CENTRE FOR DOCTORAL TRAINING IN APPLIED PHOTONICS
industry inspired: imaging | sensing | analysis

Canon
CANON MEDICAL RESEARCH EUROPE LTD.

University of Glasgow
VIA VERITAS VITA

Deep Learning for Lung Cancer Analysis

Owen Anderson

Submitted for the Degree of Doctor of Engineering

University of Glasgow
School of Computing Science

AI Centre for Excellence
Canon Medical Research Europe

Academic supervisors: Dr Paul Siebert and Dr Martin Lavery

Industrial supervisor: Dr Keith Goatman

May 2023

Abstract

This thesis describes the development and evaluation of two novel deep learning applications that tackle two cancers that affect the lungs. The first, lung cancer, is the largest cause of cancer-related deaths in the United Kingdom. It accounts for more than 1 in 5 cancer deaths; around 35,000 people every year. Lung cancer is curable providing it is detected very early. Computed tomography (CT) X-ray imaging has been shown to be effective for screening. However, the resulting 3D medical images are laborious for humans to read, and widespread adoption would put a huge strain on already stretched radiology departments. I developed a novel deep learning based approach for the automatic detection of lung nodules, potential early lung cancer, that has potential to reduce human workloads. It was evaluated on two independent datasets, and achieves performance competitive with published state-of-the-art tools, with average sensitivity of 84% to 92% at 8 false positives per scan. I developed a related invention which allows hierarchical relationships to be leveraged to improve the performance of CAD tools like this for detection and segmentation tasks.

The second cancer is malignant pleural mesothelioma. It is very different from lung cancer: rather than growing within the lung, mesothelioma grows around the outside of the lung in the pleural cavity, like the rind on an orange. It is a rare cancer, caused by exposure to asbestos fibres. It can take decades from exposure to symptoms developing. In Glasgow many mesothelioma patients worked in the ship-building industry, which relied heavily on asbestos. Although asbestos has been banned in the UK since 1999, its presence in buildings and equipment built before then mean that mesothelioma will remain a problem for years to come. Sadly, asbestos is still being mined and many countries, including the United States, have still not instigated a complete ban. For mesothelioma the main challenge is not detection, but accurate measurement —- without the ability to measure tumour size it is difficult to evaluate potential treatments. We therefore developed a fully automated volumetric assessment of malignant pleural mesothelioma. Performance of the algorithm is shown on a multi-centre test set, where volumetric predictions are highly correlated with an expert annotator ($r=0.851$, $p<0.0001$). Region overlap

scores between the automated method and an expert annotator exceed those for inter-annotator agreement across a subset of cases. Dice overlap scores of 0.64 and 0.55, by cross-validation and independent testing respectively, were achieved. Future work will progress this algorithm towards clinical deployment for the automated assessment of longitudinal tumour development.

To my parents, family and friends,
whom I love dearly.

Acknowledgements

The authors thank the National Cancer Institute for access to NCI's data collected by the National Lung Screening Trial (NLST). The statements contained herein are solely those of the authors and do not represent or imply concurrence or endorsement by NCI.

Thank you Vismantas Dilys and James Sloan for inspiring and tutoring me. Thank you Alison O'Neil and Aneta Lisowska for mentoring me. This work wouldn't have been possible without the Centre for Doctoral Training in Applied Photonics, my supervisors Martin Lavery and Paul Siebert, a fantastic group of collaborators, and Canon Medical Research Europe. Thank you especially Sandy Weir, Ian Poole and Keith Goatman — a fantastic group of leaders, without whom this thesis wouldn't be here.

Owen Anderson

March 2023

Contents

List of Figures

List of publications

Publications

Chapter **O. Anderson**, A. C. Kidd, K. A Goatman, A. J. Weir, J. P. Voisey, V. Dilys, J. P. Siebert and K. G. Blyth, "Fully Automated Volumetric Measurement of Malignant Pleural Mesothelioma from Computed Tomography Images by Deep Learning: Preliminary Results of an Internal Validation", *Biomedical Engineering Systems and Technologies. BIOSTEC 2020. Communications in Computer and Information Science. Springer, Cham.*, vol. 1400 (30 Mar 2021)
doi:/10.1007/978-3-030-72379-8_7

Article A.C. Kidd, **O. Anderson**, G. W. Cowell, A. J. Weir, J. P. Voisey, M. Evison, S. Tsim, K. A. Goatman and K. G. Blyth, "Fully automated volumetric measurement of malignant pleural mesothelioma by deep learning AI: validation and comparison with modified RECIST response criteria", *Thorax*, (02 Feb 2022)
doi:/10.1136/thoraxjnl-2021-217808

Chapter D. Zotova, A. Lisowska, **O. Anderson**, V. Dilys and A. Q. O'neil, "Comparison of Active Learning Strategies Applied to Lung Nodule Segmentation in CT Scans", *Large-Scale Annotation of Biomedical Data and Expert Label Synthesis and Hardware Aware Learning for Medical Imaging and Computer Assisted Intervention. LABELS HAL-MICCAI CuRIOUS 2019 2019 2019. Lecture Notes in Computer Science. Springer, Cham.*, vol. 11851 (24 Oct 2019)
doi:10.1007/978-3-030-33642-4_1

Proceeding **O. Anderson**, A. C. Kidd, K. A Goatman, A. J. Weir, J. P. Voisey, V. Dilys, J. P. Siebert and K. G. Blyth "Fully Automated Volumetric

Measurement of Malignant Pleural Mesothelioma from Computed Tomography Images by Deep Learning: Preliminary Results of an Internal Validation", *International Joint Conference on Biomedical Engineering Systems and Technologies, Volume 2: BIOIMAGING*, 24-26 Feb 2020
doi:10.5220/0008976100640073

Patents

Patent J. Sloan, **O. Anderson**, K. A. Goatman "Registration method and apparatus" U.S. Patent US 20210073939A1, (11 Mar 2021)

Patent **O. Anderson**, A. Lisowska, A. Q. O'neil "Data processing apparatus and method" U.S. Patent US 20210225508A1, (22 Jul 2021)

Patent **O. Anderson**, A. Lisowska, A. Q. O'neil, K. A. Goatman "Data processing apparatus and method" U.S. Patent US 20220020142A1, (20 Jan 2022)

List of Symbols and notation

Unless specified, the lowercase letters indicate the dimensionless or normalised quantities, while the uppercase letters indicate the quantities with dimensions.

lr	learning rate
w	weight
b	bias
Y	ground truth label
$\hat{Y}$	predicted label
FP	False Positive
TP	True Positive
FPR	False Positive Rate

List of Acronyms

MPM	Malignant Pleural Mesothelioma
LND	Lung Nodule Detection
DL	Deep Learning
ML	Machine Learning
NN	Neural Network
CNN	Convolutional Neural Network
LOA	Limits of Agreement
CI	Confidence Interval
CAD	Computer Assisted Detection
AUC	Area Under the Curve
ROC	Reciever Operating Characetristic
FROC	Free-response Reciever Operating Characteristic

Chapter 1

Introduction

This thesis describes the development of Deep Learning (DL) based algorithms to detect and quantify two types cancers which affect the lungs from computed tomography (CT) images: lung nodules and malignant plural mesothelioma (MPM). This chapter will provide the reader with an introduction to the anatomy of the lungs (Section 1.1), an introduction to medical imaging modalities, including CT imaging (Section 1.2). The chapter presents an overview of DL techniques (on which both of the presented algorithms are based), providing examples of a DL technology which has been approved and deployed in clinical practice (Section 1.4 and 1.3). Finally, the chapter concludes with an overview of the contributions presented in this thesis.

1.1 The Lung and Cancer

This thesis concerns algorithms for two types cancer which affect the lungs. This section provides an introduction to the anatomy of the lungs to provide underlying anatomical context, and an introduction to lung cancer.

1.1.1 The Lung

Due to the position of the heart in the body, our lungs are not symmetrical. This asymmetry is shown in figure 1.1, where the left lung is usually constructed of two lobes, and the right lung usually has three. These lobes are separated by thin fissures. Variation in the number of lobes is reasonably common among individuals, and it is important for professionals to be capable of distinguishing variations which are normal from those variations which arise from disease.

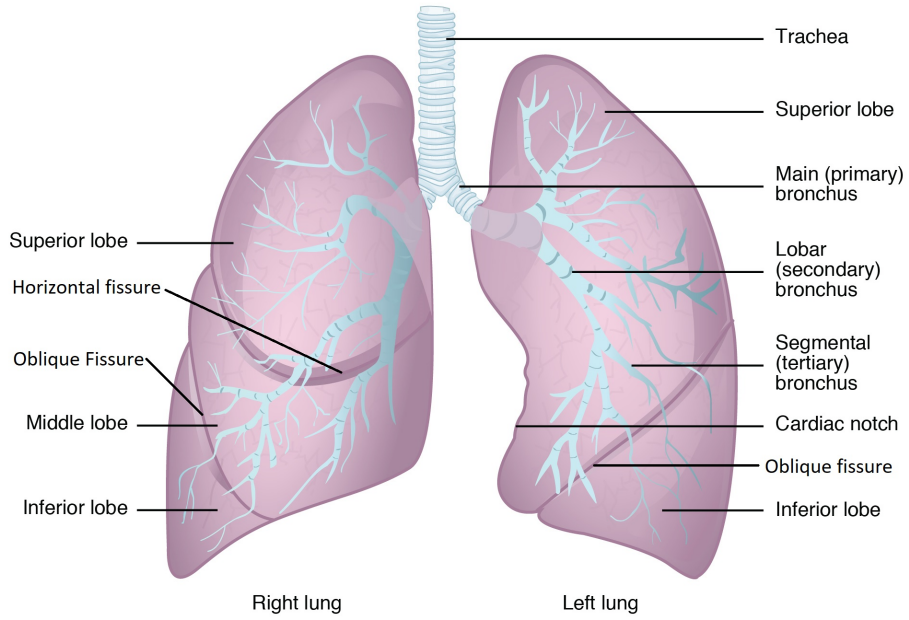


Figure 1.1: *The anatomy of the lungs, with annotated structures. Image taken from [1]*

The two lungs themselves are separated by a structure called the mediastinum, and the lung surface which faces this structure is referred to as the mediastinal surface. To facilitate contraction and expansion of the lungs, each reside within a lining with a smooth surface called the pleura. This lining usually contains a small amount of fluid, which is secreted by capillaries and cleared by the lymphatic system. An excess in this fluid is referred to pleural effusion, which results with increased pressure on the lungs, difficulty breathing and in extreme cases leads to lung collapse.

This can occur in later stage cancers.

The alveoli are small clusters of sacks within the lungs where oxygen exchange occurs, supplied by thin tubes called bronchioles which join to form larger bronchi, eventually reaching the trachia which exit the lung.

The lungs are responsible for oxygenating the blood and removing carbon dioxide, and for the reason the role of arteries and veins are reversed with respect to elsewhere in the body — here the pulmonary veins carry oxygenated blood from the periphery of the lungs to the heart for circulation, and arteries carry the de-oxygenated blood into the lungs.

1.1.2 Lung Cancer

Lung cancer is the leading cause of cancer mortality worldwide [2, 3], and accounts for almost 30% of annual cancer related deaths in Scotland [4]. Lung cancer can be divided into two categories: small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC), where the majority (85%) of cases are NSCLC [5]. These cancer types differ in biology, and generally NSCLC tumours are comprised of larger cells and are slower growing with respect to SCLC.

To guide the provision of treatment, tumours are graded by either a number staging system or the Tumour Node Metastasis (TNM) staging system [6, 4]. These staging systems aim to summarise how large the tumour is and whether the tumour has spread beyond its original site. The TNM provides a more detailed description. It is less commonly used. The number staging system is comprised of four main stages:

- stage I indicates the cancer has not spread, and remains small;
- stage II indicates that the cancer has not spread, but has grown;
- stage III indicates that the cancer may have spread to the lymph nodes or surrounding tissue;
- stage IV indicates that the cancer has spread to at least one other site.

For tracking tumour development longitudinally, the clinically standard measurement is called the RECIST (Response Evaluation Criteria in Solid Tumours) scoring system [7]. Based on the assumption that tumours are spherical, their volume

can be approximated by measuring the tumour diameter. Although tumours often start as spherical when their growth is unimpinged, spatial variations in a number of biological factors result in late stage tumours which are often irregular in shape [8]. Crucially, the RECIST score is a measurement of disease progression, and diameter measurements are compared longitudinally to asses whether the cancer has progressed. This is achieved by using thresholds to classify the measurement into four categories: disease progression, no change, partial response to treatment, and disappearance of all known disease.

1.1.3 Disease Treatment

The importance of early detection derives from the fact that early-stage disease is more treatable. Treatment is comprised of three main options: surgery, radiotherapy and chemotherapy.

Surgical intervention is the standard of care for early stage cancer in those fit enough to undergo such a procedure [9]. Surgery is less common for SCLC due to its quicker onset and increased likelihood to spread with respect to NSCLC. Surgery can involve either the removal of a small portion of diseased lung (for particularly early disease), the removal of a lobe of the lung, or the removal of an entire lung.

Radiotherapy is the process of destroying cells using ionising radiation, and is another common treatment for lung cancer, either as a palliative measure (to ease symptoms) or to attempt to remove disease. This process involves radiotherapy planning, where the location of the tumour within the body is used to assess how tumour may be targeted to minimise damage to surrounding structures. As with surgical intervention, there are a number of practical considerations to assess whether the patient and disease characteristics are suitable for such an intervention.

Chemotherapy, generally speaking, is used as an adjuvant treatment. It can be deployed prior to surgery, to shrink the tumour ahead of resection; after surgery, to ensure that disease does not return; or combined with radiotherapy, to increase efficacy of treatment. It may also be deployed for palliative care, to ease disease symptoms. There are a wide range of pharmaceutical options available, and NICE (National Institute for Heath and Care Excellence) offer a wide range of clinical care

guidelines for the treatment of lung cancer [10], which depend on the specifics of the case. Treatment efficacy can depend on the mutational status of genes. For example, for late stage non-squamous carcinoma, where the tumour exhibits a specific mutation (named T790M) in a specific gene (named EGFR) the drug Osimertinib is indicated as the optimal treatment. The pharmacological landscape is increasingly complex as new drugs are found, with novel mechanisms of action, the efficacy of which depends on the biological nature of the disease.

In order for lung cancer to be detected non-invasively, and longitudinally tracked over the course of development, medical imaging techniques are routinely used.

1.2 Medical Imaging

> When you can measure what you are speaking about, and express it in numbers, you know something about it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts advanced to the stage of science. — *Lord Kelvin*

Medical imaging is measurement of the human body to inform diagnosis, treatment and understanding. In this section we introduce X-ray, Computed Tomography and Magnetic Resonance imaging.

1.2.1 X-Ray Imaging

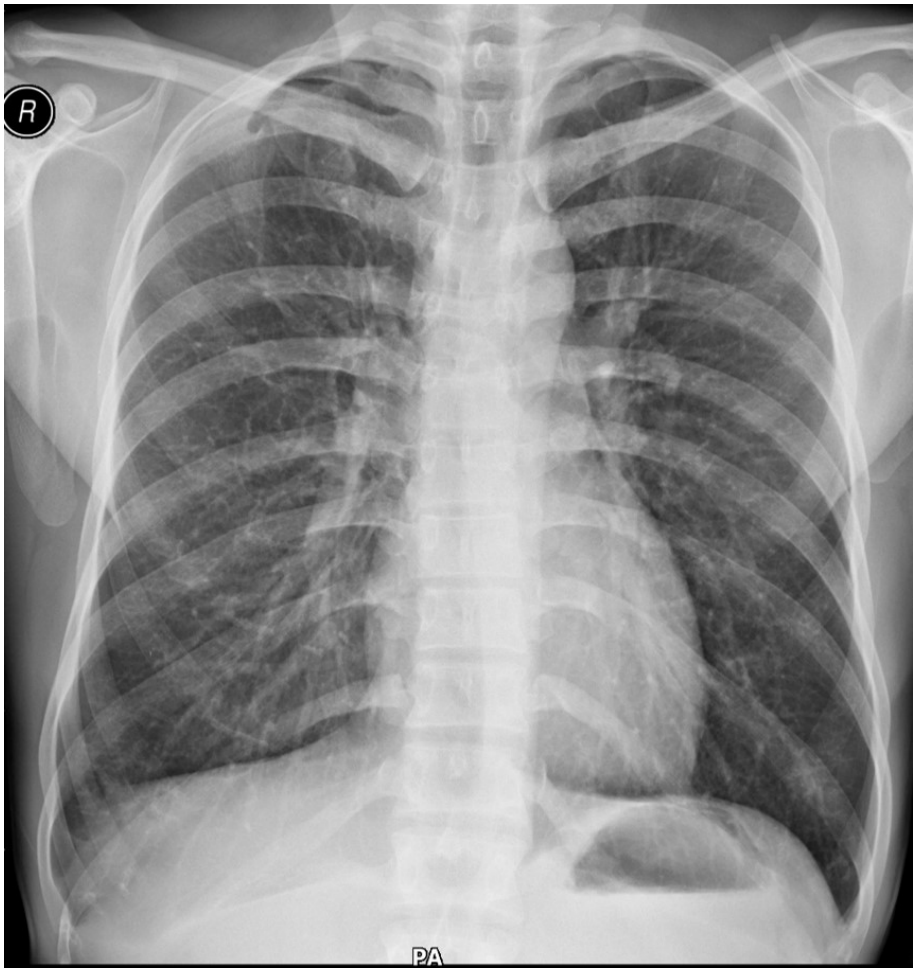


Figure 1.2: *An example X-ray image of the chest.*

X-rays are electromagnetic waves which reside in the wavelength range of 10 picometers to 10 nanometers. X-rays can penetrate much deeper into biological

tissue than natural light, and have been used for imaging since the late 19^{th} century. To generate images using X-rays, an X-ray source and X-ray detector are required.

In the early days of X-ray imaging, it was considered that this form of electromagnetic radiation was no more dangerous than natural light, however X-ray photons are many thousands of times more energetic. At certain exposure levels they can cause lasting damage to biological material, and those who were regularly exposed to this radiation developed cancer and other health issues later in life. Now, minimising radiation dose is an important factor in medical imaging involving X-rays.

X-ray images are projection images and are 2-dimensional, resulting in a level of ambiguity in discerning overlapping image features. This is shown in figure 1.2, where the complex and overlapping vasculature surrounding the heart is impossible to disambiguate by this single view. They are also uncalibrated images, and units of intensity are only meaningful with respect to other regions in the image. Despite these limitations, the modality is the most common modality for medical imaging due to the low cost and simplicity of the image acquisition process. A level of disambiguation can be achieved by taking multiple images from multiple angles, and this is the foundation of CT imaging, where many hundreds of cross-sectional X-ray measurements are integrated.

