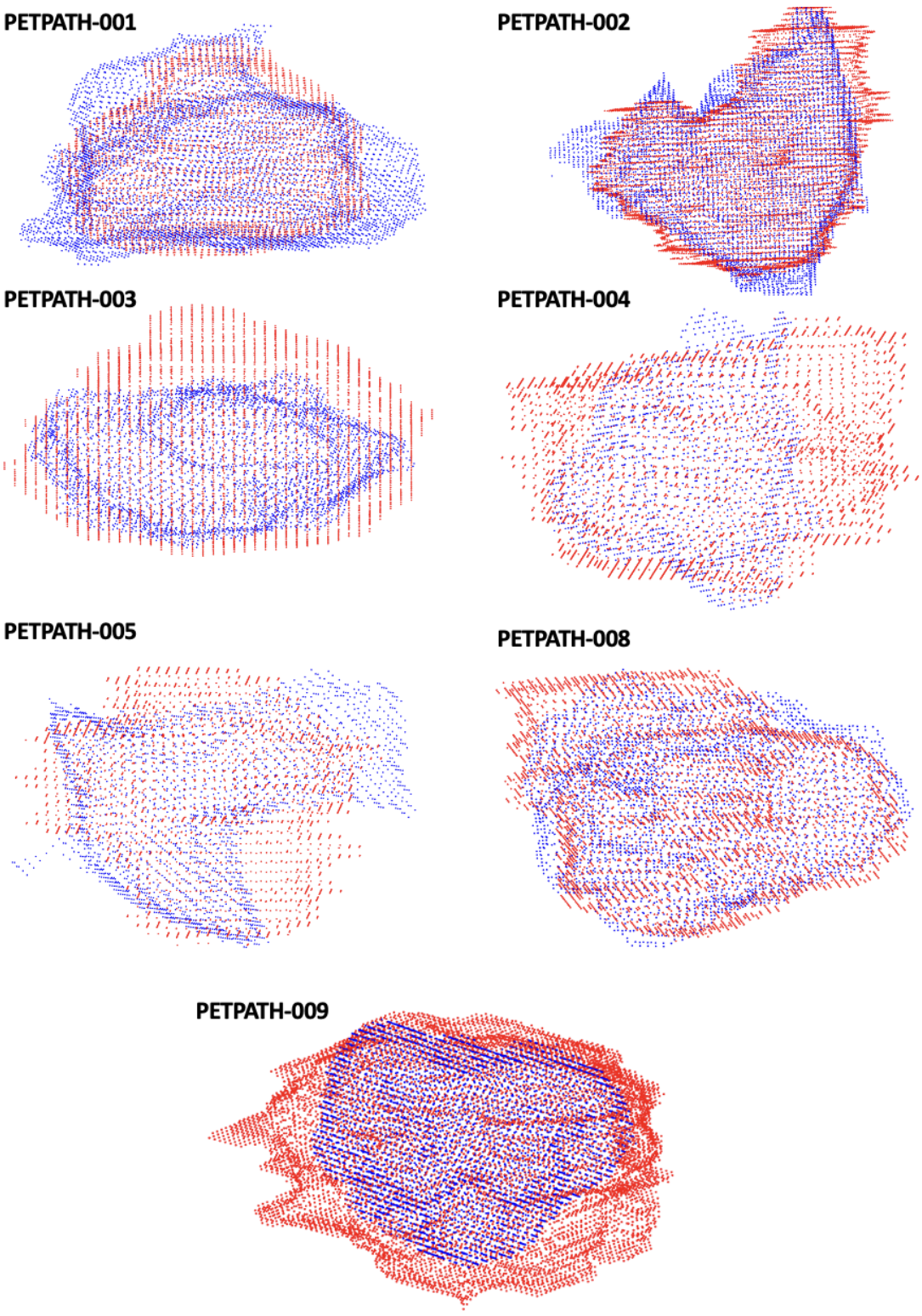


Figure 6.5: *The surface points of the registered CT and pathology point clouds. Pathology points are displayed as blue and CT points are displayed as red.*

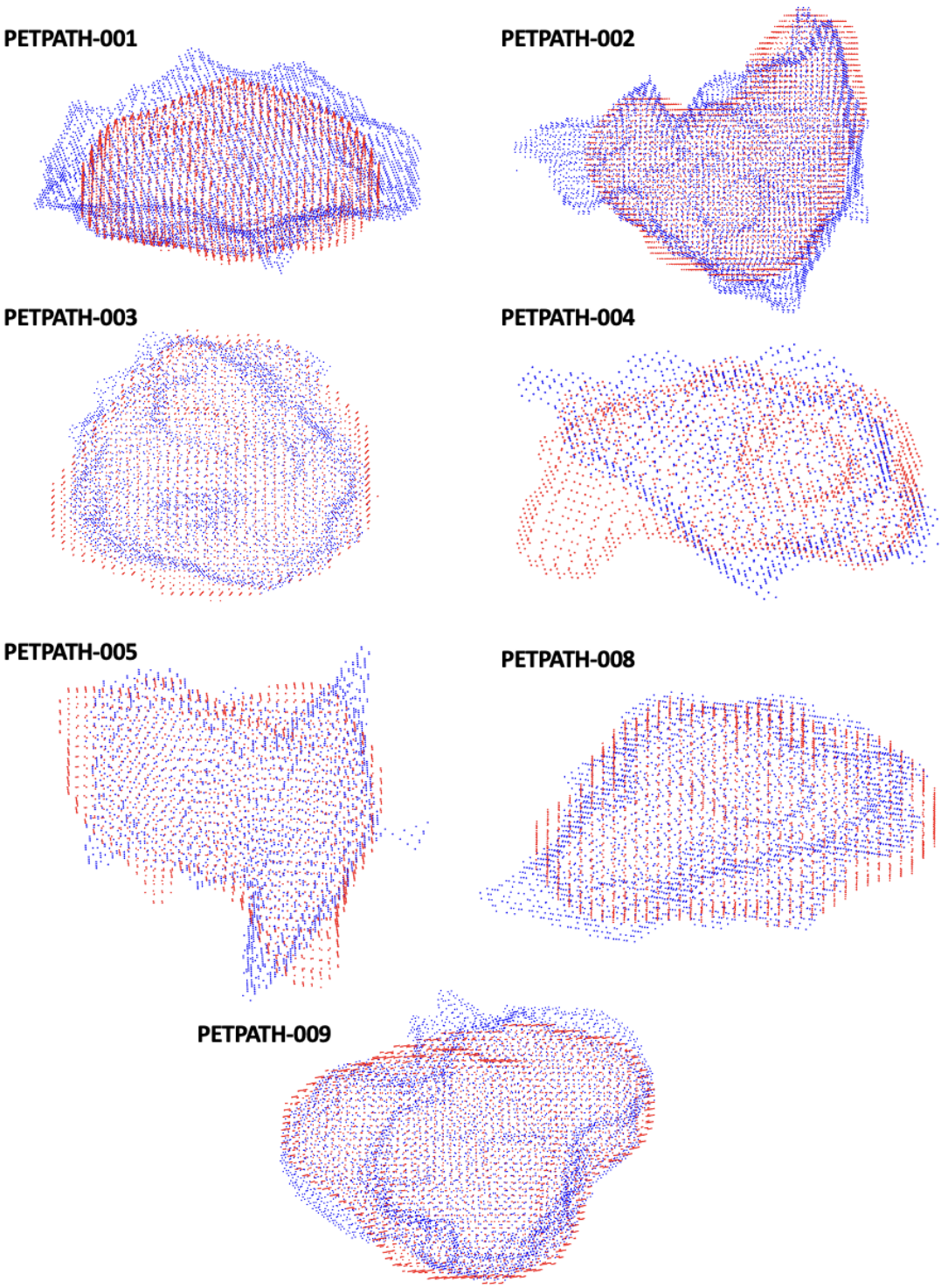


Figure 6.6: *The surface points of the registered PET and pathology point clouds when using the gated PET image. Pathology points are displayed as blue and PET points are displayed as red.*

6.4 PET/CT to Pathology Registration: Discussion

The main focus of this work has been to add respiratory motion reduction techniques to the registration workflow introduced by Reines March et al.[2]. This has been achieved by using the deviceless respiratory motion reduction methods available on the Siemens Biograph Vision PET-CT scanner. These methods are the OncoFreeze algorithm and a time-based gating algorithm. As there is no ground truth to determine if the motion reduction methods have improved the accuracy of the PET image, the evidence justifying its use must come from several sources. Qualitatively, when there is a large level of motion, such as for patient PETPATH-001, the only justification required is to visually compare the standard and gated PET images to observe the reduction in motion blurring. In most cases, this is not observed. Comparing the standard and respiratory motion reduced PET metrics of the tumour volume such as the SUV_{Max}, the size of the thresholded tumour volume and the difference in the tumour centre of mass between the PET and CT images provides evidence in these cases that the motion reduction techniques are still beneficial.

A problem with the OncoFreeze method is that the point in the breathing cycle that is used for image reconstruction is not matched to the point in the breathing cycle captured in the CT scan, this was quantitatively observed in section 6.3.2.2 where the difference of the tumour centre of mass between the PET and CT modalities was calculated. This means that the alignment between the PET and CT modalities will be poorer and may require an additional registration step. Additionally, the different points in the breathing cycle may cause the tumour shape to change, again decreasing the alignment, this will be a particular concern for patients with highly necrotic tumours such as patient PETPATH-001. Using the time-gated method allows for a closer alignment to the CT image by manually selecting the gate that has the best match to the point in the breathing cycle that the CT image was taken. Here, comparing the centre of mass between the PET and CT shows a closer alignment for all patients than for the standard PET image.

The work by Reines March et al. [2] only used a PET threshold of 0.5 SUV_{Max} to segment the PET image. The effectiveness of this method in achieving accurate tumour segmentation is uncertain, primarily due to the interpatient variability in SUV-based tumour delineation which has been documented in previous studies [220, 221]. As this

study is focused purely on the registration of pathology and PET/CT for the purpose of comparing the information in all modalities in a common coordinate system, the PET segmentation can be informed by the manual pathology segmentation which is the best approximation for the ground truth. This has allowed for the advancement of the registration workflow by including a registration based on the PET volume that has been SUV thresholded to match the pathology volume. This is likely to improve the registration performance when the CT segmentation poorly matches the pathology. There are some issues with using the PET image to define the tumour volume for the registration. One issue is that due to the nature of PET imaging, the PET volumes will have smooth edges which may hinder the registration by reducing the number and quality of surface features that can be used to guide the registration. Also, as discussed previously, the thresholding technique may erroneously include regions of non-cancerous tissue in the PET image if they have high enough uptake which can be possible due to inflammation or proximity to metabolically active organs which would again reduce the accuracy of the registration. Additionally, while this approach works when using the 18F-FDG radiotracer, this PET based segmentation would be less applicable when using more specialised radiotracers such as 18F-FMISO [222] which highlights hypoxic tumour regions. Even with these sources of error, establishing the PET volume as a viable volume for this registration provides an alternative option for registration when there is poor agreement between the CT and pathology modalities.

6.5 Gross Pathology Segmentation: Methods

The methods here detail the application of deep learning for the automatic segmentation of tumour regions in gross pathology photographs.

6.5.1 Datasets

In addition to the gross pathology data collected in section 6.2.3 there was additional data available for the automatic pathology segmentation task. All of the data consists of photographs of lung specimens that have been surgically resected from patients with NSCLC and the manual segmentations of the tumour regions. Some of this additional data comes from the work by Reines March et al. [2] which processed and imaged the pathology in the same way as earlier in this chapter. This data was combined with

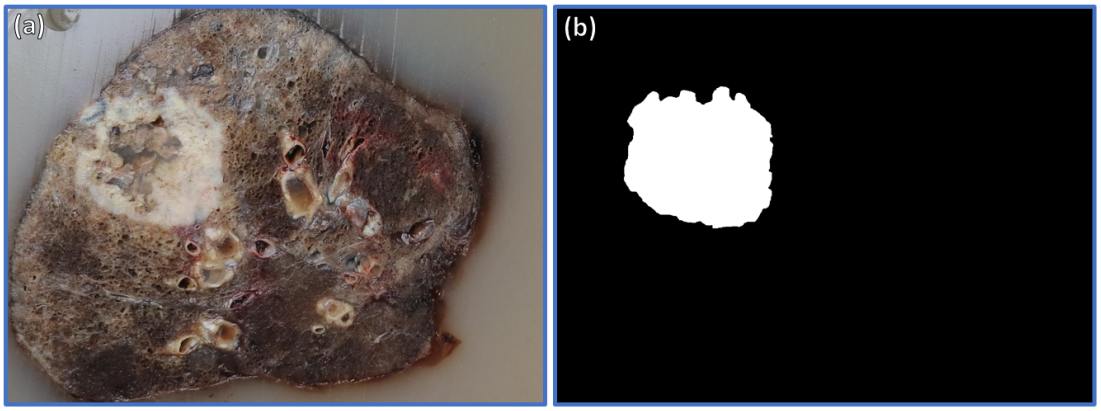


Figure 6.7: *An example of* ***(a)*** *a gross pathology photograph and* ***(b)*** *its corresponding manual tumour segmentation from dataset-A.*

Table 6.7: *Pathology photograph dataset information.*

Dataset	Number of Patients	Number of Images
Dataset-A	16	116
Dataset-B	6	52
Total	22	168

the data collected in this chapter into one larger dataset which will be referred to from now on as dataset-A. Dataset-A contains gross pathology photographs of the entire lung lobes that were inflated with agar and suspended in agar before being sliced at 5mm intervals with photographs being taken after every slice was removed. During the collection of this dataset, care was taken in the lighting of the samples as well as partially drying the samples so that minimal reflection and maximum tissue contrast could be produced. An example of a photograph and its corresponding manual segmentation can be seen in Figure 6.7.