1.2.2 Computed Tomography Imaging

Computed Tomography (CT) imaging uses a motorised X-ray source and detector pair to collect 3-Dimensional images. The source and detector array are mounted in a circular gantry, and are rotated around the subject to be imaged. This allows X-ray absorption to be measured as a function of angle in a 'sinogram'. Each sinogram acquired can be converted into a CT slice by solving a set of equations in a process called back projection. In order to generate a full 3-D image, the subject must be moved through the gantry as multiple slices are acquired.

Example CT images are shown in 1.4, provided for the axial, sagittal and coronal planes. As 3-D volumes are generated slice-wise, the axial resolution typically differs from that achieved in the sagittal and coronal planes. Typically, modern scanners may provide an axial resolution of 0.5 - 0.8 mm per voxel. Resolution in the sagittal and coronal planes is usually lower than that achieved in the axial plane, and both slice thickness and slice spacing can be varied. These parameters will be selected

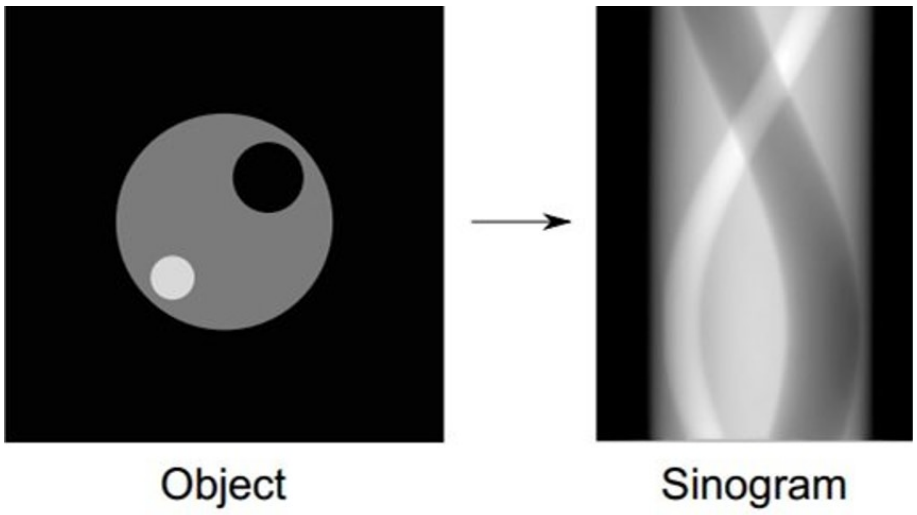


Figure 1.3: *An example dummy object and corresponding sinogram. The horizontal axis of the sinogram relates to position on detector, and the vertical axis shows angle of acquisition. Image taken from [11].*

by the radiographer based on the nature of the examination being undertaken. The radiographer will try to minimise radiation dose, whilst allowing for sufficient resolution.

Images generated by conventional back projection can be prone to blurring, and filtering is applied as a component of the image reconstruction process. The filtering is performed on the sinogram prior to back projection. Filtering kernels are chosen based on the type of examination being undertaken. A sharper (but grainier) image may aid in the analysis of bone fractures, and images with a smoother appearance are more appropriate for the analysis of soft tissues. The image reconstruction kernels are often proprietary to the scanner manufacturer, and reconstruction kernel is one example of how CT image characteristics can differ between scanners.

Hounsfield units (H.U.) are universally used to express CT imaging intensities. The units are calibrated such that 0 H.U. pertains to the radio-density of water and -1000 H.U. pertains to the radio-density of air. Intravenous contrast agent, with a high radio-density, may be used during imaging to highlight regions which may ordinarily be difficult to delineate from their surroundings, or to make functional measurements of tissues based on the uptake of contrast agent.

1.2.3 Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) is a method of acquiring 3-D images which does not depend on X-rays. The method uses strong magnetic fields and radio-frequency

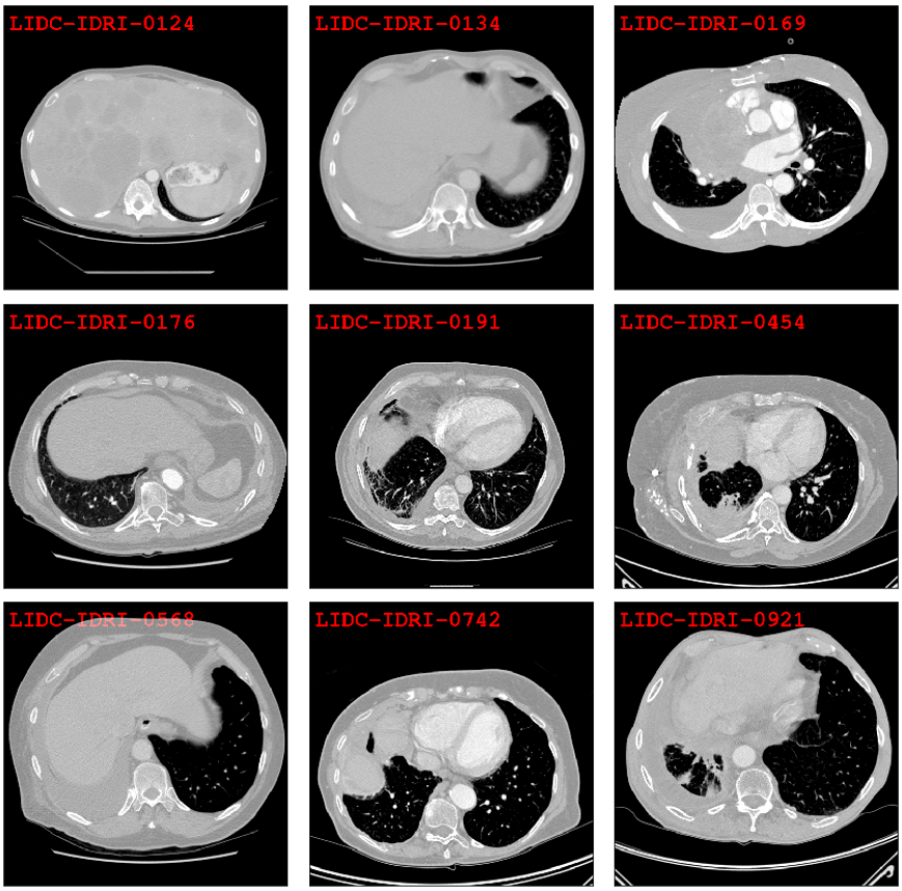


Figure 1.4: *Example axial views of CT images from nine patients, centered on the lungs, from the publicly available LIDC-IDRI dataset.*

radiation to to spatially characterise hydrogen atoms within the body. MRI is readily applied in measurements of soft tissues, where good contrast can be achieved between similar density structures such as fat and muscle.

Though MRI has a number of advantages with respect to other imaging techniques, its use is not as wide-spread as CT imaging, even for tasks where MRI would lead to less ambiguous diagnosis. MRI scanners are less common, and the imaging process is longer in duration with respect to CT, and thus more demanding for the patient.

The interpretation of medical images for diagnosis and to inform treatment requires significant expertise. The process of making measurements on such images can also be time consuming, and many automated and semi-automated tools have been developed to assist with the interpretation of medical images.

1.3 Medical Image Analysis

Whilst the theoretical and mathematical foundation for computer assisted bio-imaging has long been conceived, they are only beginning to bare fruit in the context of routine patient care. Historically the blocking factor has been technological, namely the compute required to perform such analysis at scale. Since the turn of the century, the blocking factor has become the digitisation of the medical field, or the lack thereof. Increased digitisation promises a harmonised medical experience, where a patients records can be seamlessly transposed between medical professionals and departments. Decisions and measurements may be recorded, stored, analysed and shared in an efficient manner.

This harmonisation requires development of significant infrastructure. Medical tools, systems, protocols and records remain largely disparate, with significant variation between local institutions. Only with this harmonisation can automated analysis become routine.

Another complex consideration is moral and legal in nature. Since the turn of the century we have witnessed a technological revolution. Many of the largest companies of the 21st century deal in data — a resource which is freely given, and used in ways which are not widely understood. Adverts, news articles and information can be targeted, and data is traded between private institutions in ways which have far-reaching and complex implications for society. The legal structures regarding this new digital age have taken time to emerge, and perhaps there is no data more personal and important than that data which pertains to ones health. As such, a complex landscape of data governance and regulation is symbiotically evolving as new technological capabilities are fully realised. Whilst such legislation is important, it acts to temper radical innovation. Start-up companies embarking to revolutionise healthcare are faced with a complex and ever-changing legal landscape which reduces the emergence of disruptive technologies.

For medical images, digitisation has progressed further than for other modalities of medical data. Picture Archiving and Communication systems (PACs) are widely used to store images in a standard Digital Imaging and Communications in Medicine (DICOM) format. The task of data harmonisation began early for this domain, and there is a fertile ground for the deployment of automated tools.

1.3.1 Role of Automated Medical Imaging Tools

Whilst new data-driven technologies revolutionise many aspects of modern life, there is pervasive hype and reactive contention surrounding how this will impact our collective future. Self-driving goods vehicles, automated farming technologies and drone delivery systems may make many vocations redundant. Within the last decade, many radiologists have become concerned that artificial intelligence (AI) will take their profession.

Whilst the distant future remains unclear, current technologies cannot be the sole carriers of responsibility for patient analysis and the resultant decisions. One concern may be that this insertion of technology may act to divorce the sacred relationship between doctor and patient. However, AI (as with other forms of digitisation) is likely to provide the greatest benefits where used to facilitate analysis which would otherwise be too tedious, expensive or time-consuming to be routine. Designing algorithms for problems of this nature ensures that the contribution of technology is additive, rather than a substitute for elements in the patient care pathway.

For example, if an individual receives a CT scan of the lung, it is feasible that this could be retrieved from PACs automatically, and screened for a number of incidental findings for review which may have otherwise been missed. Rather than replace the role of a radiologist in this setting, it has empowered clinicians with information which would otherwise be unavailable.

1.3.2 The need for regulatory approval of clinical applications

The deployment of AI in medical practice for image analysis requires software to pass the same level of validation as other medical devices. According to the Data Science Institute [12], the American Food and Drugs agency have currently approved 201 medical algorithms for radiology or other imaging derived tasks. Most of these algorithms target a specific body area and modality of imaging. For example, Siemens have achieved FDA approval for a Lung Computer Assisted Detection (CAD) tool for lung nodule detection from CT. Some algorithms can be used across multiple modalities, for example SubtlePET is an FDA approved noise reduction algorithm for use on PET, CT and MR imaging.

As with any new technology, the approval processes have been (and continue to

be) subject to update to allow new technologies to come to market. For example, emerging AI technologies may learn dynamically during deployment based on new data. This would result in algorithms which may change behaviour (and performance) after each new update, and updates may occur more frequently than for non-AI tools. Previously, the FDA has approved algorithms which are 'locked', and if an update is required which modifies the existing risks associated with the software, additional approval must be sought. Currently, the FDA is re-imagining this product life-cycle approach to facilitate the more dynamic technologies based on AI.

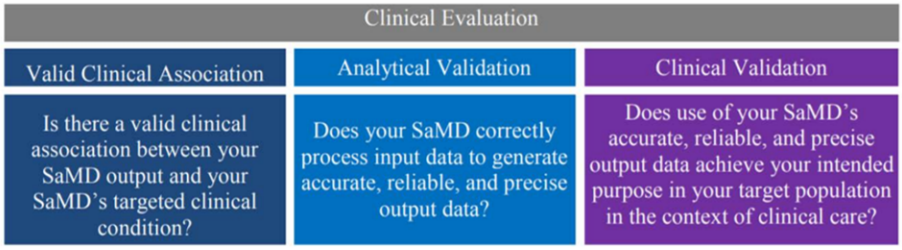


Figure 1.5: *A figure from the 'Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machience Learning Based Software as a Medical Device (SaMD)' discussion paper by the FDA. [13]*

At the heart of approving and marketing any Software as a Medical Device (SaMD) is rigorous software evaluation. Figure 1.5 shows the 3 components to an FDA evaluation. A valid clinical association is required to assess whether the software adequately match the stated clinical problem. An analytical validation is required, to determine whether the software is shown as numerically accurate by the provision of statistics. And finally, a clinical validation is required. When the software is deployed in its clinical environment, the clinical evaluation intends to ascertain whether it achieves its purpose. For example, it is possible that a poorly packaged tool with a bad interface may hinder any clinician, regardless of the numerical numerical performance of the underlying algorithm.

One algorithm for which there exists a detailed public record of the approval details is the HeartFlow algorithm, which provides the opportunity for a specific case study of a DL based technology which has been deployed at scale in clinical practice.

HeartFlow Algorithm

Clinical Problem: Coronary heart disease occurs when the flow of blood to the heart is blocked or reduced, and can be caused by a build up of calcified plaques in

the arteries. The gold-standard diagnostic measurement is called Fractional Flow Reserve (FFR), where the pressure in the arteries measured using a catheter and a small probe in a procedure called invasive coronary angiography [14]. HeartFlow use CT imaging, image analysis and simulation to automatically estimate the pressure in the blood vessels surrounding the heart in a non-invasive manner. This involves analysing a CT image of the coronary vasculature using deep learning to generate a 3D segmentation of the blood vessels. Computational fluid dynamics are are calculated based on this segmentation, which estimates the pressure at all locations within the arteries. HeartFlow provide these capabilities using cloud computing.

Validation: The National Institute for Healthcare Excellence (NICE) document the dialogue between HeartFlow and the external approval committee while seeking regulatory approval [15]. Initially, HeartFlow conducted a literature survey across 22 published studies on the diagnostic capabilities of existing diagnostic tests for coronary heart disease for benchmarking their method. In light of concerns around the specific deployment of the algorithm within the care pathway, the review committee also conducted their own survey to contribute to the meta-analysis between a HeartFlow trail [14] and several other non-invasive comparators at the per-patient and per-vessel level of analysis. Following this, reviews were performed of the clinical effectiveness evidence before approval was granted.

Impact: In the report by the National Institute for Health and Care Excellence (NICE) [15], the HeartFlow algorithm reduces false positives by as much as 50%. It provides an estimated saving to the NHS for £391 per patient relative to other noninvasive tests (based on SPECT, MRI or ECHO imaging), and has been already used on around 15,000 patients in the U.K. The report estimates that this will save the NHS £9.1m each year.

Many existing medical image analysis tools depend on pipelines with manually engineered features. AI (or deep learning) approaches involve automatically extracting these features directly from the data.

1.4 Deep Learning

Deep learning, machine learning (ML) and artificial intelligence (AI) are all terms which describe a class of algorithms which are automatically fit to a distribution of data. Early Neural Networks (NNs) were used to recognise handwritten digits in 1990 [16]. The number of publications around AI has risen from 596 in 2010 to 12,422 in 2019 [17], an uptake which can largely be attributed to advancements in computer hardware [18]. The development of these techniques continues to have significant impact on the modern world: changing the way we consume media, removing human drivers from cars, improving the accuracy of targeted advertisement and more. This impact is recent, but the fundamental concepts of deep learning were in place as early as the 1950s, starting with the Perception [19].

1.4.1 The Perceptron and Neural Networks

A perceptron is an artificial neuron which has multiple inputs and a single output. The perceptron takes inputs $(x_1, x_2, x_3...)$, and performs a weighted sum (with weights $w_1, w_2, w_3...$) of the inputs with an added bias (b). The output of the perceptron is described as:

$$output = \begin{cases} 0 & \sum_j w_j x_j + b \leq 0 \\ 1 & \sum_j w_j x_j + b > 0. \end{cases} \tag{1.1}$$

A detailed description of the perceptron is provided by Nielsen [20]. In isolation, a single perceptron can (given suitable weights and biases) perform the operation of a NAND gate, an operation which is universal — combinations of NAND gates can be used to compute all other (AND, OR and NOT) operations. When many perceptrons are combined in "layers" to create a NN (as shown in Figure 1.6) more complex functions can be approximated. NNs are universal approximators of continuous functions [21] — in theory any function can be modelled by fitting a NN. This theory holds provided there are a sufficient number of perceptrons and layers in the NN for the function to be approximated.

This is the basis of neural networks. The number of perceptrons (or neurons) in modern networks is large, and manually tuning the weights and biases (jointly termed free parameters) is unfeasible. The free parameters are automatically optimised using

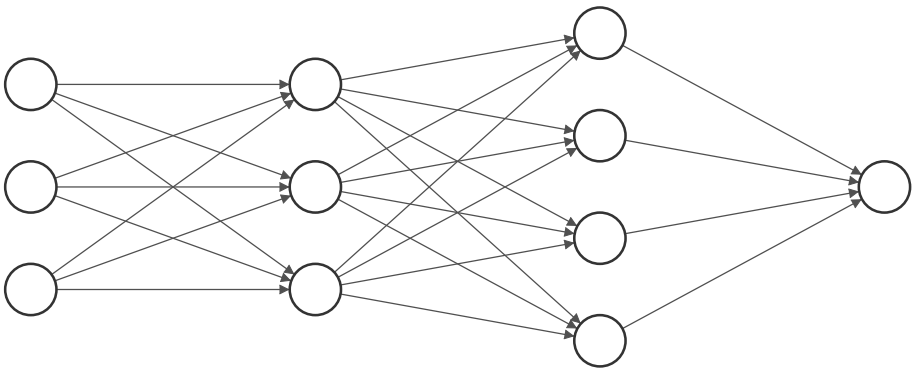


Figure 1.6: *A deep network of perceptrons. The inputs (left) are processed by the perceptrons (shown as circles), to generate an output (right). The network has four layers: an input layer, two hidden layers and an output layer.*

a method called back-propagation [20].

To optimise the free parameters, matched inputs and 'ground truth' are required. Ground truth is a term for the correct output — the output which the neural network is optimised to produce given the corresponding input. Based on comparing the output of the neural network with the ground truth using an error function, an error is calculated. This is a single scalar value which summarises network performance as measured by the ground truth. The process of back-propagation involves differentiating the error with respect to the free parameters of the network. This differential is used to alter the free parameters of the network to reduce the error. Specifically, the error is calculated over one batch of training data, this is back-propagated to provide a direction in which to tune the free parameters of the network, and a perturbation of the parameters is applied in this direction. The magnitude of the perturbation is set by the learning rate, which may be tuned manually or automatically. There are a number of common error functions which may be suitable, dependent on the the nature of the task. Examples of these are listed in Table 1.1.

The network structure in Figure 1.6 is called a dense network — a single neuron is connected to every neuron in the subsequent layer. Such a structure may be appropriate when the input is small, however when the input is large (e.g. some hundreds of thousands of pixels in an image), and the task is complex (requiring more layers), a dense network is computationally intensive. Another network structure is commonly used for the processing of images.

Name	Error Function	Description and Usage
Mean Squared Error (MSE)	$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	Networks which output continuous variables (e.g. image reconstruction).
Binary Cross-entropy	$\frac{1}{n}\sum_{i=1}^{n} Y_i \log \hat{Y}_i + \left(1 - Y_i\right) \log\left(1 - \hat{Y}_i\right)$	Networks optimised to classify an input as a binary class.
Categorical Cross-entropy	$\frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{C} Y_i \log \hat{Y}_i$	Networks optimised to classify an input as a multiple potential binary classes (the number of classes defined as C).

Table 1.1: *A list of some common error functions, where Y_i describes the ground truth, $\hat{Y}_i$ describes the output of the neural network and n is the number of training examples in each batch.*

1.4.2 Convolutional Neural Networks

When considering an image (or time series data), data points which are closer together spatially (or temporally) are more related than those which are apart. Global interaction of the data points in a single layer, as offered by a dense network structure, is an inefficient way to deal with such data. Convolutional neural networks (or CNNs) operate by sequentially filtering the input data with filters which are optimised during the training process [22]. This offers a number of benefits in the context of image analysis:

1. efficient computation — the filters are spatially localised, reducing the number of calculations required;

2. spatial invariance — the same filters are used to analyse different regions of the image;

3. parallelism — it is possible to parallelise and accelerate this process on modern Graphics Processing Units (GPUs).

Figure 1.7 shows a schematic of a typical CNN. Spatially shared weights (or kernels) are used to filter the input data. The kernel is applied at every location on the input to produce an activation, which measures how closely the input data matches the pattern captured in the kernel. When applied to the entire input, the output of this processing is an activation map which spatially describes the presence (or absence) of the pattern captured in the kernel.

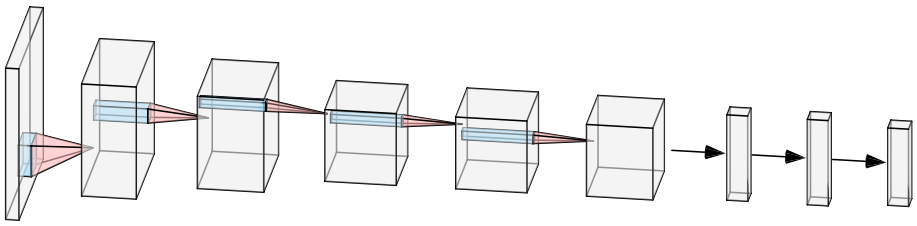


Figure 1.7: *An example CNN with 9 layers. The first layers (left) are convolutional, where shared weights are used for processing. The final three layers are dense layers, where the spatial data is flattened to a single dimension.*

Typically, a layer of processing in a CNN may contain tens to thousands of different kernels. Early layers contain filters which capture low-level information e.g. textures and edges. Deeper into the network, the kernels are filtering activation maps from earlier layers. These will detect the presence (or absence) of higher level features (e.g. combinations of textures and edges). There are many more high level features than low level features, and so the most successful CNN architectures have more kernels at deeper layers. To account for this expansion in the number of feature maps, this is often paired with a reduction of spatial resolution. This is performed using pooling layers (where reduction is performed by a local maximum or averaging operation), or by strided convolutional operations (where the kernel is applied at a subset of locations).

CNNs have been readily applied to image classification, where images are reduced to a single classification probability which describes the presence (or absence) of certain objects within the image. The benefit of convolutional architectures can be fully realised on image segmentation tasks.

1.4.3 Image Segmentation and the U-Net

In the field of computer vision, image segmentation is the process of delineating an image into multiple segments. For example, in the context of automated vehicles, these segments could describe the extent of the road or the area of the pavement. Where previously we considered the assignment of a single global value to an image (e.g. whether the pavement or road is present in the image), here we require a classification assigned at the individual pixel level (or voxel level, where considering a 3D image).

The same efficiency that convolutional methods offer to the image encoding

process can be leveraged for decoding, as popularised by the U-Net architecture in the seminal 2015 paper by Ronneberger *et al.* [23].

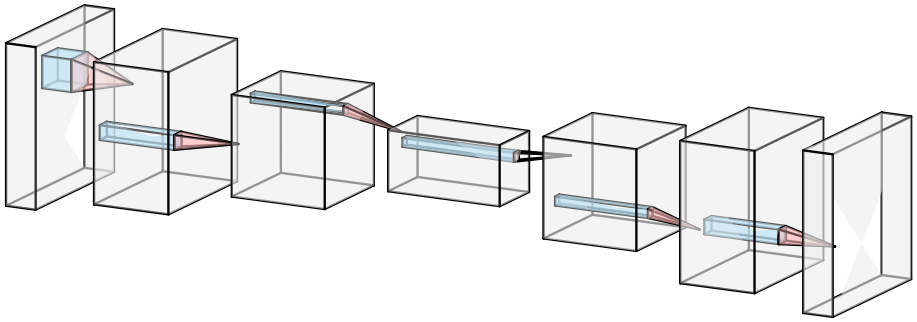


Figure 1.8: *An example of a U-Net architecture, where sequential convolutional processing reduces an input image into an intermediate embedding, which is decoded to an output which is has the same dimensions as the input.*

Figure 1.8 shows an example U-Net structure, where the input image is encoded and decoded by sequential convolutional processing. Crucially, the output has the same dimensionality as the input. Skip connections (which are not shown in this diagram) propagate feature maps from the encoding layers directly to the decoding layers, skipping the bottleneck. This allows high resolution features to be mainteined during the encoding-decoding process.

U-Nets (and similar variants on this theme) are now ubiquitous in the field of computer vision. To illustrate their strengths, it is useful to compare them to the alternative of applying a fully convolutional processing with no dimensionality reduction (pooling). In this case, a single pixel in the output predicted segmentation can only be informed by a local region from the input image, the extent of which depends on how many layers the CNN has. The number of layers determines the extent to which information can be locally 'diffused' by the CNN and allowed to interact. This concept is called the 'receptive field' of the network. Due to pooling and up-sampling, U-Nets have a vastly increased receptive field, and it is common to find architectures which can (in theory) take the entire image context into account when classifying a single pixel.

1.5 Contribution

This thesis concerns the development of DL algorithms for the measurement of two types of lung pathology: lung nodules (Chapter 2) and mesothelioma (Chapter 3).

Lung Nodule Detection by Deep Learning: Lung nodules are abnormal growths in the lung, which are often small and difficult for radiologists to identify from CT images. I present the design, implementation and testing of a novel algorithm for lung nodule detection. Though novel as a whole, the algorithm is based on components of existing technologies (e.g. CNNs and the U-Net architecture). The results of testing this algorithm showed that the most challenging nodules to segment were those located at the boundaries of the lungs.

This led to my main technical contribution to lung nodule detection: an invention which enables the utilisation of hierarchical spatial relationships to enhance DL algorithm performance on segmentation and detection tasks. Prior to this, there was no known approach to train DL algorithms to operate symbiotically together based on spatial relationships. This invention was captured as a U.S. Patent [24], and is presented in Section 2.5. Whilst the invention was conceived on the task of lung nodule detection, it is more broadly applicable to the measurement of all structures which abide by hierarchical spatial relationships.

Mesothelioma Measurement by Deep Learning: Mesothelioma, unlike other lung cancers, is not approximately spherical in shape. Rather, the cancer grows like the rind on an orange around the lungs, making a complex shape which is difficult to accurately delineate. To date, only semi-automated approaches have been conceived to assist with the segmentation of mesothelioma, relying on varying amounts of user input (e.g. manually placed seed points, or the manual delineation of neighbouring structures). In Chapter 3, I present the first fully automated segmentation algorithm for mesothelioma. The algorithm is based on a U-Net design, the specifics of which were the result of extensive experimentation on my behalf. This contribution was captured in two publications [25, 26]. Following the design, implementation and multi-fold analysis by myself, the algorithm was independently clinically evaluated by clinical collaborators. These results are presented in Section 3.7, and were published in a clinical journal [27].

Chapter 2

Lung Nodule Detection by Deep Learning

Lung cancer accounts for the majority of cancer related deaths, and the ability to treat the disease depends on how quickly the cancer is diagnosed. Lung nodule screening is a component of routine care. The manual reading process is time consuming, and radiologists benefit from automated tools which provide second-read capabilities. I developed a novel algorithm for lung nodule detection which consisted of two stages, and compared this algorithm with two publicly available high-performing benchmarks. The evaluation was conducted in three parts: an LIDC-IDRI multi-fold analysis, and LIDC-IDRI held-out test set analysis, and an independent NLST analysis. LUNA CPM scores of 0.784, 0.807 and 0.684 were achieved for these analysis respectively. Peak nodule sensitivities ranged from 84% to 91% across these analysis. The novel algorithm was found to be equivalent in performance to the benchmarks, and also competitive with other works cited in the literature. Following this work, I developed an invention whereby hierarchical spatial relationships are leveraged to increase algorithm performance, imposing the constraint that a detected lung nodule must reside within a predicted lung segmentation. The invention involves combining the results of two CNNs for different tasks during training, so that errors for one task may be back-propagated into both CNNs. I demonstrated the utility of this approach by showing a quantitative improvement on the task of lung nodule detection. For lung segmentation, improvements in performance were observed where there were pathological regions within the lung.

2.1 Overview

This chapter describes the development of a novel algorithm for lung nodule detection. An introduction to lung nodules is provided in Section 2.2. Section 2.3 describes the algorithm developed for lung nodule detection, and introduces two independent algorithms (DeepLung) which are used to benchmark the proposed algorithm across multiple datasets. The results of the benchmarking analysis are presented in Section 2.4. Finally, an invention which aims to leverage hierarchical relationships to further enable lung nodule detection is presented in Section 2.5.

2.2 Introduction

Lung nodules are an abnormal growth in the lung, which may be caused by infections or scarring of lung tissue. According to the American Thoratic Society [28], the majority (around 95%) are non-cancerous, and this may be determined by longitudinal assessment of the growth or measurement by tissue biopsy. Many lung nodules may be found incidentally in CT or X-ray images, however dedicated screening programmes may be deployed for at-risk groups. Lung nodules precede lung cancer.

Early detection of lung cancer is extremely important as the outcomes of patients diagnosed with lung cancer depend on cancer stage at diagnosis. Mithoowani *et al.* [29] report that patients identified at stage IV disease have a five-year survival of 5.8%, which increases to 68.4% for those patients identified with stage I disease. In the NELSON trial [30], the impact of screening was compared between two cohorts. Among the unscreened cohort, only 13.5% of disease was identified at stage I. For the screened cohort, this was increased to 59%, and as a result lung-cancer related mortality was reduced by 24% for males, and 33% for females. Elsewhere, reductions in mortality of 8% were found [31].

2.2.1 Lung Nodule Screening

Lung nodules may be small, and dedicated screening is a challenge requiring a radiologist to carefully browse a CT image slice-by-slice. Many lung nodules are bright in appearance, and clearly contrast the dark back-ground of the lung in CT. However, ground glass opacities (GGOs) are lung nodules which are more subtle in appearance, appearing in CT as similar to ground glass, and can be particularly

challenging to identify. The nodules of most clinical concern are those which have a mixed component, which can indicate that the nodule is in a stage of development.

The LIDC-IDRI (Lung Image Database Consortium and Image Database Resource Initiative) reference database [32] provides an excellent resource of 1018 CT cases for the development of CAD software. Each case was annotated by a two-phase (blinded followed by subsequent review) annotation process by four expert radiologists. This database contains 2669 marked lesions larger than 3 mm in diameter, however all four radiologists only agree for 928 (34.7%) for these cases, which indicates a high level of ambiguity in the annotation task, especially amongst the smaller nodule sites.

Currently there are nine FDA approved algorithms for pulmonary nodule detection [12]. Many of these offer a second read capability, whereby candidate sites are flagged for review after the radiologist has completed their initial read of the CT volume. There are significant technical challenges for CAD pipelines for lung nodule detection. Primarily, lung nodules may be difficult to distinguish from pulmonary vasculature. This is also a challenge for manual detection, and often requires the reader to leverage 3-D information by carefully browsing back and forth through axial slices, or by utilising other views. This also is a major limitation of detecting nodules from CT scans where the slice spacing is large — the ambiguity in this distinction may increase with CT slice spacing, and images of thinner slices are now mandatory in many screening programs.

Inevitably, the strategy to distinguish lung nodules from other tissue leverages knowledge about the appearance of such structures. However, additional pathologies (e.g. emphysema) may obscure or confound their appearance. Specifically for the purposes of lung nodule screening this is a challenge, because the screening population is likely to be generally unhealthier than the normal population. For example, smokers, who are at higher risk of developing lung cancer are also at higher risk of developing other conditions such as emphysema.

2.2.2 Datasets and Challenges

To obtain fair comparison between automated approaches, lung nodule detection was one of the first topics to be subjected to a Medical Imaging Grand Challenge with the ANODE09 (Automated Nodule Detection) challenge in 2009 [33]. This

was followed by the larger LUNA16 (Lung Nodule Analysis) challenge in 2016 [34], based on a subset of the LIDC-IDRI dataset. Large datasets for lung screening trials have also been published, leading to a selection of publicly available resources for algorithm development. Here, the data resources leveraged for the presented work are detailed.

The LIDC-IDRI Dataset

The LIDC-IDRI [32] dataset (mentioned in Section 2.2.1) represents a collaboration between seven academic centres and eight medical imaging companies to assemble a dataset of 1018 CT volumes. These volumes have been read by four expert radiologists independently, to mark nodules ≤ 3 mm or other pathological sites with coordinate locations. For the nodules identified as > 3 mm a segmentation of the nodule was performed. In the subsequent annotation review phase, the radiologists were allowed independently review their marks to provide a final opinion. In total, the database contains 7371 nodules. For the large nodules, further annotation was conducted by the readers to assess, via scoring, a number of features of these nodules. Specifically, the readers were asked to assess whether the nodule appeared benign or malignant, and to heuristically grade its appearance in terms of shape and structure.

LUNA Challenge Data

The dataset used for the LUNA challenge [34] is a subset of cases from the LIDC-IDRI dataset. Specifically, 888 CT volumes were selected with a slice thickness ≤ 2.5 mm. The evaluation for the challenge regards sites > 3 mm identified by two observers as a positive case. For evaluation, no penalty was incurred for algorithms which flagged small nodules, other pathological sites, or large nodules with low observer agreement as positives. The challenge scoring is further detailed in Section 2.2.3.

NLST Dataset The National Lung Screening Trial (NLST) [35] aimed to compare the efficacy of lung nodule detection by low-dose CT imaging and X-ray, and found that lung cancer mortality was reduced by 15–20% for those patients who received CT imaging. In total, 53,454 subjects were enrolled from 33 institutions who were deemed high risk for developing lung cancer (based on age and smoking history). These CT images are available by request, and as a component of the work presented here, a selection of these images were annotated over a time-boxed duration of one

month. Specifically, three contractors annotated 419 volumes, where 64 volumes were annotated by two observers, and the remainder by one observer. Annotators were provided with a protocol whereby a NLST findings report was used to identify lung nodules in the images. These reports contain information regarding lobe, laterality, type, size, and smoothness of each finding. Each nodule was segmented in the most convenient plane for the annotator, where the axial plane was recommended but not imposed. The NLST reports were limited to six findings, however for three cases there were more then six nodules present in the images. These were additionally located and segmented by the annotators. For one case, a false positive was found in the NLST findings list and subsequently excluded. A selection of 79 negative cases were also included in this dataset, which required no annotation, resulting in a total dataset of 498 cases with nodules segmented for algorithm evaluation.

2.2.3 Evaluation Metrics

Pehrson *et al.* [36] conducted a systematic review of 41 articles which apply ML and DL to the LIDC-IDRI dataset for the detection of lung nodules. They observe that there is no consensus on the method of determining algorithm performance, which raises challenges for comparison and benchmarking between methods. Generally speaking, the user of a CAD detection tool for lung nodules is likely to be interested in two primary performance metrics: sensitivity and specificity. Sensitivity, or true positive rate (TPR), is defined as

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{2.1}$$

where TP describes the number of true positives, and FN describes the number of false negatives. TPR describes the portion of positives which are correctly identified. Specificity, or the true negative rate (TNR), is described as

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} = 1 - \mathrm{FPR}, \tag{2.2}$$

where TN describes the number of true negative cases, and FP describes the number of false positive cases. This provides a measure of false alarms predicted by the algorithm. For algorithms which predict a probability associated with a candidate lung nodule site, a threshold must be applied to the algorithm output before assessing

these metrics. This threshold selection process represents a trade-off — it is possible to select a low threshold, which may provide high sensitivity at the expense of providing an increased quantity of false positives. Conversely, a high threshold may sacrifice sensitivity to reduce the number of false positive candidate sites. For this reason, the two metrics are commonly displayed together in graphical form as an ROC (Receiver Operating Characteristic) curve, and the area under this curve (AUC) provides a metric which is agnostic to the choice of operating point.

An alternative to measuring the area under the ROC is to analyse the Free-response Receiver Operating Characteristic (FROC). Rather than presenting specificity on the horizontal axis, the number of false positives at a given threshold is shown. This makes for a more interpretable presentation, as the number of false positive sites which are generated to attain a desired sensitivity can be clearly read. To summarise this analysis, several lung nodule detection challenges have utilised a target metric which is constructed at the average sensitivity across a discreet range of false positive rates per scan.

The LUNA Challenge Performance Metric

The LUNA challenge assesses algorithm performance as the average sensitivity at $\frac{1}{8}$, $\frac{1}{4}$, $\frac{1}{2}$, 1, 2, 4, and 8 false positives per scan. This range is clinically derived — CAD tools generally operate between 1 to 4 false positives per scan; some systems allow the user to vary this operating point. In this work, we refer to this metric as the LUNA CPM (Competition Metric). Due to the extensive level of annotation available on the LUNA challenge dataset, their evaluation protocol only regards candidates larger than 3 mm which have been identified by at least two observers as positive cases. The remaining candidates (including smaller nodules and other pathological structures) are excluded from the evaluation. Specifically, they are not considered false positives if they are detected by the algorithm.

2.2.4 Literature Survey and Existing Tools

Lung nodules are often a very small portion of the input images (if present at all). This means there is an extreme spatial class imbalance — non-nodule sites greatly exceed nodule sites in the CT images on a per-voxel basis. This presents challenges for algorithm development. The vast majority of the time, it is accurate to predict

non-nodule, and unless the class imbalance is addressed it is often the case that deep learning based algorithms converge to a local optimum of predicting all candidate sites as nodule negative.