Additionally, a second dataset, which will be referred to as dataset-B, was available. This consists of pathology specimens that were sliced and photographed freehand with only the standard pathology lab lighting used to light the specimens. In this dataset, less care was taken to remove excessive moisture and reflections on the samples meaning these photographs are of poorer quality than dataset-A. The number of patients and images in each dataset is summarised in table 6.7.

Examples of gross pathology photographs from four separate patients are shown in figure 6.8. This shows the variability in the tumours and some of the different features that can be seen. For example figure 6.8 (b) shows a tumour with a large necrotic core whereas figure 6.8 (d) shows a tumour with no necrotic regions. Figures 6.8 (a),(b)

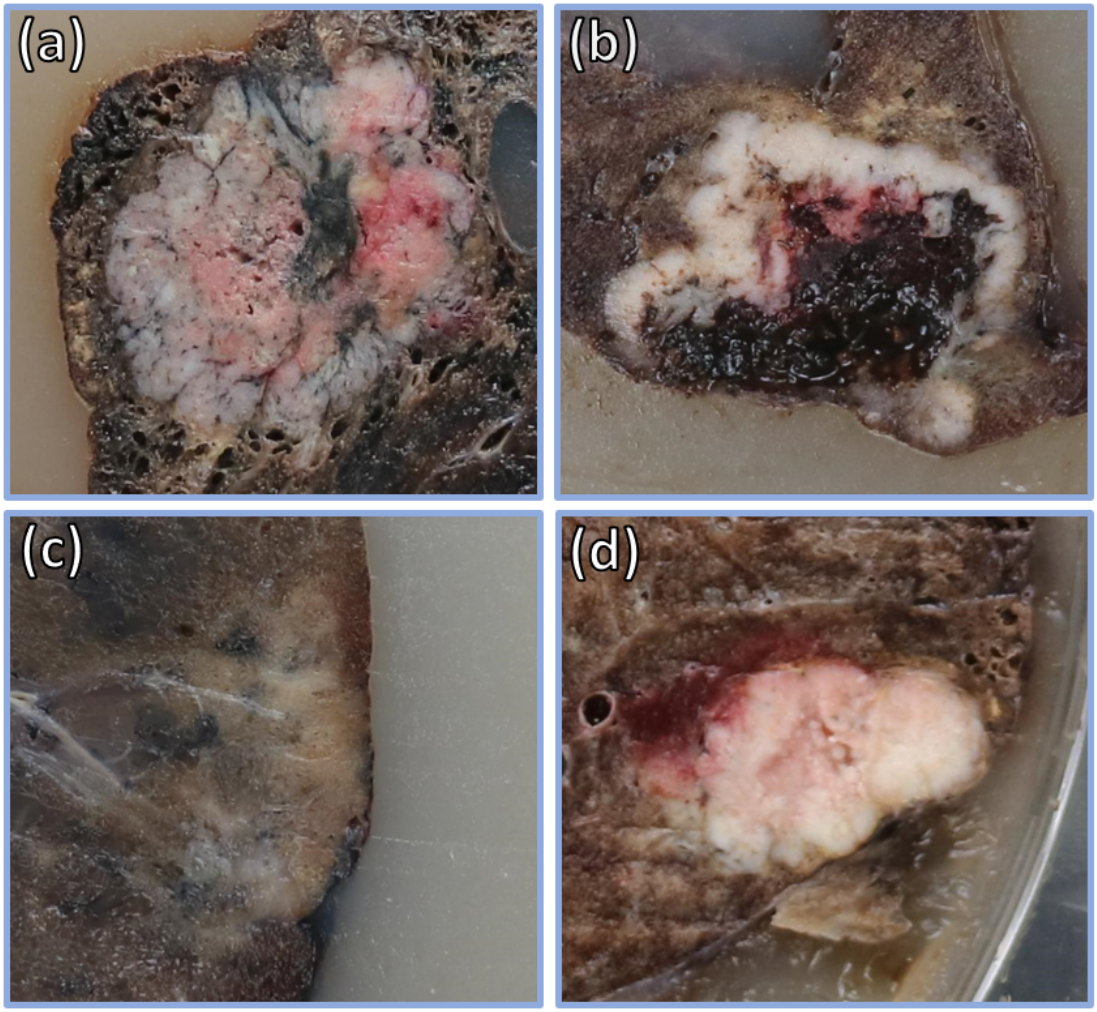


Figure 6.8: *Examples of tumour regions in gross photographs for four separate patients.*

and (d) all display regions with increased red or pink colour, this is due to the tumour reducing the quality of fixation in these regions. Figure 6.8 (c) shows an example where the tumour is displaying poorer contrast to the healthy tissue than the other examples, likely due to this example being an adenocarcinoma tumour.

6.5.2 Image Pre-processing and Data Augmentation

Before the pathology images and labels were used in training some pre-processing steps were applied.

Parameter	Value
k	2
Number of clustering repetitions	3
Max Iterations	100
Accuracy Threshold	1.00e-04

Table 6.8: *K-means segmentation parameters for the background segmentation of Dataset-B.*

6.5.2.1 Non-Tissue Background Removal

Many of the images contain a large amount of background area compared to the area of lung tissue. This was reduced by manually cropping the images down to a rectangular shape closely bounding the lung tissue. The aim of this step was to reduce the computational load of training the models by decreasing the image sizes and to get the CNN used for segmentation to focus more on areas of the lung specimen.

In all of the pathology photographs, the non-tissue background is well distinguished from the tissue regions of the image. This allows for the application of non-learning based segmentation techniques to create a mask that removes the background regions. As the pathology samples in dataset-A and dataset-B were prepared using different methodologies, the background regions in both datasets are reasonably different. Only Dataset-A contains regions of agar, these regions are of a similar colour to much of the tissue regions. The tissue has also been inflated with agar causing there to be regions of agar within the outer tissue boundary. This means a colour-based approach to background segmentation is not appropriate. Here the spectral residual saliency detection approach, as described in [223], was used. This was applied in MATLAB using methods adapted from [224].

Dataset-B contains samples imaged either on a pathology slicing board or a perforated metal pathology workstation. There are also often separate objects such as rulers contained within the images. These images are less suited to the spectral residual approach used for dataset-A but work well with a colour-based approach due to the clear colour contrast between the tissue and non-tissue regions of the image. For dataset-B a k-means clustering approach [225] was applied for colour segmentation where in this case $k = 2$. The parameters used in the k-means clustering algorithm are shown in table 6.8.

Once a mask of the non-tissue background was produced it was used to set the

pixels of the non-tissue background to intensity values of zero.

6.5.2.2 Data Augmentation

The images were then converted to patches of size 224×224 for use with the CNN. When converting the images to patches, an overlap of 50% was introduced in both the x and y image directions to conserve spatial information occurring at the borders of the patches [226]. This resulted in a quadrupled patch count compared to the scenario where overlap is not accounted for. It is important to note that, during one epoch of training on the patches, the network will therefore encounter the same data four times due to this augmentation. A random rotation of the images between 0 and 360 degrees and a random zoom between 0.8 and 1.5 times was applied to the images after every epoch of training.

6.5.3 Loss Functions

The choice of loss function when training a deep learning based semantic segmentation model can have a large impact on the performance of the model. This is especially true for problems with unbalanced datasets where a model may greatly focus on increasing the accuracy of the class with the most instances causing the accuracy of segmentation of the underrepresented class to be low. For this particular application and more generally in many oncology based segmentation problems, the tumour class is underrepresented compared to the background class but would be considered the more important class to accurately segment. Both balanced cross-entropy (BCE) loss and dice loss were applied. For binary segmentation problems, the balanced cross-entropy loss is expressed by equation 6.1 [227].

$$L_{BCE}(y, \hat{y}) = -\frac{1}{N}\sum_{n=1}^{N}(\beta y_n log(p_n) + (1-\beta)(1-y_n)log(1-p_n)) \tag{6.1}$$

Where L_{BCE} is the balanced cross-entropy loss, N is the total number of individual pixels n. y_n is a ground truth pixel value, p_n is a predicted pixel probability outcome and β is a factor used to apply a weighting to the classes.

As the problem presented in this study is a two-class classification problem, only the Dice loss function can be used. This is described in equation 6.2 where L_{Dice} is the Dice loss [227].

$$L_{Dice}(y_n, p_n) = 1 - \frac{\sum_{n=1}^{N} 2y_n p_n}{\sum_{n=1}^{N} y_n + \sum_{n=1}^{N} p_n} \tag{6.2}$$

6.5.4 Segmentation Model

For the semantic segmentation task, an ensemble-based deep learning approach was applied. This involved training multiple separate deep-learning models and combining the output segmentations into a single averaged segmentation. This was followed by post-processing of the ensemble model output through background masking and morphological steps to improve the output segmentation. The full workflow of the final model is shown in figure 6.9 and described in the following sections.

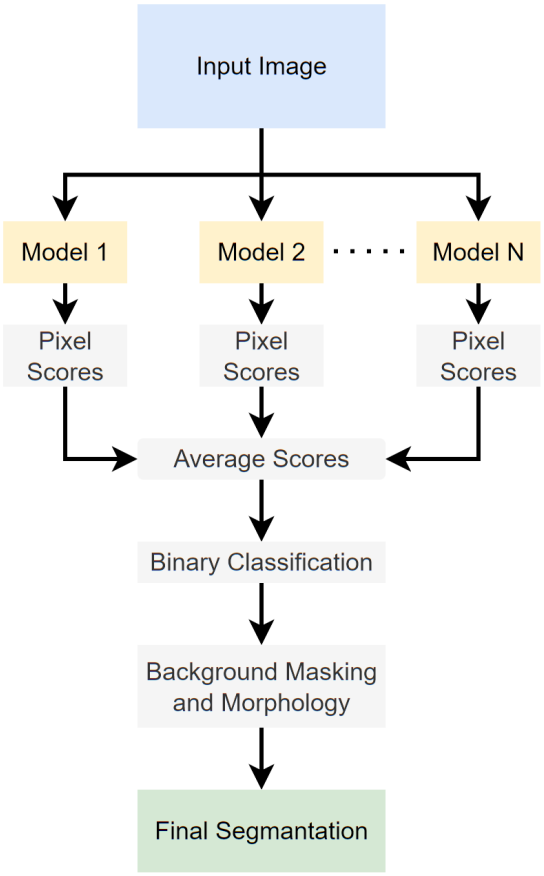


Figure 6.9: *The full ensemble model workflow.*

Several different network architectures were applied to the problem. Deeplabv3+ [77] was applied with both a ResNet50 [72] and MobileNetv2 [228] backbone as well as UNet [73] with an encoder depth of 4. All of these networks were applied with a binary pixel classification output with the pixel classifications as "tumour" or "background".

This means that both healthy lung tissue and the non-biological content (slicing table etc.) of the image were included in the background class. All of the network architectures were trained with both weighted cross entropy and DICE loss functions. All of the models used pre-trained weights from training on ImageNet [229]. As the ImageNet dataset contains no medical images the choice was taken to retrain all of the layers of the network. All of the trained networks are listed in Table 6.9

Model	**Network Architecture**	**Loss Function**
1	DeepLabV3+ ResNet50	BCE
2	DeepLabV3+ ResNet50	DICE
3	DeepLabV3+ MobileNetv2	BCE
4	DeepLabV3+ MobileNetv2	DICE
5	Unet (Encoder Depth: 4)	BCE
6	Unet (Encoder Depth: 4)	DICE

Table 6.9: *All networks trained for the gross pathology segmentation task.*

The dataset was split into training and test sets by patient so that images from one patient were only used for either training or testing. This is important as the different slices from the same patient contain similar features, such as the colour of the tumour and healthy tissue, that would bias the results if they were included in both training and test sets. Using this approach, the data was split into a training set containing 18 patients images and a test set containing 4 patients images. Using this approach causes the number of images in both the training and test sets to change depending on which patients images were used as there were more images available for some patients than others. This method generally created a split of around 96/20 images in the training/test sets. A 5-fold cross-validation approach was applied where all of the models were trained separately on different sets of 17 or 18 patients and tested on the 4 or 5 that were not included in the training set with the test patients changing every fold. This allows for the model to be tested on the full dataset. The training parameters are shown in table 6.10. All of the network training was performed on a two NVDIA GeForce GTX 1080 Ti graphics cards running in parallel.

In addition to applying the models individually, an ensemble-based approach was taken. To achieve this the individual pixel prediction probability outputs of each individual network in table 6.9 were simply averaged.

Parameter	Value
Optimisation Method	adam
Initial LR	0.001
LR Drop Rate	5
Drop Factor	0.2
Training Epochs	20

Table 6.10: *CNN training parameters.*

6.5.5 Image Post-Processing

The deep learning models often correctly segment the region of tumour in the input image but also labels some separate erroneous regions as tumour. These incorrect regions can usually be removed through some morphological operations that can be applied based on what is known about the task to improve the segmentation results. The morphology steps are detailed in the list below:

1. Small objects with a size of fewer than 5000 pixels are removed from the image (for reference, pixels are generally around 0.1x0.1mm).

2. A morphological closing operation is applied using a circular structuring element with a radius of 20 pixels.

3. Any holes in the remaining objects are filled.

4. The total number of pixels in each remaining object is calculated. Only the object consisting of the highest number of pixels is kept as the final tumour segmentation.

Step 1 is applied to remove small isolated regions that were classified as tumour as these are almost always incorrect classifications, this step also improves the performance of all of the following steps. Steps 2 and 3 in the list above are required because many NSCLC tumours contain necrotic cores. These regions are pathologically and visually different from non-necrotic areas of tumour which, combined with the fact that there are few different patient examples in the datasets, causes them to be often misclassified as non-tumour. Simply closing and filling the tumour region generally fixes this problem. Step 4 can be applied as it is known that the images in our dataset are from patients with one large NSCLC tumour.

6.6 Gross Pathology Segmentation: Results

6.6.1 Segmentation Metrics

The results of the 5-fold cross-validation are shown in tables 6.11 and 6.12 for datasets A and B respectively. For both datasets the ensemble model outperforms the individual models. Including the morphology steps improves the ensemble results across all of the metrics showing that for this particular application, they are worthwhile to include. The results on dataset-B are considerably lower than those in dataset-A. This is expected due to the lower quality of the images in dataset-B and highlights the importance of good pathology photography practice as is described in [172].

Model	Accuracy (%)	Sensitivity (%)	Tumour IoU	Background IoU
1	94.3	70.7	0.506	0.913
2	92.3	63.1	0.395	0.872
3	93.2	59.8	0.418	0.892
4	91.1	55.1	0.317	0.850
5	91.9	59.4	0.392	0.898
6	88.9	55.4	0.314	0.835
Ens	96.3	65.9	0.552	0.931
Ens + Morph	**97.3**	**70.8**	**0.632**	**0.949**

Table 6.11: *Results for Dataset-A from the 5-fold cross-validation testing scheme.*

Model	Accuracy (%)	Sensitivity (%)	Tumour IoU	Background IoU
1	82.2	69.8	0.398	0.754
2	85.6	52.1	0.357	0.776
3	87.3	67.3	0.521	0.799
4	87.6	62.1	0.448	0.798
5	83.6	62.2	0.345	0.753
6	82.3	59.8	0.371	0.762
Ens	89.7	68.9	0.503	0.837
Ens + Morph	**91.2**	**70.1**	**0.529**	**0.853**

Table 6.12: *Results for Dataset-B from the 5-fold cross-validation testing scheme.*

6.6.2 Segmentation Examples

There is a large variety in the quality of the segmentation output of the ensemble model depending on the input images, some examples of this are shown in figure 6.10. Figure

6.10 (a.i) shows an example of a correct segmentation result on a tumour with good contrast between the healthy tissue and tumour tissue. The tumour boundary in image 6.10 (a.ii) aligns very closely with the ground truth mask producing a tumour IoU of 0.956 for this image.

Figure 6.10 (b.i) shows an example of a partially correct segmentation where an area of necrosis has not been included in the segmentation. This example has an IoU of 0.439 for the tumour class. The segmentation contour in this image outlines the region of lighter tissue which corresponds to the living tumour area. The necrotic area is not included in the segmentation output but is part of the ground truth tumour area as seen in image 6.10 (b.ii). The model tends to misclassify necrotic regions as non-tumour as there are not many examples of heavily necrotic tumours in the training datasets and the coagulated blood that appears in this region also often appears in areas of healthy tissue. In other necrotic examples, this can be fixed by the post-processing morphology steps but in this case, as the living tumour area does not fully enclose the necrotic region, these steps do not solve this problem.

Figure 6.10 (c.i) shows an example of a failed segmentation with a tumour IoU of only 0.035. Upon analysing this image within the context of the dataset it is seen that the image is from one of only two patients in the datasets who had an adenocarcinoma tumour. Adenocarcinoma has a lepidic pattern of growth causing it to be less contrasted against healthy tissue in gross images than other types of NSCLC. All other images from this patient and the other patient with adenocarcinoma have a similarly failed segmentation. It is clear from this that the dataset would need to be expanded to include more adenocarcinoma examples.

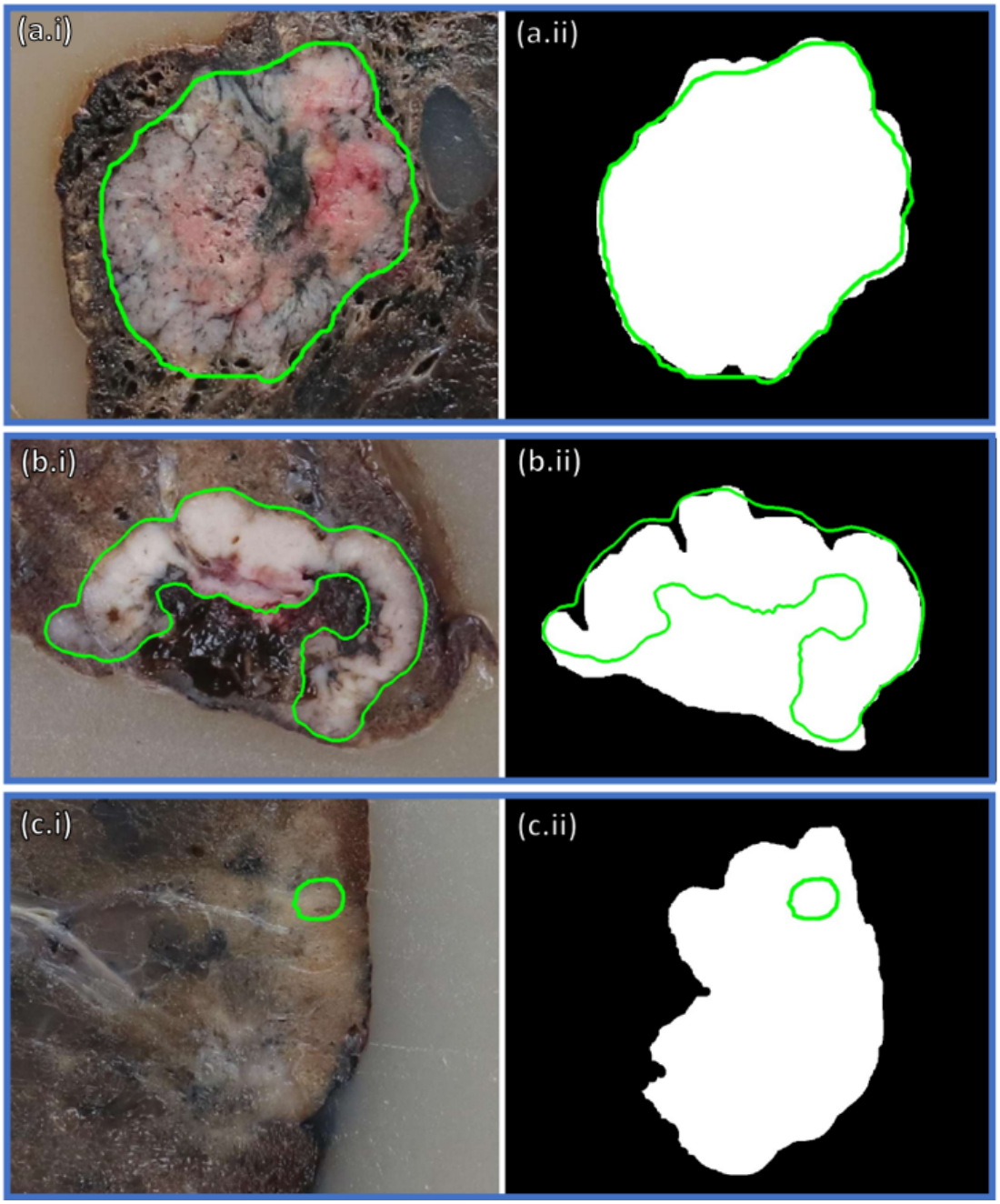


Figure 6.10: *Three segmentation examples from separate patients are displayed.* ***(a)*** *shows a good segmentation example,* ***(b)*** *shows a partially failed example due to a necrotic region and* ***(c)*** *shows a fully failed segmentation due to the tumour being an adenocarcinoma. Images denoted with* ***(i)*** *are the original test images zoomed in on the tumour area and images denoted with* ***(ii)*** *are the ground truth tumour segmentations. All images have the automatic segmentation contour overlayed (green line).*

6.7 Gross Pathology Segmentation: Discussion

The classification of the entire pathology of the lung into the two categories of tumour and non-tumour is an oversimplification that presents some problems for the segmentation model. This is most notable with adenocarcinoma tumours that are generally not recognised as tumours. Additionally, necrotic regions within the tumour are often misclassified as non-tumour regions. This can generally be fixed through the use of morphological image processing steps but it still highlights a problem with the ground truth data. This would be improved by increasing the dataset size to include more patients as the small dataset used in this study, with only 22 separate patients, included only a few examples of different pathological features such as adenocarcinomas and necrotic regions. A dataset containing a similar number of images that were all from unique patients would likely increase the performance of the trained models as this would allow the model to learn a more comprehensive array of pathological features. In the skin lesion photograph segmentation domain, large datasets such as the HAM10000 dataset [230], which contains 10000 images and ground truth segmentations of skin cancer lesions, allow for highly accurate models to be produced. In addition to increasing the dataset size, it may be beneficial to increase the number of classes used for the segmentation to include different types of tissue though this would require a time-intensive process of manual segmentation to produce the ground truth labels.

The final results for dataset-A produced better scoring metrics than those produced from dataset-B. This is unsurprising as, for the reasons described in section 6.5.1, the images in dataset-B are of poorer quality than those in dataset-A. This reduced image quality will increase the difficulty of segmentation first due to the image features being obscured and secondly due to there being fewer of these images of low quality in the overall training datasets. For further development and application of a system for the automatic segmentation of gross pathology photographs, care should be taken to ensure a high image quality by following the photography steps outlined in [172], though the inclusion of lower quality images in the training set may be beneficial to increase the robustness of the model.

For applications in clinical use, it may be beneficial to include some user input to produce a semi-automatic segmentation to decrease the chance of errors and improve overall accuracy. This could involve simply selecting the correct region from the output

of the model to remove some of the morphology steps or marking some tumour or background pixels to be input to the network. The decision to choose a fully or semi-automatic approach would depend on the specific application and pathology workflow that the model is to be included in.

6.8 Conclusions

This chapter has detailed work on the development of a system for the registration PET/CT to pathology images as well as producing a methodology for the automatic semantic segmentation of gross pathology photography. Conclusions for both of these tasks is detailed in the following subsections.

6.8.1 PET/CT to Pathology Image Registration

The previous work by Reines March et al. [2] established a methodology for the registration of CT and pathology images for NSCLC patients. This work was advanced in this chapter by the inclusion of the PET aspect of the PET/CT images. This included the use of respiratory motion reduction methods. The respiratory motion reduction methods evaluated were the Siemens OncoFreeze and time based gating algorithms. Both methods produced an improvement to the PET image quality based on the metrics evaluated with the gated method being the best fit for this project as the gate can be selected to best match the CT image. Additionally, registration methods based on SUV thresholding the PET image to obtain the PET tumour volume were produced for cases where there is a poor match between the pathology and the CT tumour volumes.

6.8.2 Gross Pathology Segmentation

Deep learning-based methods for semantic segmentation have been applied to the novel application of automatic segmentation of tumour areas in gross pathology photographs of specimens from patients with NSCLC. A pipeline for image pre-processing, model training and post-processing of the segmentation output has been detailed and validated. This work has demonstrated the possibility of achieving this goal as well as highlighting some challenges for producing a fully robust system. The main barrier to improving the performance is a lack of data. Increasing the size and diversity of the dataset would improve the model performance, especially for tumours with less

common features.

Chapter Acknowledgements

The work in this chapter was initially published as [231] and has been reproduced with permission from Springer Nature.

Chapter 7

Conclusions

7.1 Summary

The overall aim of this thesis was to apply image processing and machine learning techniques to lung cancer treatment with the aim of improving treatment outcomes. This was achieved in the more specific topics of radiotherapy induced pulmonary toxicity prediction (Chapters 4), radiotherapy induced esophageal toxicity prediction (Chapter 5), PET/CT to pathology image registration (Chapter 6) and gross pathology tumour segmentation (Chapter 6).

7.1.1 Radiotherapy induced pulmonary toxicity prediction

Chapter 4 covered the work on pulmonary toxicity prediction. Two separate datasets were used for this, one set, the Edinburgh Pneumonitis dataset, was for the prediction of pneumonitis on standard NSCLC patients receiving IMRT and the other dataset, the ASPIRE-ILD dataset, was for predicting pulmonary toxicity and other outcomes such as the FACT-L score for NSCLC patients with ILD receiving SABR treatment. Dose features, such as the V_{20}, were extracted from the lung region of the patient's dose images and available clinical features were converted to numeric representations. CT image features were extracted using radiomic approaches applied to the lung volume and through the use of a pre-trained UNet CNN model. It was shown on both datasets that both radiomic and deep learning based features extracted from CT images could provide a valuable source of information to greatly improve the accuracy of pulmonary toxicity prediction. For the ASPIRE-ILD dataset, it was additionally shown that different subsets of these features could provide information for the prediction of the FACT-L

score, EQ-5D-5L score, Cough Index and Overall Survival.

7.1.2 Radiotherapy induced esophageal toxicity prediction

Chapter 5 covered the work on the prediction of radiotherapy induced esophageal toxicity using data from the RTOG-0617 study which is currently the largest public dataset with the information required for this analysis and includes both IMRT and 3D-CRT data. A 4D-CNN was used for feature extraction and classification from dose maps. This was compared to more standard approaches of classifying dose volume histogram features using the LKB NTCP model and boosted decision trees. All models were trained as both binary classifiers and regression models. It was found that using the 4D-CNN, regression based training, combining the models as an ensemble and including all esophageal toxicities as opposed to just esophagitis provided marginal improvements to the performance which, when combined, produced a larger performance increase. Additionally, an investigation was carried out to study the robustness of boosted decision tree and ANN based models under the influence of random noise added to dose features. It was found that while often the choice of hyperparameter or data augmentation method has little effect when there is no feature noise, as noise increases, changing these variables has a large effect on the model performance.