A selection of results from the LUNA16 challenge[1] are summarised in Table 2.1, where LUNA CPM scores range from 0.608 to 0.951. Broadly speaking, these results show that the most powerful approaches are CNNs. Many of the presented approaches are two-stage in design, comprising candidate proposal and subsequent prediction refinement. However, some caution should be used when regarding these scores. LUNA scores are computed in a multi-fold fashion, and it is up to the competitors to ensure they are using the data appropriately. For example, we have observed that competitors can use the validation set to perform best model selection and other hyper-parameter selection, ultimately degrading the repressiveness of the analysis and inflating performance metrics.

[1] https://luna16.grand-challenge.org/Results/

Algorithm Name	**Candidate generation**	**False positive reduction**	**CPM**
PAtech	CNN	CNN	0.951
JianPeiCAD	Multi-scale rule-based screening	CNN	0.950
LUNA16FONOVACAD	CNN	CNN	0.947
Aidence	CNN (end-to-end)		0.871
ZNET	CNN	CNN	0.811
MOT_M5Lv1	Thresholding and morphological operations; automatically selected hand-crafted features	Hand-crafted features and NN classification	0.742
ETROCAD	Multi-scale nodule and vessel enhancement filters; nodule type classification by filtering	Hand-crafted features and SVM classification	0.676
M5LCAD LungCAM	High-intensity structure segmentation using ant colony optimisation and iterative thresholding	Hand-crafted features and NN classification	0.608

Table 2.1: *A selection of results from the LUNA challenge leaderboard. Several competitors with-hold an algorithm description, and this summary of results is not intended to be comprehensive.*

Whilst being a popular approach, DL methods do have their limitations in the context of lung nodule detection. For example, Sourlous *et al.* [37] discuss how different forms of bias may impact algorithm performance. Factors such as patient race can easily be predicted by DL methods based on imaging, and this information may be leveraged by an algorithm where instances of lung cancer have different prevalence among different racial groups. This provides an example of how underlying algorithm bias may be difficult to control for.

Since the LUNA16 challenge several alternate approaches have been published. Nasrullah *et al.* [38] use 3-D CNNs on the task of lung nodule detection using the LIDC-IDRI dataset. The use a MixNet architecture, which contains an encoder-decoder structure with skip connections which use both concatenation and summation. To reduce false positives, they include a second stage processing using a gradient boosted machine (GBM) based on the extracted features and clinical bio-markers. which provides a final classification. They report results as peak sensitivity over the range of false positive rates defined in the LUNA challenge metric, and show sensitivities reaching 94% at an FPR of 4 nodules per scan.

2.3 Methods

2.3.1 In-house Two Stage Algorithm

I present a two-stage algorithm. The first stage generates candidate nodule sites with high sensitivity, and the second stage further evaluates these candidate sites by incorporating additional 3-D information.

Nodule Detector

The first stage of the algorithm is optimised to produce a segmentation of lung nodules, and combines a 2-D U-Net with a VGG-Net [39] style encoder. The encoder is initialised using pre-trained weights from the ImageNet challenge [40], whilst the decoder is trained from a random initialisation. CT slices are rescaled to a size of 320 pixels squared. The three colour channels of VGG-Net are used to input three adjacent CT slices, to provide 3-D context. The three consecutive CT slices are resampled to a resolution of $1 \times 1 \times 2$ mm. The first stage algorithm is shown schematically in Figure 2.1.

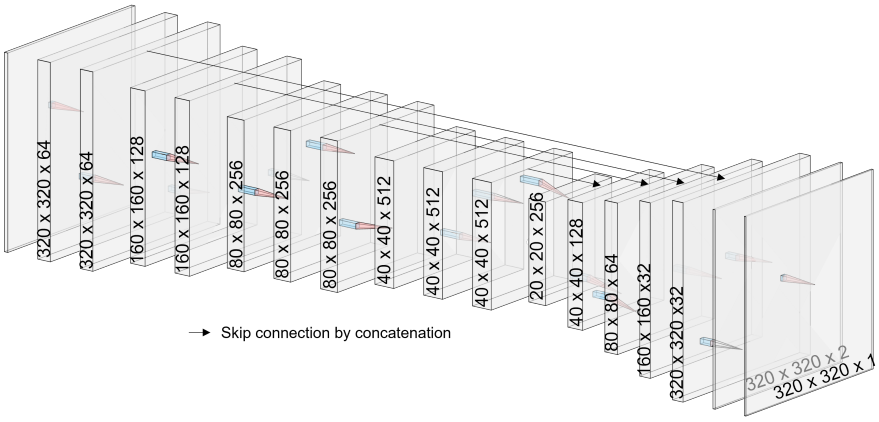


Figure 2.1: *A schematic diagram of the first stage algorithm, composed of a ResNet style encoder, and a custom decoder. Skip connections are shown as black arrows.*

Pooling is performed by the maximum pooling operation, and after each block of processing there is a Dropout [41] and batch normalisation [42] processing (not explicitly shown in Figure 2.1). To address the problem of class imbalance, the focal

loss objective function presented by Lin *et al.* [43] is used, defined by the equation:

$$FL(p_t) = -\alpha(1 - p_t)^{\gamma} \log(p_t), \tag{2.3}$$

where p_t is defined as

$$p_t = \begin{cases} p & \text{where } y = 1 \\ 1 - p & \text{elsewhere.} \end{cases} \tag{2.4}$$

Here, $y = 1$ describes the case where the ground truth label is positive. There are two hyperparameters associated with focal loss, α and γ, which are set to the default values of 0.25 and 2 respectively. Focal loss may be considered as a departure from the standard categorical cross-entropy loss, which has the deficiency in the case of class imbalance that small loss contributions for the majority class may overwhelm significant errors from the minority class. In other words, it is not worthwhile learning how to identify the minority class if this degrades performance for the majority class. In practice, the use of categorical cross-entropy for tasks with extreme class imbalance may result in classifiers which only predict the majority class. Focal loss addresses this issue by down-weighting loss contributions from examples which are already well classified, and focusing on those cases which are incorrectly classified. This method is similar, in a general sense, to the concept of curriculum learning [44], where the strategy throughout training is to gradually expose the algorithm to more difficult cases.

I chose the binary cross-entropy metric to select the best model based on the internal validation data set results. A different function was chosen to guide this process to reduce the effects of over-fitting to a single optimisation metric. For this stage, the algorithm is designed to detect anything which is nodule-like in appearance with high sensitivity, and so sites which are only identified by one radiologist in the LIDC-IDRI data is used.

The first stage model is limited in terms of nodule specificity. The input images of 3 consecutive axial slices is often not a sufficient level of 3-D context to differentiate true positives from false positives. Thus, the requirement to refine the prediction based on a more extensive level of 3-D context underpins the design of the stage 2 algorithm. Typically, for a single volume, I set the threshold so that around 100 candidate sites are proposed to the stage 2 algorithm.

Nodule Classifier

The second stage algorithm takes the candidate sites from stage one and further classifies them as nodule or non-nodule. At this stage the definition of a nodule in the ground truth is based on a consensus among at least two radiologists in the LIDC-IDRI data, in-line with the LUNA challenge evaluation protocol. Thus, a much higher specificity is encouraged in the training with respect to stage one. The network takes in $48 \times 48 \times 48$ voxel 3D blocks (at a resolution of 1 mm^3), each centred on a candidate, and classifies it as nodule or non-nodule. At training time these candidate sites are based on the output of the first stage model on the validation cases (those cases not used in first stage training). Due to the low specificity of the first-stage model with respect to the second stage ground truth (which is due to the difference in consensus level between stage one and stage two training data), there remains a large class imbalance during second stage training. To lessen this imbalance, the first stage model is also used to extract candidate sites on the data with which it was trained. This additional extraction step contributes mostly positives to the second stage training data, to better capture the variation in the ascetic characteristics of true positive nodules in this set. Focal loss is used to address class imbalance, and categorical cross-entropy is used as a validation metric for early stopping of the training. During training, the class balance is set to a ratio of 20 negatives for every positive candidate region. To measure algorithm performance on the internal validation set, the ratio is set to an equal number of positive and negative candidate sites.

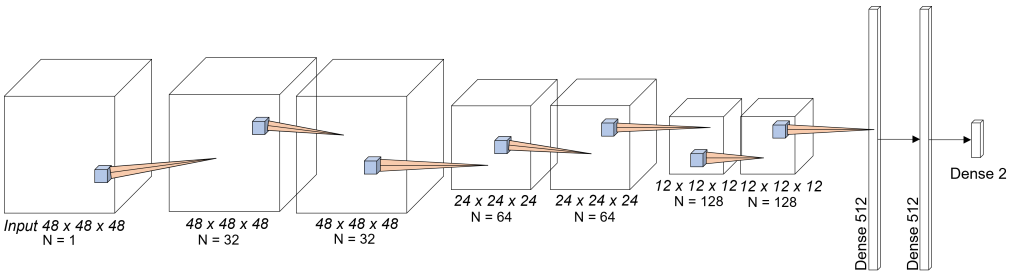


Figure 2.2: *A schematic diagram of the second stage algorithm. The input volume is shown to the left, which undergoes six layers of convolutional processing, followed by a maximum pooling operation, and two layers of dense processing to provide the output prediction.*

The architecture of the second stage network is illustrated in Figure 2.2 with three convolution then-max-pooling blocks, followed by three dense layers and a

softmax to assign final probabilities of a candidate region containing a nodule. After each block of processing, Dropout [41] and batch normalisation [42] is performed.

2.3.2 Benchmark: DeepLung

In order to benchmark our novel approach, the algorithm DeepLung was used [45]. This algorithm was first published in 2017, where it achieved state-of-the-art performance on the task of lung nodule detection. As for the algorithm proposed in Section 2.3.1, the authors propose an algorithm of two stages — the first for nodule detection and the second for nodule classification. Their contribution is the application of 3-D Dual Path Network (DPN) derived features for the second task of lung nodule classification. To benchmark their contribution, they use the 3-D equivalent of the highly successful ResNet-18 architecture to generate features for comparison. Critically, their method contains a quarter of the number of free parameters as the ResNet-18 approach [46]. For completeness, we include both the DPN and ResNet-18 based approaches as our benchmarks, which were available online [2]. For both methods, the first stage detection process is common, and comprises the application of an 3-D Faster R-CNN.

Nodule Detector

The Faster Region CNN (R-CNN) Detector [47] operates at multiple input scales to produce bounding boxes and predicted probabilities for the contained object. Fundamentally, they are an encoder-decoder architecture for the task of bounding box regression. This is the problem of producing a 5-vector output for (X, Y, width, height, probability) for each scale and pixel, for the original formulation dealing with 2-D images. For DeepLung, this formulation is expanded to 3-D. The model predicts a vector (X, Y, Z, diameter, probability) at each voxel, a formulation which encloses the assumption of approximately spherical nodules. Crucially, this light weight implementation outputs predictions at a resolution of only a quarter of the input image resolution. For processing, images patches are generated at a resolution of $96 \times 96 \times 96$. A region of $32 \times 32 \times 32$ voxels is extracted around the proposed candidate sites, which are passed to a second stage for classification.

The authors define a custom loss function for their approach, where regression is

[2] https://github.com/uci-cbcl/DeepLung

performed separately for the four spatial elements of the 5-vector output. For the predicted probability, to address the extreme class imbalance, they calculate binary cross-entropy for the positive and negative candidate sites separately. This allows them to to filter the loss contribution from negative sites, and they only evaluate those predictions which are strongly incorrectly classified. Because the loss includes contributions from all the correctly identified cases, and only the most strongly incorrectly identified cases, the contributions to loss become more class-balanced. Functionally, this approach is highly similar to Focal Loss. The best model is selected from across the epochs of training based on the value of their loss function on the monitoring subset of cases.

Nodule Classifier

The authours of DeepLung present two algorithms for subsequent classification: DPN and ResNet-18 based classifiers.

The Dual Path Network Classifier is based on the works of Chen *et al.* [48], and is a nodule classification network. It acts to encode the input image patches by sequential processing by DPN blocks and 3-D average pooling. DPN blocks have the advantage that they encourage new feature exploration whilst enabling feature re-usage throughout training, by employing parallel paths of residual and dense processing. At the penultimate layer, a vector of size 2560 is output. For training the DPN, this is classified by a fully dense layer into benign versus malignant — this is a finer grained task than nodule versus non-nodule, and encourages the learning of finer-grained features. At test time however, the 2560 long feature vector is passed to a Gradient Boosted Machine (GBM) along with other nodule statistics such as diameter which have been extracted by the R-CNN to generate a final nodule classification.

The ResNet-18 Classifier is a 3-D modification of that traditional ResNet-18 algorithm [46], modified to take the input blocks of $32 \times 32 \times 32$, and reduce these blocks to a feature vector which is further classified by a GBM. This model is trained in a consistent manner to the DPN network — in the task of benign versus malignant classification.

2.3.3 Experimental Design

In order to conduct a fair comparison between the three methods, the same data was used to train, validate and test the algorithms. The full LIDC-IDRI dataset was used to develop and validate the algorithms, which includes a selection of CT images with slice spacing that exceeds 2.5 mm. This expansion over the LUNA challenge data complicates benchmarking to other published works, however we determined that a deployed CAD tool should be capable of flagging candidate sites in CT images with a higher slice spacing.

The dataset divisions are shown in Figure 2.3, where a test set of 110 cases was partitioned. This dataset, referred to as the LIDC-IDRI test set, and was set aside and used only once when the algorithm work had concluded to evaluate performance. During development, 15-fold cross-validation was applied. This is a change to the LUNA challenge approach, which reflects the increase in the data set size. For each fold of analysis, the algorithms were fit using 672 cases, with 168 cases used for internal validation to assess over-fitting during training.

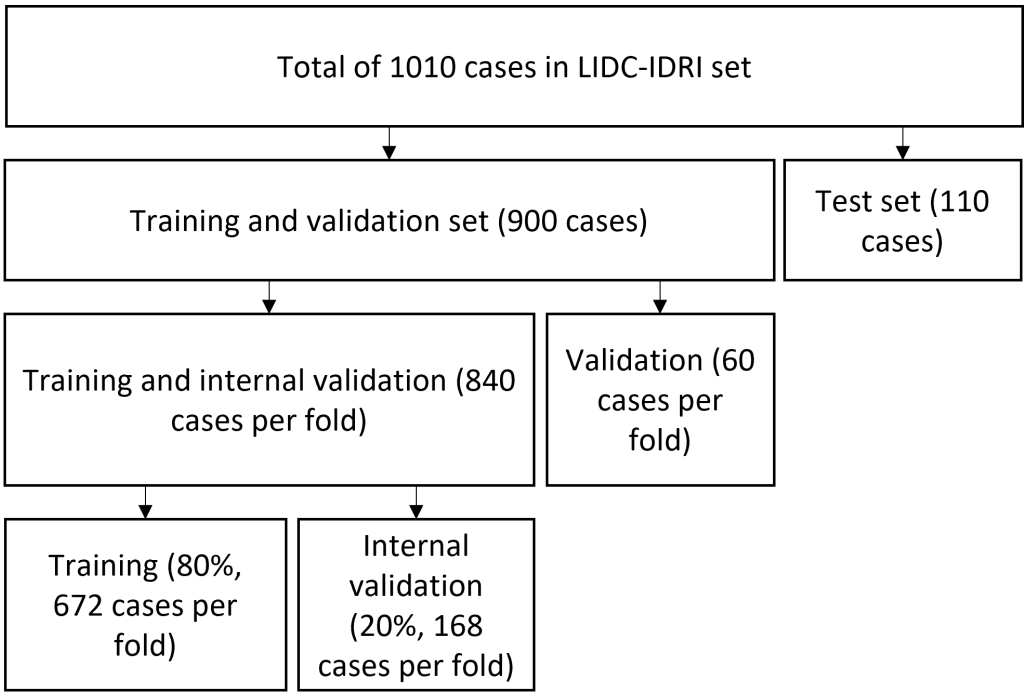


Figure 2.3: *To enable multi-fold analysis and final testing, the LIDC-IDRI dataset of 1010 cases is split into subsets: the test set, for testing the final models once; the training set, for fitting the model parameters; the internal validation set, to guide early stopping; and the validation set, for assessing performance in a multi-fold fashion.*

For the LIDC-IDRI datasets, the LUNA CPM was used as a measure of algorithm

performance. This evaluation includes the incorporation of the small and non-nodule sites as excluded from evaluation. Multi-fold analysis has its limitations, and while providing a somewhat realistic estimate of algorithm performance, a held-out set is a more reliable measure. This is because over-fitting can still occur when repeated testing guides the algorithm development process. The LIDC-IDRI test set is a useful method of testing, however this data is still likely to be similar in many aspects to the data on which the algorithms were trained. To address this, the NLST dataset described in Section 2.2.2 is used as an independent test set. Crucially, this data has no record of other pathology present in the images, and no record of small nodule findings, so such structures cannot be excluded from evaluation.

2.4 Results

The results of benchmarking our novel algorithm are divided into three subsections: Cross-Validation Results, where the scores from 15-fold cross validation is presented; LIDC-IDRI Test Results, where the results for the held-out 110 test cases are presented; and the NLST Test Results, where a set of totally independent data is used to test the algorithm.

2.4.1 Cross-Validation Results

The 15-fold cross-validation results are shown in Figure 2.4. They show that the results are very similar, with overlapping confidence intervals attained by bootstrapping. The In-house algorithm provides a Luna CPM of 0.784, which is higher than achieved by both the benchmark methods.