7.1.3 PET/CT to pathology image registration

The first half of Chapter 6 details the work on advancing the methods for registering PET/CT with pathology images. This work followed on from the work by Reines March, et al. [2]. To advance this work, respiratory motion reduction techniques were included for the PET portion of the PET/CT scan. Comparisons between time based respiratory gating, the OncoFreeze algorithm and a standard reconstruction were made. It was observed that SUV thresholding the PET image with a set SUV_{max} produced an inconsistent match to the pathology volume. The registration algorithm was advanced by adding a PET to pathology registration which used the PET volume thresholded to match the pathology volume, in many cases this provided a better match to the pathology than the CT based method.

7.1.4 Gross pathology tumour segmentation

The second half of Chapter 6 details the work developing a deep learning based method for the automatic segmentation of tumour regions in gross pathology images. pre-processing included masking of background areas using either a colour based k-means clustering or a visual saliency segmentation. Versions of DeeplabV3+ and UNet were then trained with both balanced cross entropy and dice loss functions. The best performing model was an ensemble of all of the models. Post-processing using morphological methods to remove erroneous regions and include necrotic regions proved to further increase performance. The models showed good performance on tumour types that were well represented in the dataset but performed poorly on less represented examples such as those of adenocarcinoma and highly necrotic tumours.

7.2 Challenges for clinical implementation

For all of the methods developed in this thesis, there are several challenges for clinical implementation. These are discussed in the following subsections.

7.2.1 Radiotherapy Toxicity Prediction

For all predictive models intended for clinical application, a significant challenge lies in establishing appropriate accountability in the event of model errors. This is a larger issue if the models output was solely relied upon. The best balance can be struck by implementing predictive models to provide additional information to clinicians who would then make the final treatment decisions.

Another issue for the clinical implementation of these predictive models is maintaining reliability over the constantly evolving imaging and therapy landscape. Emerging radiotherapy treatment methods, such as FLASH, will alter patient responses to treatment. This means that models trained on data from older radiotherapy techniques, such as IMRT, may fail to accurately predict outcomes when applied to newer RT techniques. This presents a large challenge due to the large volume of data required to validate these techniques for clinical settings as more data would have to be collected to validate any machine learning techniques applied to new radiotherapy methods.

For the clinical implementation of the pulmonary toxicity prediction work and use of predictive CT features, it is likely to be beneficial to impose strict guidelines on CT

imaging and reconstruction to ensure that either radiomic or deep learning methods produce high quality image features that are similar to what the model was trained on. This is especially true for RT planning CT images which are generally lower quality than diagnostic CT images so more care would have to be taken here. Additionally, changes in imaging across different centres arising from different scanning protocols and the use of different scanners can also add to these problems.

7.2.2 Automatic Gross Pathology Tumour Segmentation

The methods for automatic segmentation of tumour areas in gross pathology share some of the problems as the radiotherapy predictive models. If the segmentation models were used to automate parts of the pathology reporting process there would be issues arising if the model failed to accurately segment the tumour region. When initially applied in the clinical setting it would be beneficial to apply the methods in parallel to the current clinical standard for pathology reporting to observe if any issues or benefits would arise when compared to the current standard practices. If the models were used in real time to highlight to clinicians regions of anatomy to sample for further imaging then this again could be applied as an assistive tool as opposed to defining the final regions to sample.

7.2.3 PET/CT to Pathology Image Registration

The PET/CT to pathology registration algorithm is not intended to be directly applied to make any patient specific treatment decisions meaning there is less to discuss regarding clinical implementation. The system is intended to be used to validate PET imaging for the purpose of informing clinicians regarding what can be determined from these images. To get to this stage, further patient data would be required to be collected to both further validate the registration system and provide the data required to validate the PET imaging in relation to the histopathology.

7.3 Future Work

The largest area of future work for the topics covered by this thesis as well as the wider field of image processing and machine learning applied to medical imaging is the validation of models on large diverse datasets. The challenge here is based on data

collection and dissemination where the data collection process for medical imaging is time consuming and expensive and ethical constraints limit the open dissemination of data. The radiotherapy toxicity prediction and gross pathology segmentation would benefit greatly from a large, open, benchmark dataset allowing for the direct comparison of techniques. The PET/CT to pathology registration algorithm could be applied to investigate regions of inflammation by applying specific stains to the histopathology slides to highlight these regions. Additionally, other PET tracers could be investigated and validated using the registration workflow. Tracers such as 18F-FMISO, which highlights necrotic regions of a tumour, would be a good candidate for this histopathology based validation. One of the main challenges for all methods in this thesis is the reliability and implementation within the clinical workflow. This should be tackled by using a human-in-the-loop approach where the output of the methods are only used to inform a clinician who would make a final treatment decision. The level of trust in a specific method would have to be backed up by testing in prospective clinical settings.

Appendices

Appendix A

ASPIRE-ILD Feature Importance

During the training of the models for the prediction of the ASPIRE -ILD outcomes, the feature importance was calculated for the purpose of reducing number of features input to the final model. This process is described in section 4.2.5. This feature importance estimation produces a numerical metric indicating how strongly each feature effects the final prediction. It should be noted that although this feature importance value correlates with the weighting of the features in the boosted decision trees, this is not an exact measure of importance. Additionally, many dose, radiomic and UNet features will be highly correlated with different features from the same feature class. This means that, for many features, a low importance score may only be due to the predictive power of that feature already being included in the model from another, or a combination of other, features. A univariate method for feature importance estimation would be better for isolated feature importance. The ten features used in each outcome prediction model for the best performing feature subsets are displayed in tables A.1 to A.6. The feature importance here has been normalised so that the most important feature has an importance value of 1.

Feature	Importance
Dose - Vol55	1.000
PyRadiomic - original_glszm_SizeZoneNonUniformity	0.153
PyRadiomic - original_glcm_Correlation	0.134
UNet49	0.108
PyRadiomic - original_glcm_Imc2	0.026
PyRadiomic - original_gldm_DependenceNonUniformity	0.015
UNet18	0.013
PyRadiomic - original_firstorder_Energy	0.009
UNet19	0.006
UNet4	0.003

Table A.1: *CTCAE pulmonary toxicity feature importance for the LSBoost model with only Dose, Pyradiomic and UNet features.*

Feature	Importance
PyRadiomic - original_glcm_ClusterShade	1.000
UNet40	0.335
PyRadiomic - original_glszm_SZNonUniformityNormalized	0.276
UNet41	0.195
PyRadiomic - original_glcm_Id	0.118
PyRadiomic - original_glszm_LAHighGrayLevelEmphasis	0.095
PyRadiomic - original_firstorder_90Percentile	0.078
PyRadiomic - original_glszm_ZoneEntropy	0.073
UNet45	0.049
PyRadiomic - original_gldm_SDLowGrayLevelEmphasis	0.048

Table A.2: *FACT-L feature importance for the LSBoost model with only Dose, Pyradiomic and UNet features.*

Feature	Importance
PyRadiomic - original_glszm_GLNonUniformityNormalized	1.000
Clinical - L_packyrs	0.535
PyRadiomic - original_firstorder_Energy	0.432
Dose - MaxLD	0.129
PyRadiomic - original_firstorder_Minimum	0.113
PyRadiomic - original_firstorder_90Percentile	0.111
PyRadiomic - original_firstorder_RootMeanSquared	0.090
PyRadiomic - original_glcm_ClusterTendency	0.088
PyRadiomic - original_glcm_Idn	0.082
PyRadiomic - original_firstorder_Mean	0.077

Table A.3: *EQ-5D-5L feature importance for the LSBoost model with only Dose, clinial and Pyradiomic features.*

Feature	Importance
Clinical - dlco_base	1.000
Clinical - fvc_cr1	0.421
Clinical - L_packyrs	0.211
Clinical - t_stage	0.085
Clinical - CurrentSmoker	0.040
Clinical - dose	0.013
Clinical - consdiag	0.012
Clinical - gender	0.011
Clinical - radpattern	0.009
Clinical - nsclc	0.001

Table A.4: *Overall survival feature importance for the LSBoost model with only clinical features.*

Feature	Importance
Clinical - radpattern	1.000
Clinical - fvc_cr1	0.476
Clinical - dlco_base	0.463
Clinical - dose	0.371
Clinical - nsclc	0.320
Clinical - L_packyrs	0.056
Clinical - consdiag	0.049
Clinical - t_stage	0.029
Clinical - CurrentSmoker	0.019
Clinical - ild_subtype	0.015

Table A.5: *Cough index prediction feature importance for the LSBoost model with clinical features only.*

Feature	Importance
PyRadiomic - original_firstorder_10Percentile	1.000
UNet10	0.749
PyRadiomic - original_glcm_DifferenceVariance	0.481
UNet6	0.225
PyRadiomic - original_glrlm_RunLengthNonUniformity	0.211
Dose - V5	0.176
Dose - V60	0.121
UNet16	0.120
UNet30	0.119
PyRadiomic - original_glcm_ClusterProminence	0.101

Table A.6: *FACT-L B1 dyspnea question prediction feature importance for the LSBoost model with Dose, Pyradiomic and UNet features.*

Appendix B

ResNet50 Architecture

ResNet50 Architecture

	Name	Type	Activations	Learnable Properties
1	data 80×80×100×1 images with 'rescale-symmetric' normalization	3-D Image Input	80(S) × 80(S) × 100(S) × 1(C) × 1(B)	-
2	conv1 32 7×7×7 convolutions with stride [2 2 2] and padding [3 3 3; 3 3 3]	3-D Convolution	40(S) × 40(S) × 50(S) × 32(C) × 1(B)	Weights 7 × 7 × 7 × 1 × 32 Bias 1 × 1 × 1 × 32
3	bn_conv1 Batch normalization	Batch Normalization	40(S) × 40(S) × 50(S) × 32(C) × 1(B)	Offset 1 × 1 × 1 × 32 Scale 1 × 1 × 1 × 32
4	conv1_relu ReLU	ReLU	40(S) × 40(S) × 50(S) × 32(C) × 1(B)	-
5	pool1 3×3×3 max pooling with stride [2 2 2] and padding [1 1 1; 1 1 1]	3-D Max Pooling	20(S) × 20(S) × 25(S) × 32(C) × 1(B)	-
6	res2a_branch2a 32 3×3×3 convolutions with stride [1 1 1] and padding [1 1 1; 1 1 1]	3-D Convolution	20(S) × 20(S) × 25(S) × 32(C) × 1(B)	Weights 3 × 3 × 3 × 32 × 32 Bias 1 × 1 × 1 × 32
7	bn2a_branch2a Batch normalization	Batch Normalization	20(S) × 20(S) × 25(S) × 32(C) × 1(B)	Offset 1 × 1 × 1 × 32 Scale 1 × 1 × 1 × 32
8	res2a_branch2a_relu ReLU	ReLU	20(S) × 20(S) × 25(S) × 32(C) × 1(B)	-
9	res2a_branch2b 32 3×3×3 convolutions with stride [1 1 1] and padding [1 1 1; 1 1 1]	3-D Convolution	20(S) × 20(S) × 25(S) × 32(C) × 1(B)	Weights 3 × 3 × 3 × 32 × 32 Bias 1 × 1 × 1 × 32
10	bn2a_branch2b Batch normalization	Batch Normalization	20(S) × 20(S) × 25(S) × 32(C) × 1(B)	Offset 1 × 1 × 1 × 32 Scale 1 × 1 × 1 × 32
11	res2a Element-wise addition of 2 inputs	Addition	20(S) × 20(S) × 25(S) × 32(C) × 1(B)	-
12	res2a_relu ReLU	ReLU	20(S) × 20(S) × 25(S) × 32(C) × 1(B)	-
13	res2b_branch2a 32 3×3×3 convolutions with stride [1 1 1] and padding [1 1 1; 1 1 1]	3-D Convolution	20(S) × 20(S) × 25(S) × 32(C) × 1(B)	Weights 3 × 3 × 3 × 32 × 32 Bias 1 × 1 × 1 × 32
14	bn2b_branch2a Batch normalization	Batch Normalization	20(S) × 20(S) × 25(S) × 32(C) × 1(B)	Offset 1 × 1 × 1 × 32 Scale 1 × 1 × 1 × 32
15	res2b_branch2a_relu ReLU	ReLU	20(S) × 20(S) × 25(S) × 32(C) × 1(B)	-
16	res2b_branch2b 32 3×3×3 convolutions with stride [1 1 1] and padding [1 1 1; 1 1 1]	3-D Convolution	20(S) × 20(S) × 25(S) × 32(C) × 1(B)	Weights 3 × 3 × 3 × 32 × 32 Bias 1 × 1 × 1 × 32
17	bn2b_branch2b Batch normalization	Batch Normalization	20(S) × 20(S) × 25(S) × 32(C) × 1(B)	Offset 1 × 1 × 1 × 32 Scale 1 × 1 × 1 × 32
18	res2b Element-wise addition of 2 inputs	Addition	20(S) × 20(S) × 25(S) × 32(C) × 1(B)	-
19	res2b_relu ReLU	ReLU	20(S) × 20(S) × 25(S) × 32(C) × 1(B)	-
20	res3a_branch1 64 1×1×1 convolutions with stride [2 2 2] and padding [0 0 0; 0 0 0]	3-D Convolution	10(S) × 10(S) × 13(S) × 64(C) × 1(B)	Weights 1 × 1 × 1 × 32 × 64 Bias 1 × 1 × 1 × 64
21	bn3a_branch1 Batch normalization	Batch Normalization	10(S) × 10(S) × 13(S) × 64(C) × 1(B)	Offset 1 × 1 × 1 × 64 Scale 1 × 1 × 1 × 64
22	res3a_branch2a 64 3×3×3 convolutions with stride [2 2 2] and padding [1 1 1; 1 1 1]	3-D Convolution	10(S) × 10(S) × 13(S) × 64(C) × 1(B)	Weights 3 × 3 × 3 × 32 × 64 Bias 1 × 1 × 1 × 64
23	bn3a_branch2a Batch normalization	Batch Normalization	10(S) × 10(S) × 13(S) × 64(C) × 1(B)	Offset 1 × 1 × 1 × 64 Scale 1 × 1 × 1 × 64
24	res3a_branch2a_relu ReLU	ReLU	10(S) × 10(S) × 13(S) × 64(C) × 1(B)	-
25	res3a_branch2b 64 3×3×3 convolutions with stride [1 1 1] and padding [1 1 1; 1 1 1]	3-D Convolution	10(S) × 10(S) × 13(S) × 64(C) × 1(B)	Weights 3 × 3 × 3 × 64 × 64 Bias 1 × 1 × 1 × 64
26	bn3a_branch2b Batch normalization	Batch Normalization	10(S) × 10(S) × 13(S) × 64(C) × 1(B)	Offset 1 × 1 × 1 × 64 Scale 1 × 1 × 1 × 64
27	res3a Element-wise addition of 2 inputs	Addition	10(S) × 10(S) × 13(S) × 64(C) × 1(B)	-
28	res3a_relu ReLU	ReLU	10(S) × 10(S) × 13(S) × 64(C) × 1(B)	-
29	res3b_branch2a 64 3×3×3 convolutions with stride [1 1 1] and padding [1 1 1; 1 1 1]	3-D Convolution	10(S) × 10(S) × 13(S) × 64(C) × 1(B)	Weights 3 × 3 × 3 × 64 × 64 Bias 1 × 1 × 1 × 64
30	bn3b_branch2a Batch normalization	Batch Normalization	10(S) × 10(S) × 13(S) × 64(C) × 1(B)	Offset 1 × 1 × 1 × 64 Scale 1 × 1 × 1 × 64
31	res3b_branch2a_relu ReLU	ReLU	10(S) × 10(S) × 13(S) × 64(C) × 1(B)	-
32	res3b_branch2b 64 3×3×3 convolutions with stride [1 1 1] and padding [1 1 1; 1 1 1]	3-D Convolution	10(S) × 10(S) × 13(S) × 64(C) × 1(B)	Weights 3 × 3 × 3 × 64 × 64 Bias 1 × 1 × 1 × 64
33	bn3b_branch2b Batch normalization	Batch Normalization	10(S) × 10(S) × 13(S) × 64(C) × 1(B)	Offset 1 × 1 × 1 × 64 Scale 1 × 1 × 1 × 64
34	res3b Element-wise addition of 2 inputs	Addition	10(S) × 10(S) × 13(S) × 64(C) × 1(B)	-
35	res3b_relu ReLU	ReLU	10(S) × 10(S) × 13(S) × 64(C) × 1(B)	-
36	res4a_branch1 128 1×1×1 convolutions with stride [2 2 2] and padding [0 0 0; 0 0 0]	3-D Convolution	5(S) × 5(S) × 7(S) × 128(C) × 1(B)	Weights 1 × 1 × 1 × 64 × 128 Bias 1 × 1 × 1 × 128
37	bn4a_branch1 Batch normalization	Batch Normalization	5(S) × 5(S) × 7(S) × 128(C) × 1(B)	Offset 1 × 1 × 1 × 128 Scale 1 × 1 × 1 × 128
38	res4a_branch2a 128 3×3×3 convolutions with stride [2 2 2] and padding [1 1 1; 1 1 1]	3-D Convolution	5(S) × 5(S) × 7(S) × 128(C) × 1(B)	Weights 3 × 3 × 3 × 64 × 128 Bias 1 × 1 × 1 × 128
39	bn4a_branch2a Batch normalization	Batch Normalization	5(S) × 5(S) × 7(S) × 128(C) × 1(B)	Offset 1 × 1 × 1 × 128 Scale 1 × 1 × 1 × 128
40	res4a_branch2a_relu ReLU	ReLU	5(S) × 5(S) × 7(S) × 128(C) × 1(B)	-