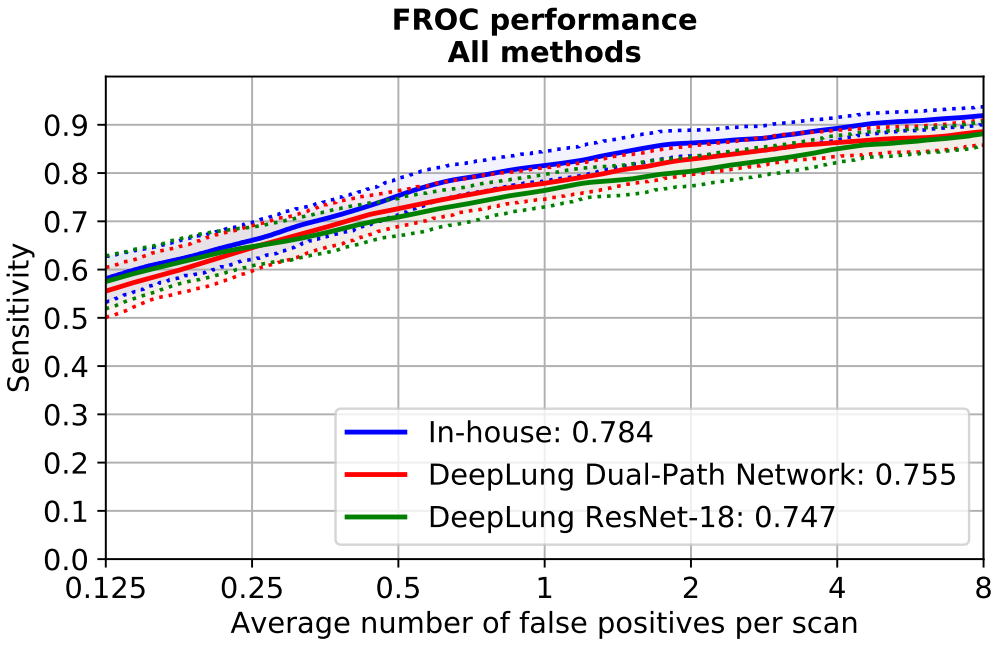


Figure 2.4: *The Free Receiver Operating Characteristic (FROC) curves for multi-fold cross validation over 900 cases, where thin dotted lines represent errors measured by bootstrapping.*

Figure 2.5 shows an example of a lung nodule detection by the three methods. The visualisation is for illustrative purposes only, and is tailored to show the bounding spheres and coordinates obtained by the DeepLung methods. Specific processing would be required on the segmentation masks provided by the In-house method to generate comparable spheres, so the predicted coordinate is shown. This coordinate

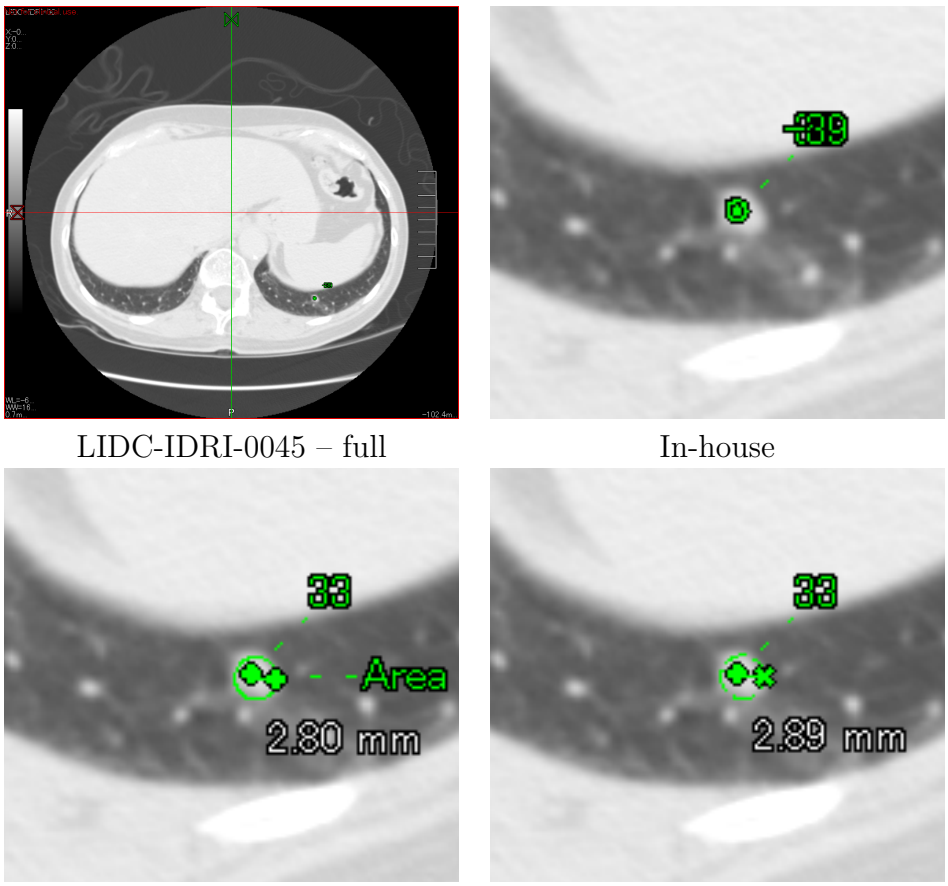


Figure 2.5: *A nodule detected in the left lung of LIDC-IDRI dataset 0045. An axial slice is shown, with the corresponding detection by the In-house, DeepLung DPN and DeepLung ResNet-18 approaches. Notably, the In-house method is not designed to predict bounding spheres, and only the predicted coordinate is shown for comparison.*

represents the maximum of the segmentation mask in the region which is above the detection threshold being evaluated.

2.4.2 Hold-out Validation Results

The LIDC-IDRI hold-out set results are shown in Figure 2.6. Here, higher values for the Luna CPM score are seen across the board, and the results remain highly similar. It is likely that this test set contains generally easier samples than are represented in the larger training sets. Crucially, each line represents the average performance of the 15-fold models for each method, and here the error bars show the best and worst models from across the folds of analysis.

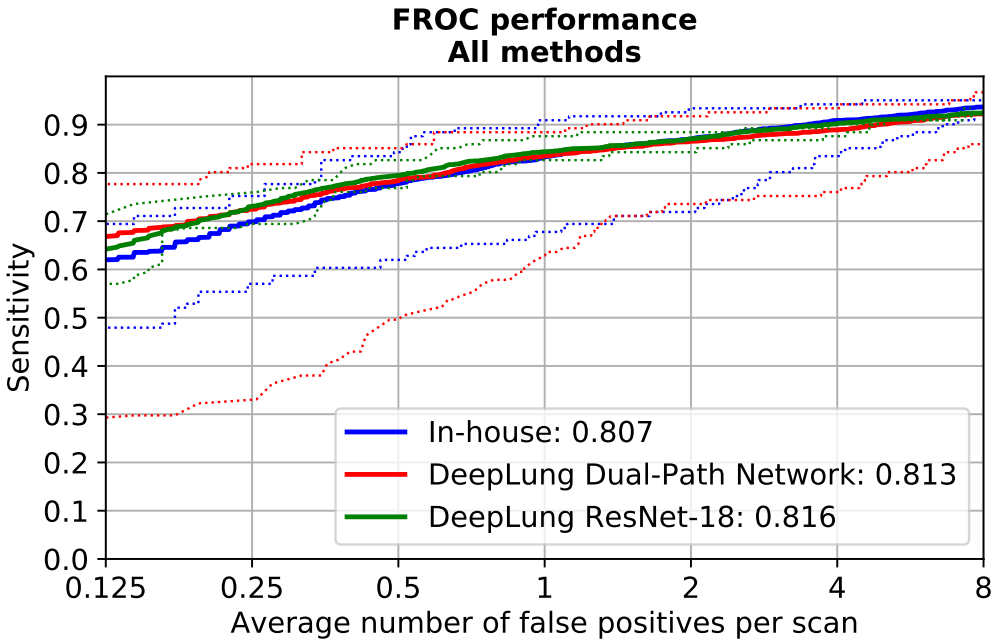


Figure 2.6: *The Free Receiver Operating Characteristic (FROC) curves for the held-out LIDC-IDRI test set of 110 cases. The thick solid lines represent the mean performance, while the thin dotted lines represent the individual models for each approach with highest and lowest LUNA16 CPM.*

2.4.3 External Validation Results

Results for the independent test set are shown in Figure 2.7. Across the board, lower Luna CPM scores are achieved. Here, The DeepLung ResNet-18 approach performs best, with a LUNA CPM score of 0.684. Peak sensitivities are also lower, where both the ResNet-18 and In-house approaches achieve 84%. Here, the DPN approach plateaus beyond an FPR of 4 false positives per scan. The algorithm employs an internal thresholding of candidate sites to evaluate by the DPN which was tuned on the LIDC-IDRI dataset. This is to say, the number of candidate sites presented is limited to only those identified by the R-CNN with high confidence. For the NLST dataset, this threshold appears to be inappropriate, and limited the number of candidate sites, and thus the maximum sensitivity achieved by the second stage.

2.4.4 Summary

The results from the different analysis are summarised in Table 2.2. It became apparent throughout this work that the most challenging nodules to detect were those which were close to the lung wall, and bordering other bright structures. This observation resulted in a body of work which aims specifically to address the technical

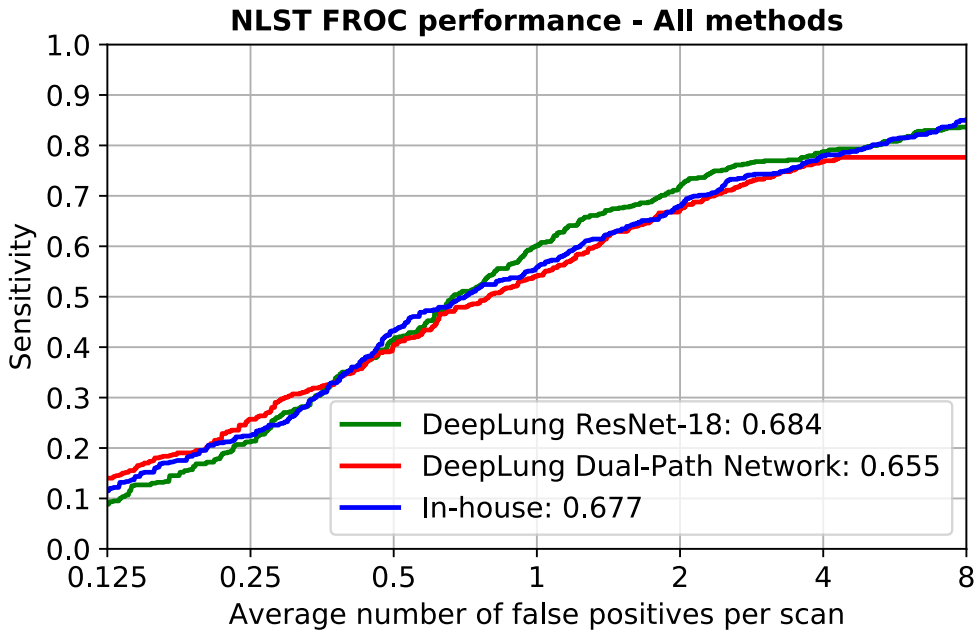


Figure 2.7: *The Free Receiver Operating Characteristic (FROC) curves for the independent NLST test set of 498 cases.*

challenges of automatically segmenting such cases using DL methods.

		LUNA16 CPM	
model	Cross-validation	LIDC-IDRI Test Set	NLST Test Set
In-house	**0.784**	0.807	0.677
DeepLung DPN	0.755	0.813	0.655
DeepLung ResNet-18	0.747	**0.816**	**0.684**

Table 2.2: *LUNA16 Challenge Performance Metric (CPM) for the three approaches on the validation, testing and independent testing datasets.*

2.5 Hierarchical Multi-task Transfer

In this section, an approach for leveraging hierarchical relationships to improve segmentation accuracy is presented, which was published in the form of a U.S. patent [24]. I developed the approach following those presented in Section 2.3 to improve the detection performance of the first stage detector algorithm described in Section 2.3.1.

2.5.1 Introduction

Given input data, it is often necessary to perform more multiple analysis. For example, it may be that automatically derived lung nodule detection and lung segmentation would be useful to inform patient diagnosis when dealing with a chest CT image. This pair of tasks exhibits a spatially hierarchical relationship — a lung nodule must reside within the lungs — as generally illustrated by Figure 2.8.

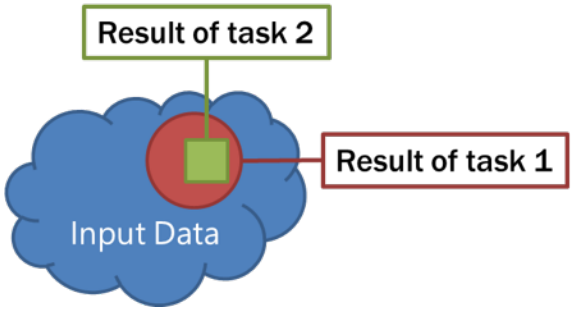


Figure 2.8: *An illustration of hierarchical data, where the result of task 1 resides within the result of task 2.*

These tasks may be performed independently by two separate deep learning models, or jointly, by a single model with multiple outputs (as described by Figure 2.9). The joint performance of multiple tasks by a single model is often referred to as 'multi-task' learning. Sometimes, modelling tasks together offers an improvement in performance. Whilst there is utility in performing related tasks together, it is rare for deep learning approaches to capture the relationships in the data which are most apparent to the human observer. Such relationships may be that certain objects within the images are always adjacent, or that when one object is present another is absent. For lung nodule detection and lung segmentation, one such relationship is that a lung nodule always resides within the lungs.

Though low-level features (such as certain textures or edges) may be applicable to multiple tasks in a synergistic manner, the design of current deep learning

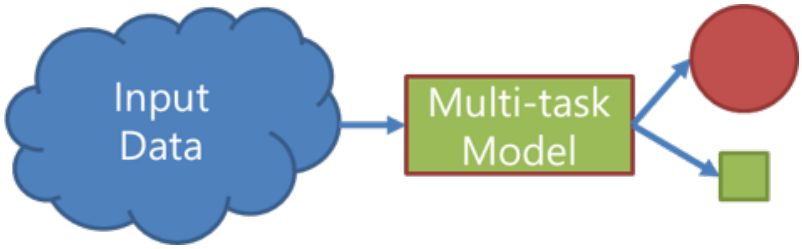


Figure 2.9: *An illustration of a multi-task model based on the data shown in Figure 2.8.*

approaches are not prone to capturing certain high-level relationships. It is not that modern algorithms lack the capability to capture these relationships, but rather these relationships may not be the most effective ones to learn in order to reduce the objective loss function given the training data. For example, an obvious error to a human observer by an automated tool would be the detection of a lung nodule in the intestines, as imaged by CT. Some regions within the intestines can sometimes appear nodular in shape, with a complex and variable structure. These were the most obvious false positive cases by the approach described in Section 2.3. If the algorithm is operating based on mostly low-level features (edges, textures and intensity), it may be that a reduction in the predicted probability of this intestinal structure comes at the expense of detecting true nodules elsewhere.

The following approach was developed to explicitly impose hierarchical spatial relationships between deep learning models, to remove the reliance to learn such relationships internally from the data. Specifically in this invention, the hierarchical spatial relationship between lung nodules and the lungs. The method:

1. guarantees congruence between lung segmentation and lung nodule detection results,

2. improves lung segmentation for regions where bright pathology is present,

3. does not require data which is annotated for both tasks,

4. allows the lung nodule detector to become more sensitive by confining it's operation to the region of the lungs.

A number of existing works consider the operation of deep learning algorithms in the context of class hierarchies.

2.5.2 Existing Works

Fu *et al.* [49] propose a "course-to-fine" network layer in the context of image classification. The layer is inspired by the Bayesian equation, and multiplies the predicted probability of a sub-class by a parent class during both algorithm training and inference. For this reason, the authors require a single dataset, which has multiple levels of annotation, because they train a single model which predicts at multiple levels of annotation.

Yan *et al.* [50] also present an approach which leverages class hierarchies. Rather than multiplying outputs, they combine class and subclass predictions generated from different depths of the classification network.

Both of these inventions are related to image classification. The contributions consider categories and sub-categories of classification e.g. a parent class of dog, with a subclass of Labrador, based on the insight that many subclasses are harder to distinguish than classes. Our contribution is primarily concerned with spatial hierarchies, and spatially limiting/extending the extent of predictions in pixel/voxel level image classification tasks.

Crucially, both approaches depend on a single fully annotated dataset. I present a method which uses multiple models and alternated training, and can utilise independently labelled datasets, removing competition between the tasks. For the task of classifying entire images, grouping of labels is relatively cheap. For segmentation tasks there is limited availability of comprehensively annotated datasets, especially in the medical domain.

2.5.3 Method

In order to impose a hierarchical spatial relationship between detected objects, I propose joining individual models by a multiplication of their segmentation outputs during training. This is combined with an alternated training process during training. For the tasks of lung segmentation and lung nodule segmentation, the stages of training are as follows:

1. the lung segmentation model is trained individually to segment the lungs,

2. the lung and lung nodule segmentation models are joined by multiplication and trained on the task of nodule segmentation.

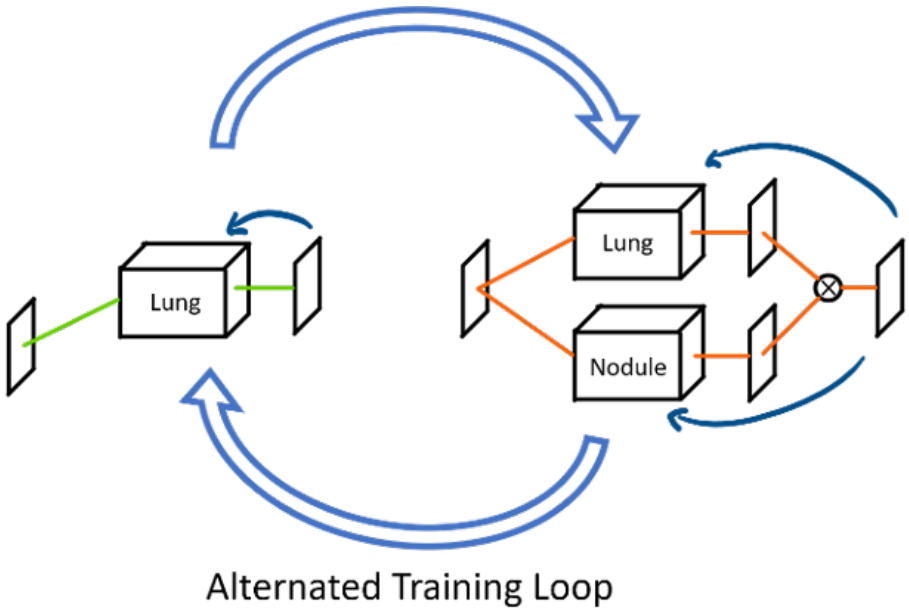


Figure 2.10: *An illustration of the combined model (right) and alternated training strategy. Errors when training on the lung segmentation task can be back-propagated into the lung model.*

This operation may be considered as calculation of the joint probability of nodule and lung:

$$P(lung \cap nodule) = P(lung)P(nodule|lung), \tag{2.5}$$

Where $P(lung \cap nodule)$ is the probability of both lung and lung nodule occurring. $P(lung)$ is the output of the lung segmentation model, and $P(nodule|lung)$ is the output of the nodule segmentation model. This results in a lung nodule segmentation model which is only applied within the bounds of the predicted lung segmentation.

This process is shown schematically in Figure 2.10. The alternated training allows errors calculated on the task of lung nodule segmentation to be back propagated into the lung segmentation model. This means that the lung segmentation is explicitly optimised to include small, high H.U. value regions which may otherwise be excluded from the predicted segmentation. Furthermore, the lung nodule segmentation can become more sensitive, because it does not need to distinguish regions which are nodular in appearance as being with or outwith the lung.