41	res4a_branch2b 128 3×3×3 convolutions with stride [1 1 1] and padding [1 1 1; 1 1 1]	3-D Convolution	5(S) × 5(S) × 7(S) × 128(C) × 1(B)	Weights 3 × 3 × 3 × 128 × 128 Bias 1 × 1 × 1 × 128
42	bn4a_branch2b Batch normalization	Batch Normalization	5(S) × 5(S) × 7(S) × 128(C) × 1(B)	Offset 1 × 1 × 1 × 128 Scale 1 × 1 × 1 × 128
43	res4a Element-wise addition of 2 inputs	Addition	5(S) × 5(S) × 7(S) × 128(C) × 1(B)	-
44	res4a_relu ReLU	ReLU	5(S) × 5(S) × 7(S) × 128(C) × 1(B)	-
45	res4b_branch2a 128 3×3×3 convolutions with stride [1 1 1] and padding [1 1 1; 1 1 1]	3-D Convolution	5(S) × 5(S) × 7(S) × 128(C) × 1(B)	Weights 3 × 3 × 3 × 128 × 128 Bias 1 × 1 × 1 × 128
46	bn4b_branch2a Batch normalization	Batch Normalization	5(S) × 5(S) × 7(S) × 128(C) × 1(B)	Offset 1 × 1 × 1 × 128 Scale 1 × 1 × 1 × 128
47	res4b_branch2a_relu ReLU	ReLU	5(S) × 5(S) × 7(S) × 128(C) × 1(B)	-
48	res4b_branch2b 128 3×3×3 convolutions with stride [1 1 1] and padding [1 1 1; 1 1 1]	3-D Convolution	5(S) × 5(S) × 7(S) × 128(C) × 1(B)	Weights 3 × 3 × 3 × 128 × 128 Bias 1 × 1 × 1 × 128
49	bn4b_branch2b Batch normalization	Batch Normalization	5(S) × 5(S) × 7(S) × 128(C) × 1(B)	Offset 1 × 1 × 1 × 128 Scale 1 × 1 × 1 × 128
50	res4b Element-wise addition of 2 inputs	Addition	5(S) × 5(S) × 7(S) × 128(C) × 1(B)	-
51	res4b_relu ReLU	ReLU	5(S) × 5(S) × 7(S) × 128(C) × 1(B)	-
52	res5a_branch1 256 1×1×1 convolutions with stride [2 2 2] and padding [0 0 0; 0 0 0]	3-D Convolution	3(S) × 3(S) × 4(S) × 256(C) × 1(B)	Weights 1 × 1 × 1 × 128 × 256 Bias 1 × 1 × 1 × 256
53	bn5a_branch1 Batch normalization	Batch Normalization	3(S) × 3(S) × 4(S) × 256(C) × 1(B)	Offset 1 × 1 × 1 × 256 Scale 1 × 1 × 1 × 256
54	res5a_branch2a 256 3×3×3 convolutions with stride [2 2 2] and padding [1 1 1; 1 1 1]	3-D Convolution	3(S) × 3(S) × 4(S) × 256(C) × 1(B)	Weights 3 × 3 × 3 × 128 × 256 Bias 1 × 1 × 1 × 256
55	bn5a_branch2a Batch normalization	Batch Normalization	3(S) × 3(S) × 4(S) × 256(C) × 1(B)	Offset 1 × 1 × 1 × 256 Scale 1 × 1 × 1 × 256
56	res5a_branch2a_relu ReLU	ReLU	3(S) × 3(S) × 4(S) × 256(C) × 1(B)	-
57	res5a_branch2b 256 3×3×3 convolutions with stride [1 1 1] and padding [1 1 1; 1 1 1]	3-D Convolution	3(S) × 3(S) × 4(S) × 256(C) × 1(B)	Weights 3 × 3 × 3 × 256 × 256 Bias 1 × 1 × 1 × 256
58	bn5a_branch2b Batch normalization	Batch Normalization	3(S) × 3(S) × 4(S) × 256(C) × 1(B)	Offset 1 × 1 × 1 × 256 Scale 1 × 1 × 1 × 256
59	res5a Element-wise addition of 2 inputs	Addition	3(S) × 3(S) × 4(S) × 256(C) × 1(B)	-
60	res5a_relu ReLU	ReLU	3(S) × 3(S) × 4(S) × 256(C) × 1(B)	-
61	res5b_branch2a 256 3×3×3 convolutions with stride [1 1 1] and padding [1 1 1; 1 1 1]	3-D Convolution	3(S) × 3(S) × 4(S) × 256(C) × 1(B)	Weights 3 × 3 × 3 × 256 × 256 Bias 1 × 1 × 1 × 256
62	bn5b_branch2a Batch normalization	Batch Normalization	3(S) × 3(S) × 4(S) × 256(C) × 1(B)	Offset 1 × 1 × 1 × 256 Scale 1 × 1 × 1 × 256
63	res5b_branch2a_relu ReLU	ReLU	3(S) × 3(S) × 4(S) × 256(C) × 1(B)	-
64	res5b_branch2b 256 3×3×3 convolutions with stride [1 1 1] and padding [1 1 1; 1 1 1]	3-D Convolution	3(S) × 3(S) × 4(S) × 256(C) × 1(B)	Weights 3 × 3 × 3 × 256 × 256 Bias 1 × 1 × 1 × 256
65	bn5b_branch2b Batch normalization	Batch Normalization	3(S) × 3(S) × 4(S) × 256(C) × 1(B)	Offset 1 × 1 × 1 × 256 Scale 1 × 1 × 1 × 256
66	res5b Element-wise addition of 2 inputs	Addition	3(S) × 3(S) × 4(S) × 256(C) × 1(B)	-
67	res5b_relu ReLU	ReLU	3(S) × 3(S) × 4(S) × 256(C) × 1(B)	-
68	pool5 3-D global average pooling	3-D Global Average...	1(S) × 1(S) × 1(S) × 256(C) × 1(B)	-
69	dropout 25% dropout	Dropout	1(S) × 1(S) × 1(S) × 256(C) × 1(B)	-
70	fc1000 2 fully connected layer	Fully Connected	1(S) × 1(S) × 1(S) × 2(C) × 1(B)	Weights 2 × 256 Bias 2 × 1
71	prob softmax	Softmax	1(S) × 1(S) × 1(S) × 2(C) × 1(B)	-
72	ClassificationLayer_fc1000 crossentropyex	Classification Output	1(S) × 1(S) × 1(S) × 2(C) × 1(B)	-

Bibliography

[1] *Cancer Research UK*. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer. Accessed: 13-06-2024.

[2] G. Reines March. "Registration of pre-operative lung cancer PET/CT scans with post-operative histopathology images". PhD thesis. University of Strathclyde, 2020.

[3] S. Ali et al. "Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy". In: *Medical image analysis* 70 (2021), p. 102002.

[4] J. D. Hipp et al. "Spatially Invariant Vector Quantization: A pattern matching algorithm for multiple classes of image subject matter including pathology". In: *Journal of pathology informatics* 2.1 (2011), p. 13.

[5] *OpenStax College Anatomy and Physiology ii*. https://courses.lumenlearning.com/suny-ap2/chapter/the-lungs/. Accessed: 2023-08-23.

[6] K. E. Pinkerton et al. "Architecture and cellular composition of the air–blood tissue barrier". In: *Comparative biology of the normal lung*. Elsevier, 2015, pp. 105–117.

[7] *Cancer Research UK Lung Cancer*. https://www.cancerresearchuk.org/about-cancer/lung-cancer. Accessed: 2023-08-14.

[8] A. Depeursinge et al. "Building a reference multimedia database for interstitial lung diseases". In: *Computerized medical imaging and graphics* 36.3 (2012), pp. 227–238.

[9] J. Thariat et al. "Past, present, and future of radiotherapy for the benefit of patients". In: *Nature reviews Clinical oncology* 10.1 (2013), pp. 52–60.

[10] J. Winiecki. “Principles of radiation therapy”. In: *Physical Sciences Reviews* 7.12 (2020), pp. 1501–1528.

[11] N. Ghaderi et al. “A century of fractionated radiotherapy: How mathematical oncology can break the rules”. In: *International Journal of Molecular Sciences* 23.3 (2022), p. 1316.

[12] M. Teoh et al. “Volumetric modulated arc therapy: a review of current literature and clinical use in practice”. In: *The British journal of radiology* 84.1007 (2011), pp. 967–996.

[13] S. Senthi et al. “Outcomes of stereotactic ablative radiotherapy for central lung tumours: a systematic review”. In: *Radiotherapy and Oncology* 106.3 (2013), pp. 276–282.

[14] D. S. Chang et al. *Basic radiotherapy physics and biology*. Tech. rep. Springer, 2014.

[15] A. P. Chen et al. “Grading dermatologic adverse events of cancer treatments: the Common Terminology Criteria for Adverse Events Version 4.0”. In: *Journal of the American Academy of Dermatology* 67.5 (2012), pp. 1025–1039.

[16] D. Savarese. “Common terminology criteria for adverse events”. In: *UpToDate Waltham, MA: UpToDate* (2013), pp. 1–9.

[17] T. J. Bledsoe, S. K. Nath, and R. H. Decker. “Radiation pneumonitis”. In: *Clinics in chest medicine* 38.2 (2017), pp. 201–208.

[18] T. I. Lingos et al. “Radiation pneumonitis in breast cancer patients treated with conservative surgery and radiation therapy”. In: *International Journal of Radiation Oncology* Biology* Physics* 21.2 (1991), pp. 355–360.

[19] D. Murro and S. Jakate. “Radiation esophagitis”. In: *Archives of Pathology and Laboratory Medicine* 139.6 (2015), pp. 827–830.

[20] A. Barrett et al. *Practical radiotherapy planning*. CRC Press, 2023.

[21] H. Hoy, T. Lynch, and M. Beck. “Surgical treatment of lung cancer”. In: *Critical Care Nursing Clinics* 31.3 (2019), pp. 303–313.

[22] M. Herdman et al. “Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L)”. In: *Quality of life research* 20 (2011), pp. 1727–1736.

[23] D. F. Cella et al. “The Functional Assessment of Cancer Therapy scale: development and validation of the general measure”. In: *J Clin Oncol* 11.3 (1993), pp. 570–579.

[24] M. J. Willemink and P. B. Noël. “The evolution of image reconstruction for CT—from filtered back projection to artificial intelligence”. In: *European radiology* 29 (2019), pp. 2185–2195.

[25] A. Kalra. “Developing fe human models from medical images”. In: *Basic finite element method as applied to injury biomechanics*. Elsevier, 2018, pp. 389–415.

[26] K. Lameka, M. D. Farwell, and M. Ichise. “Positron emission tomography”. In: *Handbook of clinical neurology* 135 (2016), pp. 209–227.

[27] P. E. Kinahan and J. W. Fletcher. “Positron emission tomography-computed tomography standardized uptake values in clinical practice and assessing response to therapy”. In: *Seminars in Ultrasound, CT and MRI*. Vol. 31. 6. Elsevier. 2010, pp. 496–505.

[28] V. Kapoor, B. M. McCook, and F. S. Torok. “An introduction to PET-CT imaging”. In: *Radiographics* 24.2 (2004), pp. 523–543.

[29] A. Pepin et al. “Management of respiratory motion in PET/computed tomography: the state of the art”. In: *Nuclear medicine communications* 35.2 (2014), p. 113.

[30] C. Ozhasoglu and M. J. Murphy. “Issues in respiratory motion compensation during external-beam radiotherapy”. In: *International Journal of Radiation Oncology* Biology* Physics* 52.5 (2002), pp. 1389–1399.

[31] P. J. Schleyer et al. “Retrospective data-driven respiratory gating for PET/CT”. In: *Physics in Medicine & Biology* 54.7 (2009), p. 1935.

[32] P. Schleyer et al. “Data-driven respiratory gating whole body PET using continuous bed motion”. In: *2018 IEEE Nuclear Science Symposium and Medical Imaging Conference Proceedings (NSS/MIC)*. IEEE. 2018, pp. 1–5.

[33] F. Büther et al. “Clinical Evaluation of a Data-Driven Respiratory Gating Algorithm for Whole-Body PET with Continuous Bed Motion”. In: *Journal of Nuclear Medicine* 61.10 (2020), pp. 1520–1527.

[34] M. Dawood et al. “Optimal number of respiratory gates in positron emission tomography: a cardiac patient study”. In: *Medical physics* 36.5 (2009), pp. 1775–1784.

[35] M. Dawood et al. “Respiratory gating in positron emission tomography: a quantitative comparison of different gating schemes”. In: *Medical physics* 34.7 (2007), pp. 3067–3076.

[36] J. Daouk et al. “Respiratory-gated positron emission tomography and breath-hold computed tomography coupling to reduce the influence of respiratory motion: methodology and feasibility”. In: *Acta Radiologica* 50.2 (2009), pp. 144–155.

[37] V. Bettinardi, E. Rapisarda, and M. Gilardi. “Number of partitions (gates) needed to obtain motion-free images in a respiratory gated 4D-PET/CT study as a function of the lesion size and motion displacement”. In: *Medical physics* 36.12 (2009), pp. 5547–5558.

[38] M. Mustra, K. Delac, and M. Grgic. “Overview of the DICOM standard”. In: *2008 50th International Symposium ELMAR*. Vol. 1. IEEE. 2008, pp. 39–44.

[39] G. Kumar and P. K. Bhatia. “A detailed review of feature extraction in image processing systems”. In: *2014 Fourth international conference on advanced computing & communication technologies*. IEEE. 2014, pp. 5–12.

[40] W.-C. Siu and K.-W. Hung. “Review of image interpolation and super-resolution”. In: *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE. 2012, pp. 1–10.

[41] D. Abdullah et al. “Application of interpolation image by using bi-cubic algorithm”. In: *Journal of Physics: Conference Series*. Vol. 1114. IOP Publishing. 2018, p. 012066.

[42] L. Najman and H. Talbot. *Mathematical morphology: from theory to applications*. John Wiley & Sons, 2013.

[43] H. Matsumoto, M. Ohtani, and I. Washitani. “Tree crown size estimated using image processing: A biodiversity index for sloping subtropical broad-leaved forests”. In: *Tropical Conservation Science* 10 (2017), p. 1940082917721787.

[44] Y. E. Erdi et al. “Segmentation of lung lesion volume by adaptive positron emission tomography image thresholding”. In: *Cancer: Interdisciplinary International Journal of the American Cancer Society* 80.S12 (1997), pp. 2505–2509.

[45] P. Roy et al. “Adaptive thresholding: A comparative study”. In: *2014 International conference on control, Instrumentation, communication and Computational Technologies (ICCICCT)*. IEEE. 2014, pp. 1182–1186.

[46] N. Dhanachandra, K. Manglem, and Y. J. Chanu. “Image segmentation using K-means clustering algorithm and subtractive clustering algorithm”. In: *Procedia Computer Science* 54 (2015), pp. 764–771.

[47] J. C. Bezdek, R. Ehrlich, and W. Full. “FCM: The fuzzy c-means clustering algorithm”. In: *Computers & geosciences* 10.2-3 (1984), pp. 191–203.

[48] D. Arthur and S. Vassilvitskii. “K-means++ the advantages of careful seeding”. In: *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. 2007, pp. 1027–1035.

[49] J. J. Van Griethuysen et al. “Computational radiomics system to decode the radiographic phenotype”. In: *Cancer research* 77.21 (2017), e104–e107.

[50] C. Nioche et al. “LIFEx: a freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity”. In: *Cancer research* 78.16 (2018), pp. 4786–4789.

[51] *MATLAB version ’23.2.0.2515942 (R2023b)*. The Mathworks, Inc. Natick, Massachusetts, 2021.

[52] W. H. Nailon. “Texture analysis methods for medical image characterisation”. In: *Biomedical imaging* 75 (2010), p. 100.

[53] C. D. Cantrell. *Modern mathematical methods for physicists and engineers*. Cambridge University Press, 2000.

[54] A. Zwanenburg et al. “The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping”. In: *Radiology* 295.2 (2020), pp. 328–338.

[55] B. A. Altazi et al. “Reproducibility of F18-FDG PET radiomic features for different cervical tumor segmentation methods, gray-level discretization, and reconstruction algorithms”. In: *Journal of applied clinical medical physics* 18.6 (2017), pp. 32–48.

[56] G. Srinivasan and G. Shobha. “Statistical texture analysis”. In: *Proceedings of world academy of science, engineering and technology*. Vol. 36. December. 2008, pp. 1264–1269.

[57] R. M. Haralick, K. Shanmugam, and I. H. Dinstein. “Textural features for image classification”. In: *IEEE Transactions on systems, man, and cybernetics* 6 (1973), pp. 610–621.

[58] M. M. Galloway. “Texture analysis using gray level run lengths”. In: *Computer graphics and image processing* 4.2 (1975), pp. 172–179.

[59] A. Krizhevsky, I. Sutskever, and G. E. Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems* 25 (2012).

[60] O. Russakovsky et al. “Imagenet large scale visual recognition challenge”. In: *International journal of computer vision* 115 (2015), pp. 211–252.

[61] M. Elgendy. *Deep learning for vision systems*. Simon and Schuster, 2020.

[62] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. `http://www.deeplearningbook.org`. MIT Press, 2016.

[63] A. F. Agarap. “Deep learning using rectified linear units (relu)”. In: *arXiv preprint arXiv:1803.08375* (2018).

[64] N. Srivastava et al. “Dropout: a simple way to prevent neural networks from overfitting”. In: *The journal of machine learning research* 15.1 (2014), pp. 1929–1958.

[65] L. R. Dice. “Measures of the amount of ecologic association between species”. In: *Ecology* 26.3 (1945), pp. 297–302.

[66] J. Terven et al. “Loss Functions and Metrics in Deep Learning. A Review”. In: *arXiv preprint arXiv:2307.02694* (2023).

[67] S. Ruder. “An overview of gradient descent optimization algorithms”. In: *arXiv preprint arXiv:1609.04747* (2016).

[68] D. P. Kingma and J. Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).

[69] P. Chlap et al. “A review of medical image data augmentation techniques for deep learning applications”. In: *Journal of Medical Imaging and Radiation Oncology* 65.5 (2021), pp. 545–563.

[70] M. Xu et al. “A comprehensive survey of image augmentation techniques for deep learning”. In: *Pattern Recognition* 137 (2023), p. 109347.

[71] B. Bischl et al. “Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 13.2 (2023), e1484.

[72] K. He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[73] O. Ronneberger, P. Fischer, and T. Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer. 2015, pp. 234–241.

[74] L.-C. Chen et al. “Semantic image segmentation with deep convolutional nets and fully connected crfs”. In: *arXiv preprint arXiv:1412.7062* (2014).

[75] L.-C. Chen et al. “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs”. In: *IEEE transactions on pattern analysis and machine intelligence* 40.4 (2017), pp. 834–848.

[76] L.-C. Chen et al. “Rethinking atrous convolution for semantic image segmentation”. In: *arXiv preprint arXiv:1706.05587* (2017).

[77] L.-C. Chen et al. “Encoder-decoder with atrous separable convolution for semantic image segmentation”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 801–818.

[78] L. Breiman. *Classification and regression trees*. Routledge, 2017.

[79] J. R. Quinlan. “Induction of decision trees”. In: *Machine learning* 1 (1986), pp. 81–106.

[80] G. V. Kass. “An exploratory technique for investigating large quantities of categorical data”. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 29.2 (1980), pp. 119–127.

[81] C. Bentéjac, A. Csörgő, and G. Martınez-Muñoz. “A comparative analysis of gradient boosting algorithms”. In: *Artificial Intelligence Review* 54 (2021), pp. 1937–1967.

[82] Y. Freund and R. E. Schapire. “A desicion-theoretic generalization of on-line learning and an application to boosting”. In: *European conference on computational learning theory*. Springer. 1995, pp. 23–37.

[83] A. Sotiras, C. Davatzikos, and N. Paragios. “Deformable medical image registration: A survey”. In: *IEEE transactions on medical imaging* 32.7 (2013), pp. 1153–1190.

[84] G. Haskins, U. Kruger, and P. Yan. “Deep learning in medical image registration: a survey”. In: *Machine Vision and Applications* 31.1 (2020), pp. 1–18.

[85] Y. Fu et al. “Deep learning in medical image registration: a review”. In: *Physics in Medicine & Biology* 65.20 (2020), 20TR01.

[86] S.-Y. Guan et al. “A review of point feature based medical image registration”. In: *Chinese Journal of Mechanical Engineering* 31.1 (2018), pp. 1–16.

[87] F. P. Oliveira and J. M. R. Tavares. “Medical image registration: a review”. In: *Computer methods in biomechanics and biomedical engineering* 17.2 (2014), pp. 73–93.

[88] T. Lindeberg. “Feature detection with automatic scale selection”. In: *International journal of computer vision* 30.2 (1998), pp. 79–116.

[89] D. G. Lowe. “Distinctive image features from scale-invariant keypoints”. In: *International journal of computer vision* 60.2 (2004), pp. 91–110.

[90] P. J. Besl and N. D. McKay. “Method for registration of 3-D shapes”. In: *Sensor fusion IV: control paradigms and data structures*. Vol. 1611. International Society for Optics and Photonics. 1992, pp. 586–606.

[91] S. Granger and X. Pennec. “Multi-scale EM-ICP: A fast and robust approach for surface registration”. In: *European Conference on Computer Vision*. Springer. 2002, pp. 418–432.

[92] A. W. Fitzgibbon. “Robust registration of 2D and 3D point sets”. In: *Image and vision computing* 21.13-14 (2003), pp. 1145–1153.

[93] S. Klein, M. Staring, and J. P. Pluim. “Evaluation of optimization methods for nonrigid medical image registration using mutual information and B-splines”. In: *IEEE transactions on image processing* 16.12 (2007), pp. 2879–2890.

[94] W. R. Crum, O. Camara, and D. L. Hill. “Generalized overlap measures for evaluation and validation in medical image analysis”. In: *IEEE transactions on medical imaging* 25.11 (2006), pp. 1451–1461.

[95] J. M. Fitzpatrick and J. B. West. “The distribution of target registration error in rigid-body point-based registration”. In: *IEEE transactions on medical imaging* 20.9 (2001), pp. 917–927.

[96] S. M. Bentzen et al. “Quantitative Analyses of Normal Tissue Effects in the Clinic (QUANTEC): an introduction to the scientific issues”. In: *International Journal of Radiation Oncology* Biology* Physics* 76.3 (2010), S3–S9.

[97] J. T. Lyman. “Complication probability as assessed from dose-volume histograms”. In: *Radiation Research* 104.2s (1985), S13–S19.

[98] G. Kutcher et al. “Histogram reduction method for calculating complication probabilities for three-dimensional treatment planning evaluations”. In: *International Journal of Radiation Oncology* Biology* Physics* 21.1 (1991), pp. 137–146.

[99] B. Emami et al. “Tolerance of normal tissue to therapeutic irradiation”. In: *International Journal of Radiation Oncology* Biology* Physics* 21.1 (1991), pp. 109–122.