The de-coupled nature of the models allows separate datasets with either lung segmentation or lung nodule segmentation to be used to train the lung and joined models individually. There is no requirement for data to be annotated for all target classes.

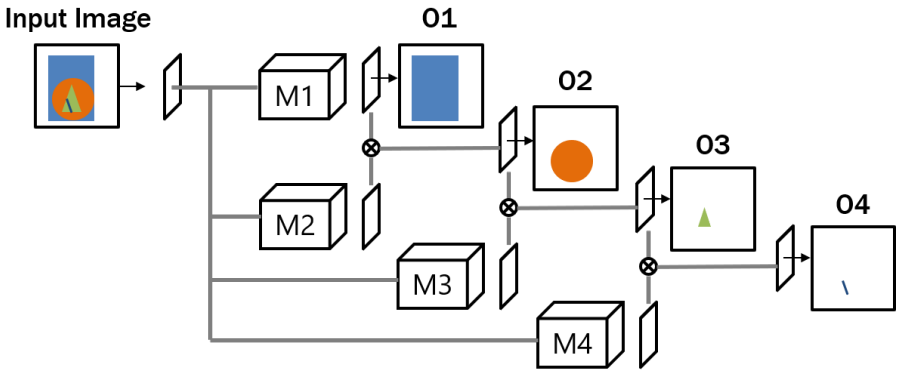


Figure 2.11: *four models (M1 to M4) cascaded at inference time to generate four outputs (O1 to O4) on example hierarchical segmentation tasks. The cyclical alternated training process is not described here.*

Though not shown in detail here, the proposed method may be applied to an arbitrary number of hierarchical segmentation tasks (shown Figure 2.11).

2.5.4 Results

Using the approach described for the first stage detector described in Section 2.3.1, performance improved slightly from a sensitivity of 87.5% to 90.1% at an FPR of 50 per scan. The improvement to lung segmentation was also assessed. Figures 2.12 and 2.13 show an example of a region of lung segmentation where a nodule is present, and example results whereby alternated training forces the lung segmentation to include this previously excluded pathological region.

For the case shown in Figure 2.13, other regions of bright pathology, especially in the thin regions of the lung are also included in the lung segmentation, which were previously excluded. This result also clearly shows a nodule segmentation result which is not confined to the lung region.

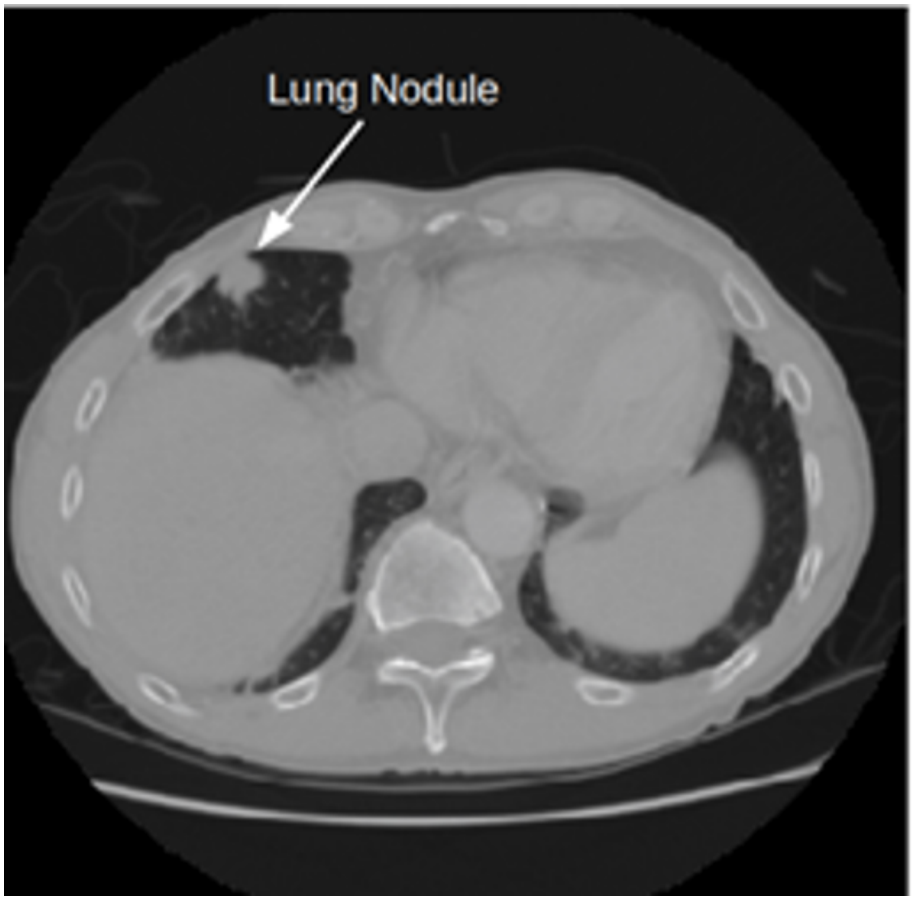


Figure 2.12: *A lung nodule which found in the LIDC-IDRI dataset which is close in proximity to the lung wall.*

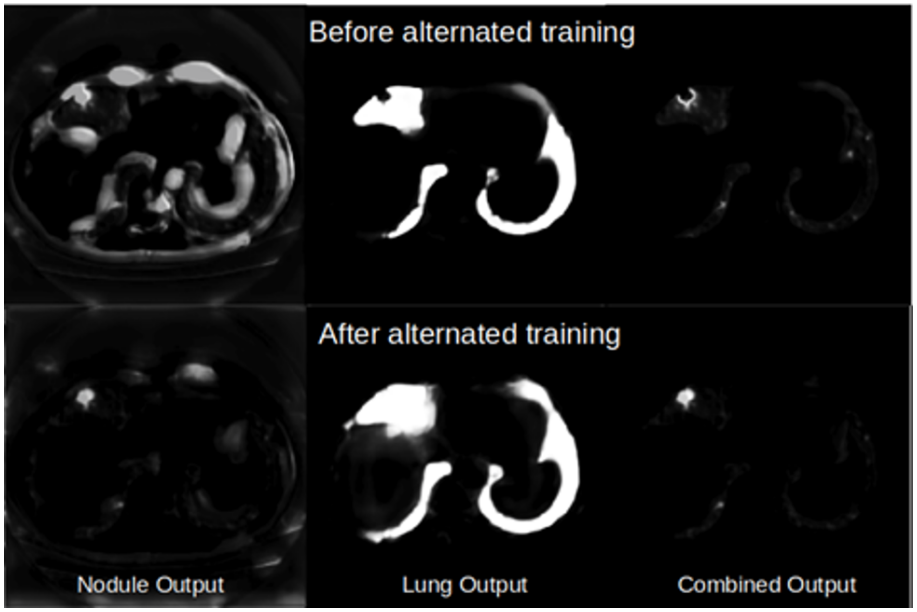


Figure 2.13: *The output of the model(s) before (top) and after (bottom) alternated training, shown for a single CT slice (Figure 2.12). The left panel shows the output of the nodule segmentation model only. The central panel shows the lung segmentation output, and the right panel shows the combined output. As a result of the described approach, the lung segmentation sub-model now includes the lung nodule within the area of the lung.*

2.6 Discussion

With comparison to the paper by Zhu *et al.* [45], we see a lower score achieved by the DPN of 0.755 compared with the 0.842 presented in their paper for the LUNA dataset. There are a number of differences in our experiments which may cause this discrepancy. Namely, we use a different number of folds for analysis (increased to 15 from 10), however this should not degrade performance. We also include a number of additional cases in this analysis with larger slice spacing, which may contribute to the discrepancy. The open-source implementation of their method reported results across 10-folds of analysis, however the data which was used to guide early stopping was also used to report the result. To make comparison fair, this approach was changed so that a separate set was used for early stopping, in an identical manner to the In-house approach. This is most likely to be the source of the discrepancy, and it is probable that the self-reported score in the paper is inflated due to over-fitting.

All of the approaches yield highly similar results, and results which are dependent on the data used to test the algorithms. Across 15-folds of analysis, the In-house method is most performant by the Luna CPM. However, on the two test sets, we see that the DeepLung ResNet-18 approach performs best. Still, the results are highly similar, with overlapping confidence intervals, and it is difficult to draw meaningful conclusions.

Though a direct comparison is not possible due to differing data, all of the algorithms presented here appear competitive with the literature. For example, Gu *et al.* [51] achieve a LUNA CPM of 0.797 on the LUNA challenge data, which is lower than any of the averages we attain across the same data but falls within the uncertainty we find through the multi-fold analysis.

With comparison to the LUNA challenge results shown in 2.1, we achieve a lower than many of the submissions. The differing data may contribute to this, and the inclusion of images with larger slice spacing are likely to be more difficult cases to analyse. Another factor is the multi-fold nature of the challenge. As we observed when deploying DeepLung, data hygiene was a factor which inflated the reported score of this algorithm. Once the methodology was changed to provide a more realistic analysis of performance, all of the methods presented here performed similarly. We suspect that those submissions which score extremely highly by the LUNA CPM would be unlikely to show the same performance on independent data.

The potential of over-fitting complicates comparison between the LUNA approaches, and represents a limitation of the challenge design.

Both Gong *et al.* [52] and Wang *et al.* [53] only report a single operating level. Gong *et al.* report a sensitivity of 79.3% at an FPR of 4 per scan using the LUNA dataset and 10-fold cross-validation, which is nearly 10% below the level of our models at that FPR on that data. This, however, is a highly similar sensitivity which is achieved on the NLST data at that FPR. Wang *et al.* report 95.8% sensitivity at an FPR of 2 on the LUNA data by 10-fold cross-validation, which is nearly 10% above our performance at that FPR.

We note that for the NLST evaluation, results achieved by the DeepLung DPN were hampered by the sub-optimal setting of the internal threshold. It would be possible to adjust this threshold and improve the score, however we consider this threshold as a hyperparameter which was fit to the LIDC-IDRI data. Interfering with the method to improve the score by adjusting such parameters would diminish the value of independently testing the methods.

It is highly likely that publicly available challenge data-sets cannot be easily used to ascertain real world performance due to the body of research and magnitude of solutions which are deployed on this data. While at first inspection the dataset appears large, with CT volumes from over 1000 subjects, these images only contain 2669 nodules in total, and it is likely that the highest scoring methods on this dataset are overfit (to a greater or lesser extent) to the qualities of these cases. This was apparent in our own work when applying a selection of algorithms developed on the LIDC-IDRI challenge dataset to an independent set. All methods performed significantly lower on the independent NLST test set.

Whilst over-fitting may be one source of this discrepancy, the NLST dataset may be more challenging for a number of reasons. One reason may be the inclusion of volumes where there are no lung nodules present. Sensitivity is measured across the entire dataset, and the inclusion of cases with no nodules present can only act to increase the average number of false positives per scan. The other source of discrepancy may be due to the exclusion of small nodules and pathological structures from the LIDC-IDRI evaluation scores. This was not possible for the NLST data, and shows that caution must be used when using competition metrics and data to assess real world algorithm performance.

We showed how task hierarchies could be leveraged to improve the performance

of lung nodule detection algorithms. The results we provide show benefit to both lung nodule detection and lung segmentation. Whilst highly encouraging, further experimentation would be necessary to determine the benefit to the pipeline as a whole (as only the first stage detector was augmented with the novel approach). For numerically evaluating the performance on lung segmentation, results will depend on the cohort captured in the data set. Segmentation of healthy lungs is not a difficult process, due to the large difference in H.U. units at the boundary of the lung high performance can be achieved by traditional approaches (based on thresholding and morphology). Lungs which contain pathological regions are much more difficult to accurately delineate, and this is where the presented approach would provide most benefit.

Chapter 3

Mesothelioma Measurement by Deep Learning

Mesothelioma is an aggressive cancer with a poor prognosis. Unlike lung nodules, which are approximately spherical in shape, mesothelioma grows around the lungs like the rind of an orange. Disease measurement is highly challenging due to the irregular shape of the tumour, and it is difficult to routinely assess how a patient is responding to treatment. Previous works have attempted to semi-automate the process of mesothelioma measurement. Here, I present the first algorithm to fully-automate the process of MPM measurement based on CT images. In contrast to the work presented in Chapter 2 which involved detection of small (and often subtle) pathological regions, MPM measurement involves accurately delineating tumour from other (visually apparent) pathological regions, and as such the presented algorithm is of a single U-Net design. Multi-fold analysis across 123 CT images containing MPM showed a mean Dice coefficient of 0.64 for segmentation accuracy and volumetric measurements which were not significantly different from zero, with a 95% LOAs between -417 and +363 cm^3. The algorithm performance on a multi-centre test set of 120 CT volumes from 60 patients showed the algorithm bias for volumetric measurements remained low (+31 cm^3), and 95% LOAs of -345 to +407 cm^3 were achieved. Inter-observer analysis on a subset of cases demonstrated that the segmentation accuracy of the algorithm is within the agreement between observers. This work comprises the largest study of this type to date in the context of automated mesothelioma analysis, and to our knowledge, provides the first fully-automated method of volumetric measurement of MPM tumours. In future work, the algorithm

will be developed further towards clinical deployment, to enhance the ability of clinicians to routinely and accurately characterise mesothelioma cases.

3.1 Overview

Section 3.2 provides the reader with an introduction to the mesothelioma disease, treatments and measurement methods.

3.2 Introduction

Malignant Pleural Mesothelioma (MPM) is a cancer associated with exposure to asbestos fibres. The outlook for patients with this disease is poor and care is often palliative. The irregular shape and thin nature of the tumour makes response to treatment difficult to measure, impacting both the outcome for individual patients and the development of new treatments. Whilst mesothelioma can develop in the testes, the lining of the abdomen and the lining of the heart, the vast majority of cases occur in the lining of the lungs where asbestos fibers become lodged after inhalation, and this is called pleural mesothelioma.

3.2.1 Asbestos and Disease Prevalence

Asbestos is a naturally occurring natural mineral (shown in Figure 3.1) formed of microscopic silicate fibres. The material has been widely applied industrially due to several useful properties including low heat, sound and electrical conductivity, water and fire resistance. It is also a cheap and strong material.

Asbestos has been used throughout ancient history, including in the Roman and Egyptian civilisations. Around the early 1900s its mining and application was increased, and over the following years asbestos became the material of choice for roofing panels, spray on fire retardant, ceiling tiles, floor tiles, insulation for pipes, walling material, and as woven fabrics for fireproof garments.

Though a relationship between fibrosis of the lungs and asbestos was speculated for many years [54], it was 1960 before the relationship between pleural mesothelioma was conclusively described in the literature [55].

There are several forms of asbestos, differentiated by their composition. Blue and brown asbestos — the most harmful types — were banned first in the UK in 1985, followed by white asbestos in 1999 [57]. Due to this ban, case rates are beginning to level-off in the UK, and are projected to decline into the future (shown in Figure 3.2). Despite the ban, asbestos remains in many buildings built before this time. For

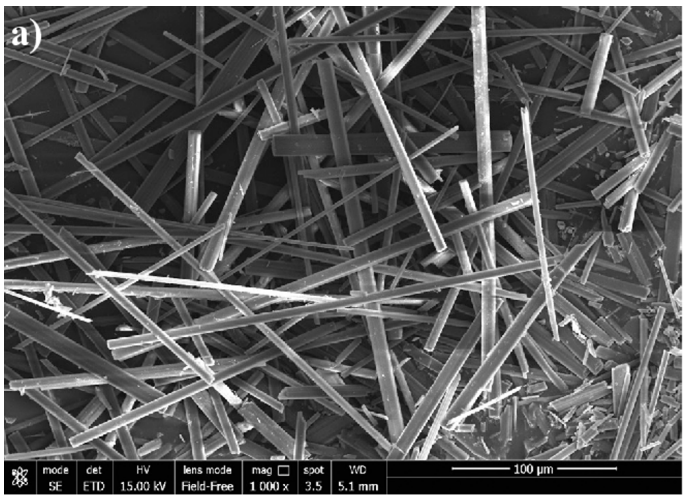


Figure 3.1: *An SEM image of asbestos fibres, which are similar in length to the diameter of many mammalian cells (Figure from [56])*

example, it is estimated that as many as 90% of public school buildings in England still contain asbestos [58].

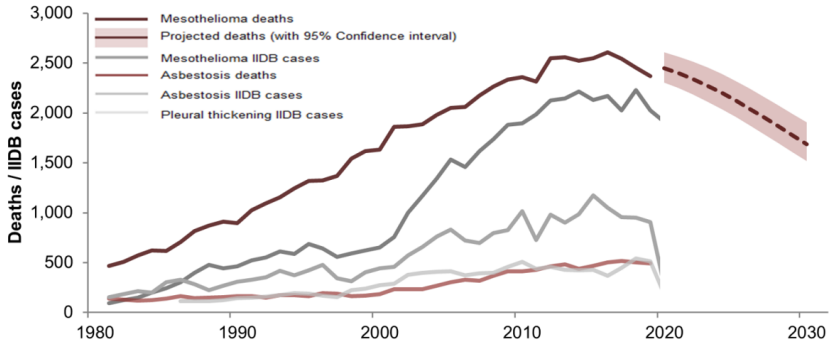


Figure 3.2: *UK mesothelioma, asbestosis and pleural thickening deaths and Industrial Injuries Benefit Disablement (IIDB) cases (figure from [59]).*

Elsewhere in the world, the mesothelioma epidemic is only starting [60]. In countries such as Zambia, Colombia, Russia, India, China and Kazakhstan there remains little or no legislation around asbestos use, and a rise in mesothelioma cases could be expected in the future [60, 61, 62, 63].

In most countries, legislation and regulation surrounding asbestos use has taken time to emerge as the impact to health has been realised. This realisation was gradual due to the long latency of the disease — symptoms do not develop until some decades after exposure, and it is only years later that the damage to public health may be fully realised.

3.2.2 Disease Development

After a period of prolonged inflammation caused by asbestos fibres lodged in the lung wall, heart or diaphragm MPM tumour may develop. The fibres may gradually move further into the pleural lining, causing chronic inflammation and genetic changes which cause cells to become cancerous [64]. Often there is a significant period between exposure to asbestos and contraction of the disease, as shown by Faig *et al.* [65] with an observed mean latency of 49.8 years. A patient with mesothelioma is likely to experience chest pain, shortness of breath, fatigue, fever, reduced appetite and swollen fingertips.