[100] C. Burman et al. “Fitting of normal tissue tolerance data to an analytic function”. In: *International Journal of Radiation Oncology* Biology* Physics* 21.1 (1991), pp. 123–135.

[101] S. G. Ellsworth et al. “Declarations of Independence: How Embedded Multicollinearity Errors Affect Dosimetric and Other Complex Analyses in Radiation Oncology”. In: *International Journal of Radiation Oncology* Biology* Physics* 117.5 (2023), pp. 1054–1062.

[102] K. Boonyawan et al. “Clinical and dosimetric factors predicting grade 2 radiation pneumonitis after postoperative radiotherapy for patients with non-small cell lung carcinoma”. In: *International Journal of Radiation Oncology* Biology* Physics* 101.4 (2018), pp. 919–926.

[103] D. A. Palma et al. “Predicting radiation pneumonitis after chemoradiation therapy for lung cancer: an international individual patient data meta-analysis”. In: *International Journal of Radiation Oncology* Biology* Physics* 85.2 (2013), pp. 444–450.

[104] L. B. Marks et al. “Radiation dose–volume effects in the lung”. In: *International Journal of Radiation Oncology* Biology* Physics* 76.3 (2010), S70–S76.

[105] E. D. Yorke et al. “Correlation of dosimetric factors and radiation pneumonitis for non–small-cell lung cancer patients in a recently completed dose escalation study”. In: *International Journal of Radiation Oncology* Biology* Physics* 63.3 (2005), pp. 672–682.

[106] M. Oertel et al. “Pulmonary toxicity after total body irradiation—an underrated complication? Estimation of risk via normal tissue complication probability calculations and correlation with clinical data”. In: *Cancers* 13.12 (2021), p. 2946.

[107] B. Liang et al. “Dosiomics: extracting 3D spatial features from dose distribution to predict incidence of radiation pneumonitis”. In: *Frontiers in oncology* 9 (2019), p. 269.

[108] T. Adachi et al. “Multi-institutional dose-segmented dosiomic analysis for predicting radiation pneumonitis after lung stereotactic body radiation therapy”. In: *Medical Physics* 48.4 (2021), pp. 1781–1791.

[109] B. Liang et al. “Prediction of radiation pneumonitis with dose distribution: a convolutional neural network (CNN) based model”. In: *Frontiers in oncology* 9 (2020), p. 1500.

[110] L. Bin et al. “A deep learning-based dual-omics prediction model for radiation pneumonitis”. In: *Medical Physics* 48.10 (2021), pp. 6247–6256.

[111] B. Parashar et al. “Chemotherapy significantly increases the risk of radiation pneumonitis in radiation therapy of advanced lung cancer”. In: *American journal of clinical oncology* 34.2 (2011), pp. 160–164.

[112] J. Dang et al. “Analysis of related factors associated with radiation pneumonitis in patients with locally advanced non-small-cell lung cancer treated with three-dimensional conformal radiotherapy”. In: *Journal of cancer research and clinical oncology* 136 (2010), pp. 1169–1178.

[113] J. Wang, X. Qiao, Y. Cao, et al. “Analysis of correlated factors of radiation pneumonitis after three-dimensional conformal radiotherapy for non-small cell lung cancer”. In: *Chin J Clin Oncol* 36.19 (2009), pp. 1086–1089.

[114] J. M. Monson et al. “Clinical radiation pneumonitis and radiographic changes after thoracic radiation therapy for lung carcinoma”. In: *Cancer: Interdisciplinary International Journal of The American Cancer Society* 82.5 (1998), pp. 842–850.

[115] F. J. Núñez-Benjumea et al. “Benchmarking machine learning approaches to predict radiation-induced toxicities in lung cancer patients”. In: *Clinical and Translational Radiation Oncology* 41 (2023), p. 100640.

[116] A. Cunliffe et al. “Lung texture in serial thoracic computed tomography scans: correlation of radiomics-based features with radiation therapy dose and radiation pneumonitis development”. In: *International Journal of Radiation Oncology* Biology* Physics* 91.5 (2015), pp. 1048–1056.

[117] D. A. Palma et al. “Lung density changes after stereotactic radiotherapy: a quantitative analysis in 50 patients”. In: *International Journal of Radiation Oncology* Biology* Physics* 81.4 (2011), pp. 974–978.

[118] U. Bernchou et al. “Prediction of lung density changes after radiotherapy by cone beam computed tomography response markers and pre-treatment factors for non-small cell lung cancer patients”. In: *Radiotherapy and Oncology* 117.1 (2015), pp. 17–22.

[119] A. Bertelsen et al. “Radiation dose response of normal lung assessed by Cone Beam CT–a potential tool for biologically adaptive radiation therapy”. In: *Radiotherapy and Oncology* 100.3 (2011), pp. 351–355.

[120] S. P. Krafft et al. “The utility of quantitative CT radiomics features for improved prediction of radiation pneumonitis”. In: *Medical physics* 45.11 (2018), pp. 5317–5324.

[121] C. Puttanawarut et al. “Radiomic and dosiomic features for the prediction of radiation pneumonitis across esophageal cancer and lung cancer”. In: *Frontiers in Oncology* 12 (2022), p. 197.

[122] Z. Zhang et al. “Radiomics and dosiomics signature from whole lung predicts radiation pneumonitis: A model development study with prospective external validation and decision-curve analysis”. In: *International Journal of Radiation Oncology* Biology* Physics* 115.3 (2023), pp. 746–758.

[123] Z. Zhang et al. “Computed tomography and radiation dose images-based deep-learning model for predicting radiation pneumonitis in lung cancer patients after radiation therapy”. In: *Radiotherapy and Oncology* (2023), p. 109581.

[124] J. F. Bradley. “Data from NSCLC-Cetuximab”. In: *The Cancer Imaging Archive, DOI: http://doi.org/10.7937/TCIA.2018.jze75u7v* (2018).

[125] R. R. Colen et al. “Radiomics to predict immunotherapy-induced pneumonitis: proof of concept”. In: *Investigational new drugs* 36.4 (2018), pp. 601–607.

[126] K. Tsujino et al. “Combined analysis of V20, VS5, pulmonary fibrosis score on baseline computed tomography, and patient age improves prediction of severe radiation pneumonitis after concurrent chemoradiotherapy for locally advanced non–small-cell lung cancer”. In: *Journal of Thoracic Oncology* 9.7 (2014), pp. 983–990.

[127] J. Belderbos et al. “Acute esophageal toxicity in non-small cell lung cancer patients after high dose conformal radiotherapy”. In: *Radiotherapy and Oncology* 75.2 (2005), pp. 157–164.

[128] M. Werner-Wasik et al. “Radiation dose-volume effects in the esophagus”. In: *International Journal of Radiation Oncology* Biology* Physics* 76.3 (2010), S86–S93.

[129] L. R. Coia, R. J. Myerson, and J. E. Tepper. “Late effects of radiation therapy on the gastrointestinal tract”. In: *International Journal of Radiation Oncology* Biology* Physics* 31.5 (1995), pp. 1213–1236.

[130] S.-J. Ahn et al. “Dosimetric and clinical predictors for radiation-induced esophageal injury”. In: *International Journal of Radiation Oncology* Biology* Physics* 61.2 (2005), pp. 335–347.

[131] Z. Nesheiwat et al. “Radiation Esophagitis”. In: *StatPearls [Internet]*. StatPearls Publishing, 2022.

[132] S. Baker and A. Fairchild. “Radiation-induced esophagitis in lung cancer”. In: *Lung Cancer: Targets and Therapy* (2016), pp. 119–127.

[133] J. D. Bradley et al. “Long-term results of NRG oncology RTOG 0617: standard- versus high-dose chemoradiotherapy with or without cetuximab for unresectable stage III non–small-cell lung cancer”. In: *Journal of Clinical Oncology* 38.7 (2020), p. 706.

[134] L. B. Marks et al. “Use of normal tissue complication probability models in the clinic”. In: *International Journal of Radiation Oncology* Biology* Physics* 76.3 (2010), S10–S19.

[135] A. Ozgen, M. Hayran, and F. Kahraman. “Mean esophageal radiation dose is predictive of the grade of acute esophagitis in lung cancer patients treated with concurrent radiotherapy and chemotherapy”. In: *Journal of Radiation Research* 53.6 (2012), pp. 916–922.

[136] P. Paximadis et al. “Dosimetric predictors for acute esophagitis during radiation therapy for lung cancer: Results of a large statewide observational study”. In: *Practical radiation oncology* 8.3 (2018), pp. 167–173.

[137] J. Rose et al. “Systematic review of dose–volume parameters in the prediction of esophagitis in thoracic radiotherapy”. In: *Radiotherapy and Oncology* 91.3 (2009), pp. 282–287.

[138] A. K. Singh, M. A. Lockett, and J. D. Bradley. “Predictors of radiation-induced esophageal toxicity in patients with non-small-cell lung cancer treated with three-dimensional conformal radiotherapy”. In: *International Journal of Radiation Oncology* Biology* Physics* 55.2 (2003), pp. 337–341.

[139] W.-B. Qiao et al. “Clinical and dosimetric factors of radiation-induced esophageal injury: radiation-induced esophageal toxicity”. In: *World journal of gastroenterology: WJG* 11.17 (2005), p. 2626.

[140] R. Wijsman et al. “Multivariable normal-tissue complication modeling of acute esophageal toxicity in advanced stage non-small cell lung cancer patients treated

with intensity-modulated (chemo-) radiotherapy". In: *Radiotherapy and Oncology* 117.1 (2015), pp. 49–54.

[141] P. D. Maguire et al. "Clinical and dosimetric predictors of radiation-induced esophageal toxicity". In: *International Journal of Radiation Oncology* Biology* Physics* 45.1 (1999), pp. 97–103.

[142] M. Kwint et al. "Acute esophagus toxicity in lung cancer patients after intensity modulated radiation therapy and concurrent chemotherapy". In: *International Journal of Radiation Oncology* Biology* Physics* 84.2 (2012), e223–e228.

[143] T. H. Kim et al. "Dose-volumetric parameters of acute esophageal toxicity in patients with lung cancer treated with three-dimensional conformal radiotherapy". In: *International Journal of Radiation Oncology* Biology* Physics* 62.4 (2005), pp. 995–1002.

[144] P. G. Hawkins et al. "Prediction of radiation esophagitis in non–small cell lung cancer using clinical factors, dosimetric parameters, and pretreatment cytokine levels". In: *Translational oncology* 11.1 (2018), pp. 102–108.

[145] X. Zheng et al. "Multi-omics to predict acute radiation esophagitis in patients with lung cancer treated with intensity-modulated radiation therapy". In: *European Journal of Medical Research* 28.1 (2023), pp. 1–10.

[146] S. Wang et al. "A model combining age, equivalent uniform dose and IL-8 may predict radiation esophagitis in patients with non-small cell lung cancer". In: *Radiotherapy and Oncology* 126.3 (2018), pp. 506–510.

[147] J. S. Niedzielski et al. "A novel methodology using CT imaging biomarkers to quantify radiation sensitivity in the esophagus with application to clinical trials". In: *Scientific reports* 7.1 (2017), p. 6034.

[148] C. Ladbury et al. "Explainable artificial intelligence to identify dosimetric predictors of toxicity in patients with locally advanced non-small cell lung cancer: a secondary analysis of RTOG 0617". In: *International Journal of Radiation Oncology* Biology* Physics* (2023).

[149] S. Mostafaei et al. "CT imaging markers to improve radiation toxicity prediction in prostate cancer radiotherapy by stacking regression algorithm". In: *La radiologia medica* 125 (2020), pp. 87–97.

[150] A. Pella et al. “Use of machine learning methods for prediction of acute toxicity in organs at risk following prostate radiotherapy”. In: *Medical physics* 38.6Part1 (2011), pp. 2859–2867.

[151] A. L. D. Araújo et al. “Machine learning for the prediction of toxicities from head and neck cancer treatment: A systematic review with meta-analysis”. In: *Oral oncology* 140 (2023), p. 106386.

[152] C. Wei et al. “Development and validation of an interpretable radiomic nomogram for severe radiation proctitis prediction in postoperative cervical cancer patients”. In: *Frontiers in Microbiology* 13 (2023), p. 1090770.

[153] M. D. Zeiler and R. Fergus. “Visualizing and understanding convolutional networks”. In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Springer. 2014, pp. 818–833.

[154] D. Varshni et al. “Pneumonia detection using CNN based feature extraction”. In: *2019 IEEE international conference on electrical, computer and communication technologies (ICECCT)*. IEEE. 2019, pp. 1–7.

[155] C. Cortes and V. Vapnik. “Support-vector networks”. In: *Machine learning* 20 (1995), pp. 273–297.

[156] T. K. Ho. “Random decision forests”. In: *Proceedings of 3rd international conference on document analysis and recognition*. Vol. 1. IEEE. 1995, pp. 278–282.

[157] K. P. Murphy et al. “Naive bayes classifiers”. In: *University of British Columbia* 18.60 (2006), pp. 1–8.

[158] A. Kirillov et al. “Segment anything”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 4015–4026.

[159] P. Shi et al. “Generalist vision foundation models for medical imaging: A case study of segment anything model on zero-shot medical segmentation”. In: *Diagnostics* 13.11 (2023), p. 1947.

[160] M. Y. Lu et al. “Visual language pretrained multiple instance zero-shot transfer for histopathology images”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 19764–19775.

[161] J. W. Soh, S. Cho, and N. I. Cho. "Meta-transfer learning for zero-shot super-resolution". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 3516–3525.