The tumour develops in the pleural space around the lungs, taking a shape like the rind on an orange (illustrated in Figure 3.3). Often fluid (pleural effusion or PE) may build up in the affected lung, leading to the possibility of eventual lung collapse. In the advanced stages of the disease, it is possible that malignant cells may spread through the lymph system and form new tumours at distant sites.

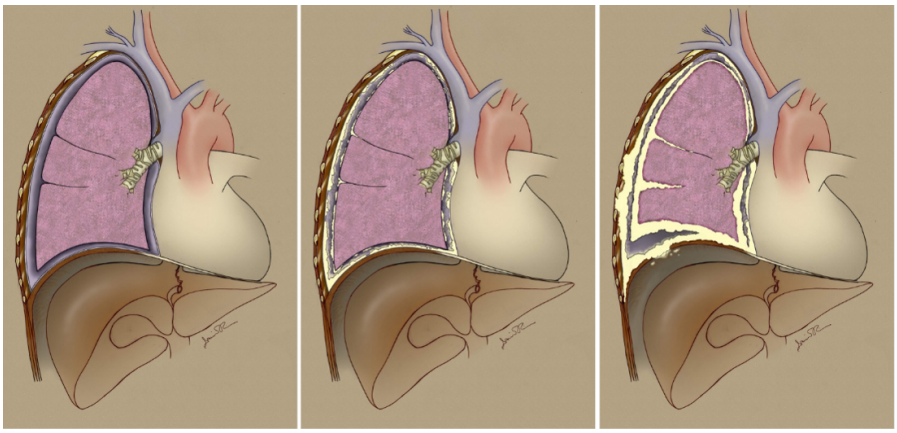


Figure 3.3: *An illustration of MPM development. The left panel shows the lung, heart and liver of a healthy individual. Tumour starts to develop in the central panel shown to enclose the lung in pale yellow. After time, the tumour may grow between the lobes of the lung, as shown in the right panel. Illustration by David C. Rice.*

Those diagnosed with the disease have a median survival of around 13-15 months [66] beyond their diagnosis, and the efficacy of current treatments is poor.

3.2.3 Disease Treatment

There is no curative treatment for MPM, however surgery, radiotherapy and chemotherapy or a combination thereof may extend median survival times. Due to the often disjoint and delocalised nature of the tumour, targeting by radiotherapy is often

challenging, and is likely to be advised in cases where the tumour is situated in a localised mass. Several therapeutic options for chemotherapy exist, for example the combination of Pemetrexed with Cisplatin has been shown to increase median survival to 12.1 months, compared with 9.3 months by Cicplatin alone (which considered the previous standard of care) [67]. Chemotherapy alone is used most often amongst patients considered to have unresectable disease by surgery, and where possible a multi-modal treatment approach is applied.

Clinical trials aimed to investigate new treatments are expensive. One component is related to the rarity of the disease — it can be challenging to source cohorts of a large size without a prolonged duration study or multi-site collaboration. Another factor is the nature of the disease — the typically high surface to volume ratio of the tumour makes it extremely difficult and time consuming to measure volumetric change, and the efficacy of new treatments are often difficult to determine by cheaper, routine measurements.

3.2.4 Manual Mesothelioma Measurement

Mesothelioma may be suspected on presentation if the patient has a known history of asbestos exposure. Imaging will be performed of the lungs. Typically this will be be acquired by CT, however some confounding structures are better differentiated in MR. Specifically, pleural effusion is a common feature which shares very similar H.U. values with tumour in CT, and distinction is more easily made in MR. However due to the relative rarity of MR scanners in the UK, imaging by this modality is not clinically routine. Following the imaging, a biopsy will be taken. It is impossible to determine solely from imaging whether mesothelioma is present. If a biopsy confirms mesothelioma, a treatment regime may commence, and the routine measurement to track patient development is known as the modified-RECIST Score.

RECIST Scoring System

The RECIST (Response Evaluation Criteria in Solid Tumours) scoring system ([7]) was originally developed to measure the change in tumours which are approximately spherical in shape e.g. lung nodules. For spherical tumours, uni-dimensional measurements of diameter across two time points are sufficient to track patient progression. This measurement can be made extremely quickly. For spherical tumours, there is

often no benefit in performing a full volumetric segmentation.

Modified RECIST Scoring System

Mesothelioma tumour is not spherical in shape. A modified version of the RECIST score (mRECIST) was developed for mesothelioma tumours. Rather than diameter, the thickness of tumour is compared across two time points. Firstly, three regions where the tumour appears thickest are selected from the CT image. The thickness of these regions is measured perpendicular to the lung wall, and these three measurements are summed. Examples of such measurements are shown in Figure 3.4. In the follow-up scan, corresponding regions are identified and the tumour thickness is again measured and summed. The two thickness sums are compared and the patient is classified as follows [7, 68]:

1. Complete Response (CR), indicating a disappearance of all known disease;
2. Partial Response (PR), indicating a 30% or more decrease in the mRECIST score;
3. Stable disease/No change, indicating that no new lesions have appeared, and the mRECIST score has not significantly changed;
4. Progressive Disease (PD), indicating a 20% or more increase in the mRECIST score, or the appearance of new lesions.

The mRECIST score is prone to poor inter- and intra-annotator agreements [69]. Yoon *et al.* show that the inter-observer 95% limit of agreement (LOA) values exceed the cut-offs described in the RECIST classification [70], indicating that it is challenging to achieve agreement in classifying tumour progression. The identification of regions where the tumour is thickest is a heuristic task — different experts may select different regions. The measurement is sparse — it does not measure the entirety of the tumour. Additionally, there is inherent ambiguity in differentiating the complex structures in the images, and finding corresponding regions in images which may have significantly changed in appearance between baseline and follow-up is a challenge.

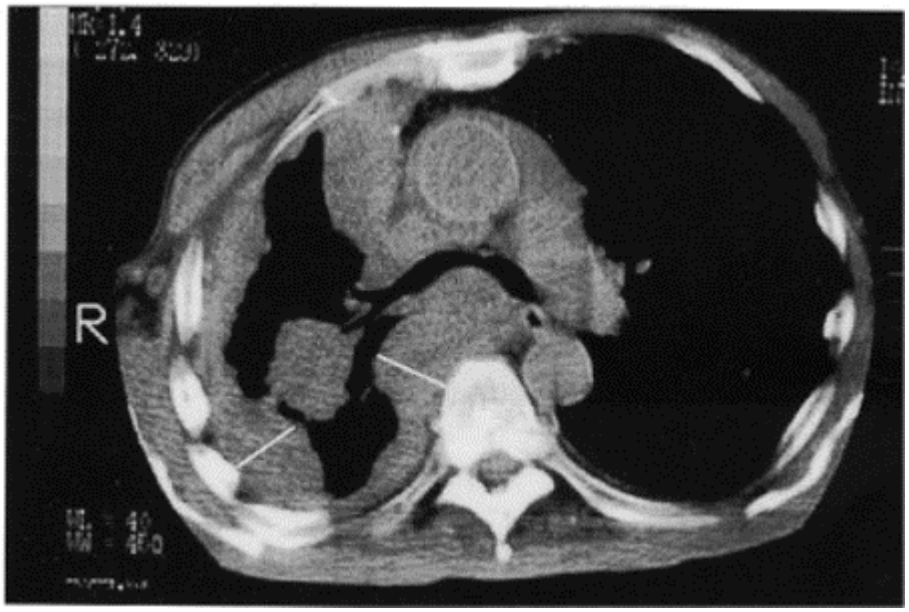


Figure 3.4: *A slice from a CT image of a patient with mesothelioma. Two example mRECIST thickness measurements are shown as white lines on the tumor. Figure from [68].*

Manual Volumetric Segmentation

Full volumetric segmentation is the gold-standard mesothelioma measurement. Frauenfelder *et al.* [71] measure agreement in tumour progression classifications between three observers across 30 MPM cases. They show that that where the mRECIST scoring system results in 14/30 cases classified with total agreement, this number is increased to 27/30 when using a volumetric approach.

Due to inherent ambiguities in appraising the cases, disagreement between annotators is likely to remain. Labby *et al.* [72] compare five independent expert annotators and report 95% CIs of 311% and 111% for volumetric measurements performed on baseline and follow-up images respectively. Access to baseline contours when annotating the follow-up images introduced bias, and the authors propose that this is the source of a lower CI for the follow-up images. Figure 3.5 shows an example of the differing segmentations provided by the observers.

Volumetric tumour measurement it extremely time consuming and expensive to perform. For the most accurate measurements, an expert must delineate the tumour on each axial CT slice. The high surface-to-volume ratio and irregular nature of the tumour means that the volumetric measurements are sensitive to interpolation (either between annotation on slices or between annotation points) provided by standard annotation software. The time to perform a tumour segmentation is highly

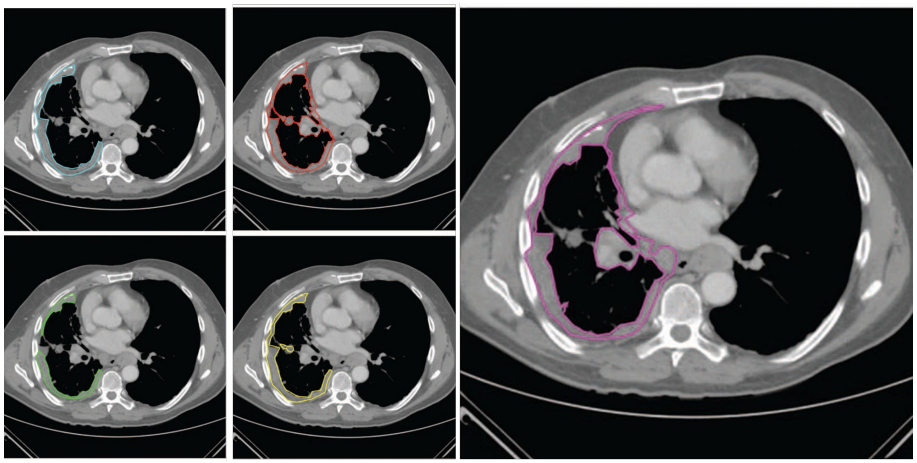


Figure 3.5: *Five independent tumour segmentations produced for a corresponding axial CT slice. Figure adapted from Labby* et al. *[72].*

variable — depending largely on the volume of tumour present — and times may range from 60-90 minutes per volume. Measurements of volumetric change require two full volumetric segmentations to be performed for two time-points.

The development of a fully automated tool for mesothelioma measurement may greatly reduce the time-to-annotate for an expert — segmentation by an automated method could be used as a starting point for an expert review. In case where such an automated tool produces results which are within the agreement bounds of expert human annotators, it may be used to evaluate the efficacy of new treatments and statistics derived from the tool could be used to monitor patient care and inform treatment decisions.

3.2.5 Automated Mesothelioma Measurement

Due to the clear opportunity for automation of MPM, several methods have been developed to semi-automate the measurement of mesothelioma to varying degrees.

Gudmundsson *et al.* [73] use CNNs to segment MPM tumour, pleural effusion and pleural plaques as a single segment. They train two U-Nets for separate analysis of the left and right lung, and the laterality of the disease must be manually input to select which CNN is used to analyse the images. Over a test set comprising 131 axial slices (taken from 41 patients) with MPM tumour segmented they achieve Dice coefficients ranging between 0.662 to 0.800. Crucially, their test set is annotated with MPM tumour segmentation, rather than the task for which the algorithm was trained. It is likely that these tasks are congruent due to the level of ambiguity in differentiating many of the structures, and the score is similar to inter-observer

Dice coefficients. In later publications, they aim to exclude pleural effusion from the measurements, as the volume of pleural effusion is unrelated to tumour volume [74], acting to confound the predictions.

Chen *et al.* [75] propose a semi-automated method based on a random walk segmentation of the tumour, initialised by a minimum of 20 manually placed seed points per slice within the tumour region. The median time-to-annotate for their method is reduced by semi-automation from 68.1 min to 23.1 min. They achieve a Dice coefficient of 0.825 over a test set of 15 subjects.

Earlier work by Sensakovic *et al.* [76] firstly aim to detect the hemithoratic cavity, based upon a segmentation of the liver. This liver segmentation involves the manual delineation of the liver in several axial CT sections, which are interpolated to define the liver. Once the hemithoratic cavity is identified, the user must input the laterality of the disease. The pleural space is then identified automatically, and a k-means classifier is used to segment MPM tumour within the pleural space based on H.U. intensity. Based on an evaluation comprising 5 CT sections taken from 31 subjects, the automatic segmentation achieved a median Dice coefficient of 0.484 compared with 5 observers. This is similar to the median Dice coefficient across the 5 observers of 0.517.

Brahim *et al.* [77] localise the thoratic cavity and perform subsequent texture analysis to derive the MPM tumour. Over a test set of 10 CT images, they achieve a Dice coefficient of 0.88.

3.3 Tumour Shape Analysis

It a challenge to achieve high agreement between observers when segmenting long and thin regions such as MPM tumour, both at the image-level (e.g. as measured by the DICE coefficient) and at the total volume level. This may be partially due to confounding image features, and partially due to the shape of the region. Analysis of tumour shapes may allow us to determine how sensitive volumetric measurements are to the tumour delineation. I designed a simulation to determine what change in volume could be expected from an expansion or contraction in the segmented region of a few voxels. This provides an indication of the variability one may expect from measurement error alone, without the confounding effects of image interpretation.

Method: Several binary masks of 3-D shapes were analysed:

- a binary mask of a sphere, containing a volume comparable to that of MPM tumour,
- a binary lung mask of a single lung, generated from a subject from the LIDC-IDRI dataset,
- a binary mask of a lung annulus, generated by applying a threshold to the distance transform of the lung mask described above, and subsequently dilating to a thickness of (n) mm,
- and a binary mask of MPM tumour, generated from a subject from the PRISM study.

The analysis comprised performing either the binary dilation or binary erosion morphological operation and subsequently evaluating the volume ratio V_f/V_i, where V_i pertains to the volume before morphological processing, and V_f pertains to the resultant volume of the region following the processing. Morphological processing was performed in the 2-D axial plane (rather than 3-D processing), in order to reflect the nature of annotation process. The volume ratio V_f/V_i is plotted as a function of number of pixels change in the boundary location (where erosion is described as a negative change, and dilation as positive change).

Results: The results are shown in Figure 3.6. The sphere and lung annulus show a low sensitivity of volume to the boundary location. These shapes have a low

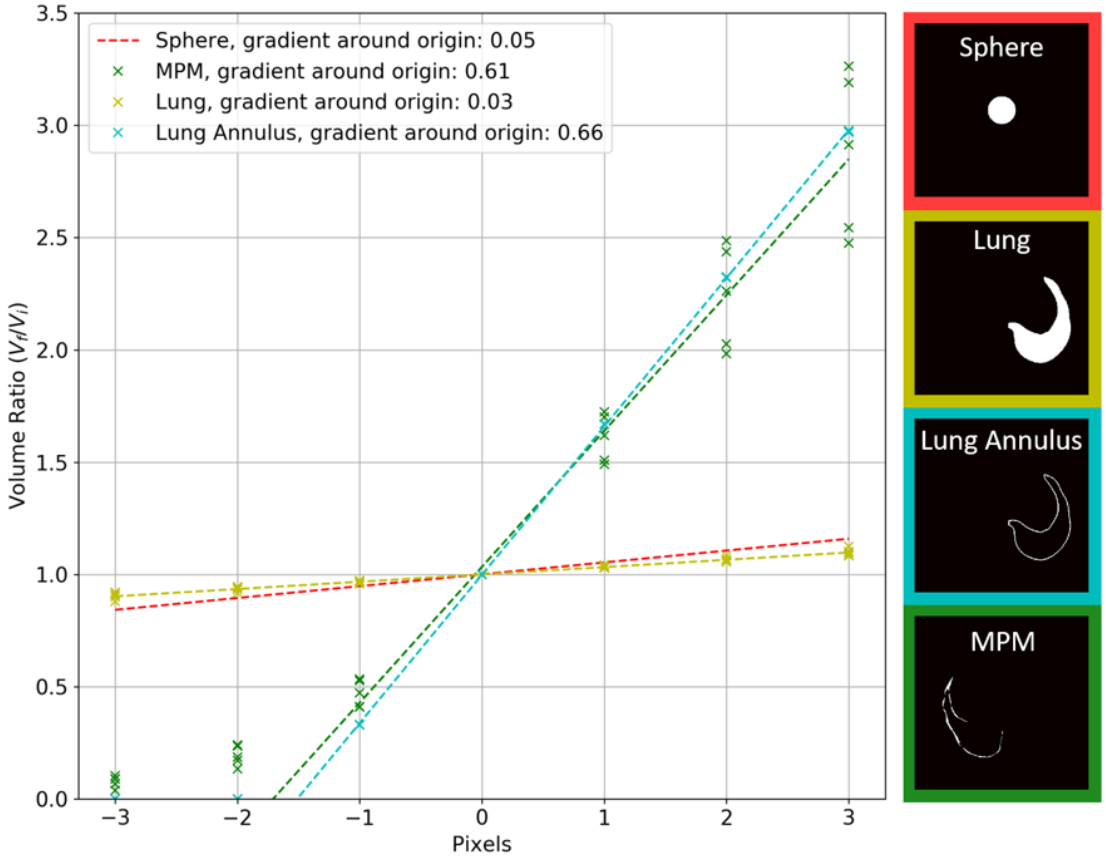


Figure 3.6: *The volume ratio following binary erosion or dilation for four shapes: a sphere, in red; a lung segmentation mask, in gold; a lung annulus, in blue; and a MPM segmentation mask, in green.*

surface-to-volume ratio. For the synthetically generated lung annulus and MPM tumour segmentation masks, a variation of one pixel (or voxel) in the placement of the boundary lines can results in volume changes of 60%. Measurement error may be approximated as half of the smallest measurement, and this experiment shows that when dealing with extremely thin tumours the measurement error from the shape alone may be approximately ± 30%.

Across the literature, segmentation agreement (as measured by the Dice coefficient) is extremely variable for both manual and semi-automated tumour measurements. This variability comes from the nature of the data — images of early stage subjects are likely to contain narrower regions of tumour, and for these cases high agreement is more difficult to achieve. This data-dependent variability makes comparisons between methods a challenge.

3.4 Preliminary Experimentation

I conducted a number of preliminary experiments throughout the development of the algorithm presented later in this chapter. As shown in Figure 3.7, many early results showed qualitative promise. For example, regions of pleural thickening were often identified by the algorithm. However, Figure 3.7 also shows how the promising results often showed extremely limited overlap with the reference ground truth, which can lead to poor numerical performance. This section intends to provide a brief summary of early experiments prior to the publication of a functional method.