[162] A. S. Kirov and L. M. Fanchon. "Pathology-validated PET image data sets and their role in PET segmentation". In: *Clinical and Translational Imaging* 2.3 (2014), pp. 253–267.

[163] T. Puri et al. "A method for accurate spatial registration of PET images and histopathology slices". In: *EJNMMI research* 5.1 (2015), pp. 1–11.

[164] R. Garcia-Parra et al. "Investigation on tumor hypoxia in resectable primary prostate cancer as demonstrated by 18 F-FAZA PET/CT utilizing multimodality fusion techniques". In: *European journal of nuclear medicine and molecular imaging* 38.10 (2011), pp. 1816–1823.

[165] H. Park et al. "Registration methodology for histological sections and in vivo imaging of human prostate". In: *Academic radiology* 15.8 (2008), pp. 1027–1039.

[166] C. Meyer et al. "Challenges in accurate registration of 3-D medical imaging and histopathology in primary prostate cancer". In: *European journal of nuclear medicine and molecular imaging* 40.1 (2013), pp. 72–78.

[167] W. Shao et al. "ProsRegNet: A deep learning framework for registration of MRI and histopathology images of the prostate". In: *Medical image analysis* 68 (2021), p. 101919.

[168] J. Stroom et al. "Feasibility of pathology-correlated lung imaging for accurate target definition of lung tumors". In: *International Journal of Radiation Oncology* Biology* Physics* 69.1 (2007), pp. 267–275.

[169] J. van Loon et al. "Microscopic disease extension in three dimensions for non–small-cell lung cancer: development of a prediction model using pathology-validated positron emission tomography and computed tomography features". In: *International Journal of Radiation Oncology* Biology* Physics* 82.1 (2012), pp. 448–456.

[170] J. Yu et al. "Comparison of tumor volumes as determined by pathologic examination and FDG-PET/CT images of non–small-cell lung cancer: a pilot study".

In: *International Journal of Radiation Oncology* Biology* Physics* 75.5 (2009), pp. 1468–1474.

[171] M. Wanet et al. "Gradient-based delineation of the primary GTV on FDG-PET in non-small cell lung cancer: a comparison with threshold-based approaches, CT and surgical specimens". In: *Radiotherapy and Oncology* 98.1 (2011), pp. 117–125.

[172] B. A. Rampy and E. F. Glassy. "Pathology gross photography: the beginning of digital pathology". In: *Surgical Pathology Clinics* 8.2 (2015), pp. 195–211.

[173] W. D. Travis et al. "IASLC multidisciplinary recommendations for pathologic assessment of lung cancer resection specimens after neoadjuvant therapy". In: *Journal of Thoracic Oncology* 15.5 (2020), pp. 709–740.

[174] A. Samani, J. Zubovits, and D. Plewes. "Elastic moduli of normal and pathological human breast tissues: an inversion-technique-based investigation of 169 samples". In: *Physics in medicine & biology* 52.6 (2007), p. 1565.

[175] P.-h. Wu et al. "Feature-based automated segmentation of ablation zones by fuzzy c-mean clustering during low-dose computed tomography". In: *Medical physics* 48.2 (2021), pp. 703–714.

[176] O. Okasha et al. "Myocardial involvement in patients with histologically diagnosed cardiac sarcoidosis: a systematic review and meta-analysis of gross pathological images from autopsy or cardiac transplantation cases". In: *Journal of the American Heart Association* 8.10 (2019), e011253.

[177] E. J. Baltussen et al. "Hyperspectral imaging for tissue classification, a way toward smart laparoscopic colorectal surgery". In: *Journal of biomedical optics* 24.1 (2019), pp. 016002–016002.

[178] Y. Yuan, M. Chao, and Y.-C. Lo. "Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance". In: *IEEE transactions on medical imaging* 36.9 (2017), pp. 1876–1886.

[179] Q. Ha, B. Liu, and F. Liu. "Identifying melanoma images using efficientnet ensemble: Winning solution to the siim-isic melanoma classification challenge". In: *arXiv preprint arXiv:2010.05351* (2020).

[180] N. Codella et al. “Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)”. In: *arXiv preprint arXiv:1902.03368* (2019).

[181] M. K. Hasan et al. “A survey, review, and future trends of skin lesion segmentation and classification”. In: *Computers in Biology and Medicine* (2023), p. 106624.

[182] D. Palma et al. “Assessment of Precision Irradiation in Early Non-Small Cell Lung Cancer and Interstitial Lung Disease (ASPIRE-ILD): Primary Analysis of a Phase II Trial”. In: *International Journal of Radiation Oncology, Biology, Physics* 117.2 (2023), S28–S29.

[183] A. Trotti et al. “CTCAE v3. 0: development of a comprehensive grading system for the adverse effects of cancer treatment”. In: *Seminars in radiation oncology*. Vol. 13. 3. Elsevier. 2003, pp. 176–181.

[184] G. Kothari et al. “The impact of inter-observer variation in delineation on robustness of radiomics features in non-small cell lung cancer”. In: *Scientific Reports* 12.1 (2022), p. 12822.

[185] T. P. Coroller et al. “Radiomic phenotype features predict pathological response in non-small cell lung cancer”. In: *Radiotherapy and oncology* 119.3 (2016), pp. 480–486.

[186] J. Hofmanninger et al. “Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem”. In: *European Radiology Experimental* 4.1 (2020), pp. 1–13.

[187] Y. Freund and R. E. Schapire. “A decision-theoretic generalization of on-line learning and an application to boosting”. In: *Journal of computer and system sciences* 55.1 (1997), pp. 119–139.

[188] N. S. McClure et al. “Minimally important difference of the EQ-5D-5L index score in adults with type 2 diabetes”. In: *Value in Health* 21.9 (2018), pp. 1090–1097.

[189] C. M. Nolan et al. “The EQ-5D-5L health status questionnaire in COPD: validity, responsiveness and minimum important difference”. In: *Thorax* 71.6 (2016), pp. 493–500.

[190] A. P. Y. Tsai et al. "Minimum important difference of the EQ-5D-5L and EQ-VAS in fibrotic interstitial lung disease". In: *Thorax* 76.1 (2021), pp. 37–43.

[191] D. Cella et al. "What is a clinically meaningful change on the functional assessment of Cancer therapy–lung (FACT-L) questionnaire?: results from eastern cooperative oncology group (ECOG) study 5592". In: *Journal of clinical epidemiology* 55.3 (2002), pp. 285–295.

[192] A. A. Raj and S. S. Birring. "Clinical assessment of chronic cough severity". In: *Pulmonary pharmacology & therapeutics* 20.4 (2007), pp. 334–337.

[193] J. D. Bradley et al. "Standard-dose versus high-dose conformal radiotherapy with concurrent and consolidation carboplatin plus paclitaxel with or without cetuximab for patients with stage IIIA or IIIB non-small-cell lung cancer (RTOG 0617): a randomised, two-by-two factorial phase 3 study". In: *The lancet oncology* 16.2 (2015), pp. 187–199.

[194] K. Clark et al. "The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository". In: *Journal of digital imaging* 26 (2013), pp. 1045–1057. DOI: 10.1007/s10278-013-9622-7. URL: https://doi.org/10.1007/s10278-013-9622-7.

[195] J. Bradley and K. Forster. "Data from NSCLC-Cetuximab". In: *The Cancer Imaging Archive* (2018). DOI: 10.7937/TCIA.2018.jze75u7v. URL: https://doi.org/10.7937/TCIA.2018.jze75u7v.

[196] A. Ebrahimi, S. Luo, and R. Chiong. "Introducing transfer learning to 3D ResNet-18 for Alzheimer's disease detection on MRI images". In: *2020 35th international conference on image and vision computing New Zealand (IVCNZ)*. IEEE. 2020, pp. 1–6.

[197] J. S. Niedzielski et al. "18F-Fluorodeoxyglucose Positron Emission Tomography can quantify and predict esophageal injury during radiation therapy". In: *International Journal of Radiation Oncology* Biology* Physics* 96.3 (2016), pp. 670–678.

[198] P. Roca et al. "Artificial intelligence to predict clinical disability in patients with multiple sclerosis using FLAIR MRI". In: *Diagnostic and Interventional Imaging* 101.12 (2020), pp. 795–802.

[199] D. Tricarico et al. “Deep regression by feature regularization for COVID-19 severity prediction”. In: *International Conference on Image Analysis and Processing*. Springer. 2022, pp. 496–507.

[200] A. Ghosh, H. Kumar, and P. S. Sastry. “Robust loss functions under label noise for deep neural networks”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 31. 1. 2017.

[201] T. Hastie et al. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009.

[202] M. Pavlou et al. “How to develop a more accurate risk prediction model when there are few events”. In: *Bmj* 351 (2015).

[203] G. A. Seber and A. J. Lee. *Linear regression analysis*. John Wiley & Sons, 2012.

[204] D. W. Marquardt. “An algorithm for least-squares estimation of nonlinear parameters”. In: *Journal of the society for Industrial and Applied Mathematics* 11.2 (1963), pp. 431–441.

[205] M. A. Ganaie et al. “Ensemble deep learning: A review”. In: *Engineering Applications of Artificial Intelligence* 115 (2022), p. 105151.

[206] O. Sagi and L. Rokach. “Ensemble learning: A survey”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8.4 (2018), e1249.

[207] J. M. Luna et al. “Machine learning highlights the deficiency of conventional dosimetric constraints for prevention of high-grade radiation esophagitis in non-small cell lung cancer treated with chemoradiation”. In: *Clinical and translational radiation oncology* 22 (2020), pp. 69–75.

[208] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

[209] N. V. Chawla et al. “SMOTE: synthetic minority over-sampling technique”. In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

[210] A. Appelt et al. “Deep learning for radiotherapy outcome prediction using dose data–a review”. In: *Clinical Oncology* 34.2 (2022), e87–e96.

[211] J. M. Borras et al. “How many new cancer patients in Europe will require radiotherapy by 2025? An ESTRO-HERO analysis”. In: *Radiotherapy and Oncology* 119.1 (2016), pp. 5–11.

[212] *Annual Report and Financial Statements 2021*. Tech. rep. Prostate Cancer UK, 2021.

[213] D. S. Bitterman et al. "Clinical natural language processing for radiation oncology: a review and practical primer". In: *International Journal of Radiation Oncology* Biology* Physics* 110.3 (2021), pp. 641–655.

[214] J. C. Hong et al. "Natural language processing for abstraction of cancer treatment toxicities: accuracy versus human experts". In: *JAMIA open* 3.4 (2020), pp. 513–517.

[215] S. Chen et al. "Deep Learning-Based Natural Language Processing to Automate Esophagitis Severity Grading from the Electronic Health Records". In: *International Journal of Radiation Oncology, Biology, Physics* 117.2 (2023), S18.

[216] M. Okada et al. "Effect of tumor size on prognosis in patients with non–small cell lung cancer: the role of segmentectomy as a type of lesser resection". In: *The Journal of thoracic and cardiovascular surgery* 129.1 (2005), pp. 87–93.

[217] W. D. Travis et al. "The IASLC lung cancer staging project: proposals for coding T categories for subsolid nodules and assessment of tumor size in part-solid tumors in the forthcoming eighth edition of the TNM classification of lung cancer". In: *Journal of Thoracic Oncology* 11.8 (2016), pp. 1204–1223.

[218] M. E. Casey and D. R. Osborne. "Siemens biograph vision 600". In: *Advances in PET: the latest in instrumentation, technology, and clinical practice*. Springer, 2020, pp. 71–91.

[219] C. A. Schneider, W. S. Rasband, and K. W. Eliceiri. "NIH Image to ImageJ: 25 years of image analysis". In: *Nature methods* 9.7 (2012), pp. 671–675.

[220] M. A. Lodge. "Repeatability of SUV in oncologic 18F-FDG PET". In: *Journal of Nuclear Medicine* 58.4 (2017), pp. 523–532.

[221] M. Tamal. "Intensity threshold based solid tumour segmentation method for Positron Emission Tomography (PET) images: a review". In: *Heliyon* 6.10 (2020).

[222] J. G. Rajendran and K. A. Krohn. "F-18 fluoromisonidazole for imaging tumor hypoxia: imaging the microenvironment for personalized cancer therapy". In: *Seminars in nuclear medicine*. Vol. 45. 2. Elsevier. 2015, pp. 151–162.

[223] X. Hou and L. Zhang. "Saliency detection: A spectral residual approach". In: *2007 IEEE Conference on computer vision and pattern recognition*. Ieee. 2007, pp. 1–8.

[224] B. Schauerte and R. Stiefelhagen. "Quaternion-based spectral saliency detection for eye fixation prediction". In: *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part II 12*. Springer. 2012, pp. 116–129.

[225] D. Arthur and S. Vassilvitskii. *k-means++: The advantages of careful seeding*. Tech. rep. Stanford, 2006.

[226] S. R. Hashemi et al. "Asymmetric loss functions and deep densely-connected networks for highly-imbalanced medical image segmentation: Application to multiple sclerosis lesion detection". In: *IEEE Access* 7 (2018), pp. 1721–1735.

[227] C. H. Sudre et al. "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations". In: *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 2017, pp. 240–248.

[228] M. Sandler et al. "MobileNetV2: Inverted Residuals and Linear Bottlenecks". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 4510–4520. DOI: 10.1109/CVPR.2018.00474.

[229] J. Deng et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[230] P. Tschandl, C. Rosendahl, and H. Kittler. "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions". In: *Scientific data* 5.1 (2018), pp. 1–9.

[231] M. Gil et al. "A Deep Learning Based Approach to Semantic Segmentation of Lung Tumour Areas in Gross Pathology Images". In: *Annual Conference on Medical Image Understanding and Analysis*. Springer. 2023, pp. 18–32.