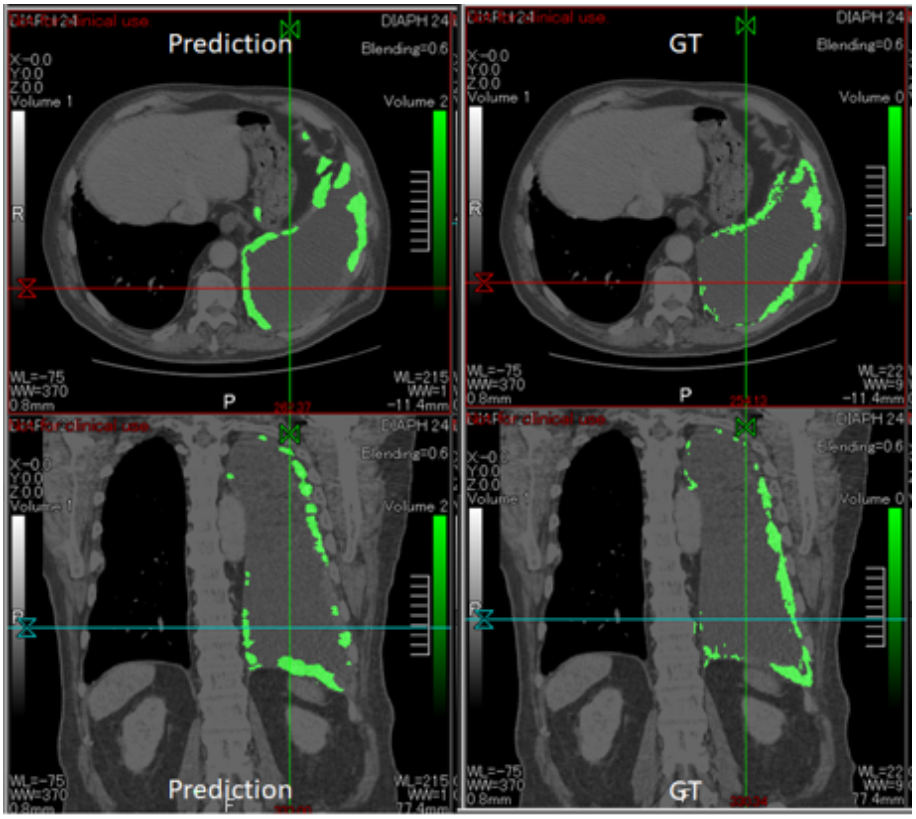


Figure 3.7: *Axial (top) and coronal (bottom) CT views with tumour segmentation shown in green overlay. Manual segmentation is shown in the right column, and a predicted segmentation by an early method is shown left. False positive regions may be seen by the automated method.*

Fully Convolutional Approach: Fully convolutional approaches to image segmentation are desirable due to their simplicity and efficiency. These methods do not involve any down-sampling, thus receptive field is solely dependant on network depth. Generally, these methods are more suited to segmentation tasks which can be completed using lower-level features. Several novel fully convolutional architectures

were developed to process 2-D CT slices for mesothelioma segmentation, however none of these methods converged during training. It was concluded from these experiments that the accurate delineation of mesothelioma tumour is highly dependent on higher level anatomical features. At a local image level, distinction between tumour and neighbouring structures and pathology is an ill-posed problem, and successful approaches should be capable of utilising high-level anatomical features.

Input Image Resolution: Input image resolution was designed as a variable parameter in the experiment pipeline, and a selection of resolutions were tested for automatic segmentation. These experiments showed that, especially for images of early stage disease, the segmentation masks of thin and irregular shapes were extremely difficult to resample accurately. Some regions may be as narrow as a voxel, and the segmentation may be lost entirely upon downsampling. Similarly, there are inherent ambiguities to upsampling such narrow regions. These added confounding factors to interpretation of the results. To avoid these complexities, it was concluded that resolution should be maintained at the resolution at which the images were carefully annotated.

Hard Dropout for Uncertainty Estimation: For clinical application of deep learning algorithms, it is desirable to provide a measure of model certainty with an output prediction. This would allow for better interpretation of the results — for instance, it would be possible to show that the algorithm is more uncertain where image ambiguity is high. For this purpose, hard dropout as a Bayesian approximation of uncertainty was investigated [41]. This method involves using the model to generate a segmentation mask multiple times using a subset of different kernels. It was determined that uncertainty generated by this method had no discernible correlation to model error as calculated from the reference ground truth by a single observer. This may be due to a variety of reasons. It is possible that in the case where agreement between observers is low (as is the case with MPM tumour segmentation) the agreement between multiple annotations may be necessary to benchmark an uncertainty score.

Pre-trained Network Weights: When training neural networks, it is a common approach to use pre-trained weights as opposed to a random initialisation. We com-

pared both pre-trained weights (based on ImageNet [40]) and random initialisation, and found that overall there was no significant difference in performance between the two approaches. However, it was clear that using pre-trained weights reduced the number of epochs required for the algorithm to reach the optimum stopping point during training.

3.5 Methods

I have developed an automated approach for the segmentation of MPM tumour from CT images as part of a retrospective cohort study funded by the Cancer Innovation Challenge (Scottish Health Council). The work in these sections has been published as a conference proceeding as "Fully Automated Volumetric Measurement of Malignant Pleural Mesothelioma from Computed Tomography Images by Deep Learning: Preliminary Results of an Internal Validation" [25] and Springer book chapter titled "Estimating the False Positive Prediction Rate in Automated Volumetric Measurements of Malignant Pleural Mesothelioma"[26], on which the following material is based.

3.5.1 Data

123 volumetric CT datasets from 108/403 subjects recruited to the DIAPHRAGM and PRISM research studies were used to train and cross-validate the automated method, all of which had a confirmed histological diagnosis of MPM. A further subset of CT datasets from the NLST archive were utilised to test the automated detector.

PRISM (Prediction of ResIstance to chemotherapy using Somatic copy number variation in Mesothelioma) [78] is a retrospective cohort study to determine a genomic classifier that predicts chemo-resistance in MPM. The study involves retrieval of tumour blocks pre- and post-chemotherapy from 380 subjects across five UK centres. 123 CT images from 85/380 PRISM subjects are included in this study (43 images acquired pre-treatment, and 80 images acquired post-treatment), from centres in Glasgow.

DIAPHRAGM (Diagnostic and Prognostic Biomarkers in the Rational Assessment of Mesothelioma) [79] was a 3 year prospective observational study, which involved 747 patients from 23 UK sites. Subjects were recruited to the study upon first presentation of MPM. A subcohort of 23/747 subjects from Glasgow centres (who had both MRI and CT images) were selected. All the selected images were acquired pre-treatment. Contemporaneous MRI images are useful in disambiguating some confounding features in CT images.

NLST (National Lung Screening Trial) [80] was a multicentre study which aimed to compare low-dose CT with chest radiography for lung cancer screening. The study targeted older (55–74 years) ex- and current smokers. 46,613 CT images from 14,965 subjects are used to test detector specificity.

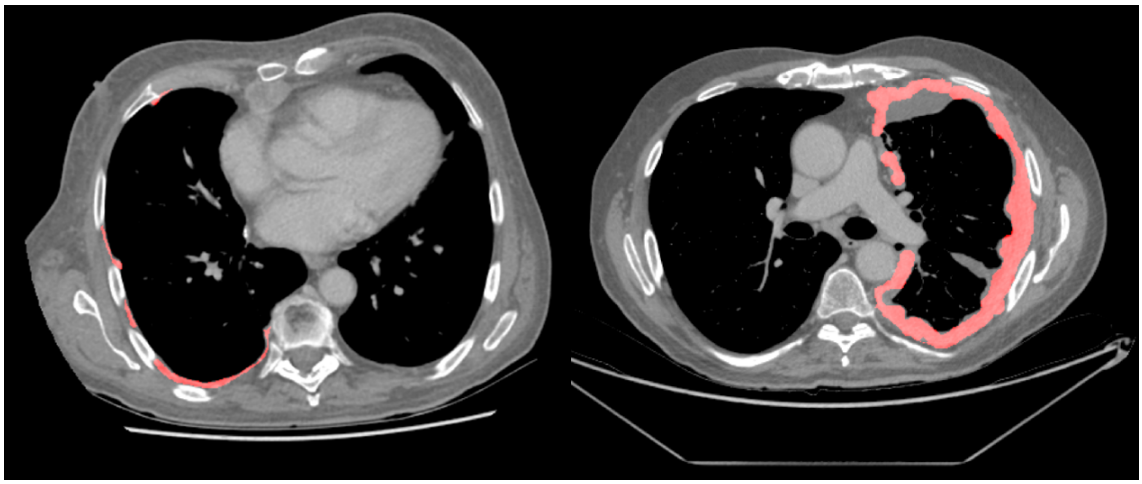


Figure 3.8: *Two axial CT slices from two subjects in the cohort, with manually derived MPM tumour segmentation shown in red. Left: A slice from a CT image taken in the DIAPHRAGM study. Right: A slice from a CT image taken in the PRISM study. The unsegmented areas (in grey) represent adjacent pleural fluid. Figure from [25].*

The images from the DIAPHRAGM study were acquired earlier in development of MPM with respect to those from the PRISM study, and consequently the tumour volumes tend to be smaller and thinner in the DIAPHRAGM study. Slices from a PRISM and DIAPHRAGM dataset are compared in figure 3.8.

Ground truth generation:

A respiratory clinician with training in image analysis and mesothelioma identification manually segmented the MPM tumour in 123 CT images from the PRISM and DIAPHRAGM studies. Tumour segmentation was performed in the axial plane using Myrian software (Intrasense, Paris). Segmentations were performed in all slices containing tumour for 80/123 images. For 43/123 images a more sparse annotation was performed where every fifth slice was annotated. Consecutive slices are highly correlated — both in appearance and in terms of the tumour characteristics. Annotating a subset of slices allowed a greater number of subjects to be included in the training set, increasing the diversity of this cohort. Although beneficial to training the algorithm, a sparser annotation resulted in datasets which could not be used to evaluate volume accuracy, and were not included in the evaluation of the

algorithm.

The manual segmentation was drawn free-hand, rather than using any interpolation or semi-automated tools (e.g. region growing techniques). This was to avoid sources of bias in the ground truth generation process. For example, region growing techniques may bias the annotation process towards inclusion of regions with certain ascetic qualities. Given the ambiguity in delineating the boundaries of mesothelioma tumour, these included regions may not be incorrect per-se, but consistent over-segmentation would result in a dataset which did not fully capture this interpretation ambiguity. Similarly, interpolation of the segmentation between slices would bias the annotator to more closely replicate their own interpretation of the previous image slice.

Ground truth inter-slice consistency processing:

The MPM tumour was manually segmented in the axial plane. A free-hand segmentation was required to capture the complex shape of the tumour, and inevitably this leads to some annotation inconsistencies between slices. These appear as a discontinuities of the tumour segmentation in the orthogonal, sagittal and coronal planes, contrasting with the continuous nature of the tumour viewed in the axial plane of annotation (figure 3.8). For many measurements inconsistencies of this nature are negligible, however for MPM measurement the between-slice inconsistency can have a significant effect on volumetric measurements. To improve between-slice consistency, a three-dimensional binary closing operation (figure 3.10) was performed using an $11 \times 11 \times 11$ voxel structuring element. A limitation of this approach is that any genuine holes in the tumour smaller than five voxels will be removed.

3.5.2 Cross-validation

The algorithm was evaluated by k-fold cross validation, where a setting of $k = 7$ was found to provide robust group statistics for each test set, whilst maximising the amount of training data at each fold. As described in section 3.5.1, 43/123 datasets were sparsely annotated, and could not be used to evaluate volumetric accuracy. These datasets were used in the training set for all seven folds. The 80/123 datasets with full annotation were randomly assigned to seven folds, to provide a validation set of 11 or 12 datasets per fold. The 68 or 69 remaining datasets are

further sub-divided by a 30:70 split, where 30% is used to determine the best model and select an optimal model threshold (referred to as the internal validation set), and 70% is used as the training set, to which the 43 sparsely annotated datasets are added.

Neighbouring CT slices are highly correlated, and including all the slices from a CT images biased the algorithm towards maximising performance on the images with the greatest number of slices. To counter this, fully segmented CT images were also subsampled to 100 slices when training the algorithm.

Performance metrics:

Absolute volume correspondence and segmentation accuracy are used to evaluate agreement between the automated method and manual observer. Given only single time-point images were available in this preliminary evaluation, we were unable to evaluate volume *change* accuracy.

Bland-Altman analysis [81] is used to evaluate volumetric agreement between the automated and manual segmentations. This plots the difference of two measurements against the mean of the two measurements, together with the mean difference and the 95% limits of agreement. The following summarises the volumetric agreement statistics:

1. The mean difference (or bias) between the two measurement methods
2. A test for whether the mean difference between the two measurement methods is significantly different from zero, determined using a two-sided paired t-test (MATLAB statistics toolbox, Mathworks, Natick).
3. The 95% limits of agreement [81].
4. A test whether the difference between the measurement methods increases (or decreases) as the tumour volume increases. This was determined from the slope of a least squares regression fit to the points in the Bland-Altman plot. Specifically, it tests whether the slope is statistically different from a zero gradient, based on t-statistics (MATLAB statistics toolbox, Mathworks, Natick).

The Dice Score is used to measure region overlap between the manual and automated measurements. Although volumetric agreement is the primary property of

interest, it does not show whether the same regions have been delineated, or whether the regions intersect. The Dice score provides a measure of these properties.

3.5.3 Algorithm

To automatically segment MPM tumour, a Convolutional Neural Network (CNN) was trained.

Architecture:

The CNN was a U-Net architecture [23] — similar to the method used by Gudmundsson *et al.* [73, 82]. Our network (figure 3.9) takes three axial slices at a time, and predicts a segmentation at the central of these slices. The encoder is pre-trained in a VGG classifier on the ImageNet challenge data [83]. For pre-training, the three-channel input was used to consume three-colour natural images.

All network activations are rectified linear units, aside from the ultimate layer of the network, which was a softmax activation. Dropout (with a rate of 0.2) [84] was used to prevent over-fitting and batch normalisation [85] was used at the locations illustrated in figure 3.9 to improve the training characteristics. The network was implemented and trained using the Keras framework [86].

The main benefit of the U-Net, as opposed to a standard CNN with no down-sampling, is the increased receptive field. For this analysis, voxel intensities from the entire axial slice (512×512) may be used to inform the segmentation output at any voxel. For mesothelioma, where the appearance of the tumour may be subtle, a manual annotator may rely on their understanding of the surrounding anatomy and knowledge of how the disease progresses to inform a manual segmentation. So a U-Net with a large receptive field was found most suitable, allowing the automated approach to leverage any spatial priors and surrounding structures to inform the segmentation.

Though similar, this algorithm differs from the first stage detector presented in Chapter 2. Specifically, it operates with larger (higher resolution) input images, to allow use of fine-grained textural features to disambiguate tumour from other pathology. It also outputs two segmentation channels — one pertaining to background, and the other pertaining to segmented tumour.

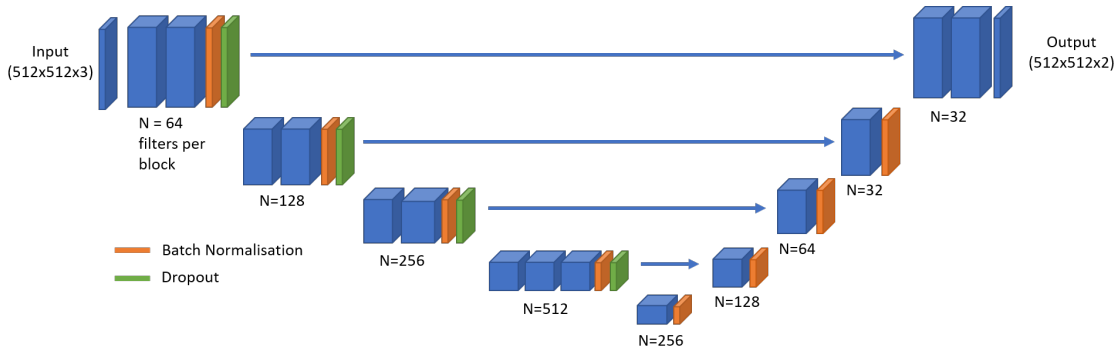


Figure 3.9: *A schematic of the U-Net model architecture. The blue boxes represent a stack of convolutional filters, with the number of filters per stack shown to the left of each box. All filters have a dimensionality of* 3×3*. Green and orange boxes represent dropout and batch normalisation layers respectively. The blue arrows represent skip connections by feature concatenation. Figure from [25].*

Image pre-processing:

CT image intensities input to the network are clipped to [-1050, +1100] Hounsfield Units, and normalised to range [-1, +1]. The images are retained at their original resolution (which is typically within the range 0.71 mm to 1.34 mm).

Training:

The network was trained for 30 epochs, after which the best performing model is selected across the epochs. This model was chosen based on highest average voxel-level accuracy for the internal validation set. For training, the Adam optimiser was used, with a cyclic learning rate [87], where the learning rate (lr) has been set to oscillate between $lr = 0.0001$ and $lr = 0.003$, with a full cycle duration of one epoch. A batch size of 8 slices (with context) per batch allowed the model (10,019,874 parameters) to train on the available GPU.

Despite MPM tumour segmentation being a binary classification task, categorical cross-entropy was used as the objective function. Therefore, the output of the network was two-channel: one for tumour segmentation, and one for background segmentation. This objective function was found to improve convergence with respect to binary cross-entropy. The slices during training were randomly ordered, and it was possible that the class balance in the first batch was highly unbalanced. When batches were predominantly tumour negative in the first few batches, weights near the decoder of the network were optimised to zero, and training stopped as errors could no longer back-propagated. Categorical cross-entropy was used to overcome

this, a non-zero signal is always required in one of the two output channels, regardless of the class balance of the example slice/batch. This regularising effect of categorical cross-entropy increased experiment repeatability between runs and folds of analysis.

Binarisation:

The output of the CNN was a probability map, showing the probability of MPM tumour at every voxel in the input CT slice. This output was binarised by applying a threshold. The optimal threshold for the CNN was chosen to maximise the mean Dice coefficient between the binarised prediction and manual annotations in the internal validation set. The optimal threshold varied slightly between models — different training datasets had varying levels of complexity, resulting in models which predicted in varying probability ranges. The internal validation sets at each fold also contained different disease characteristics, which added variance to the optimal threshold between folds.

Tumour volume:

For validation, the algorithm was used to segment the MPM tumour in every slice of the input CT images. Tumour volume was then calculated [25]:

$$M(x,y,z) = \begin{cases} 1 & \forall P(x,y,z) > t \\ 0 & \forall P(x,y,z) \leq t \end{cases} \tag{3.1}$$

where M describes a segmentation image of same dimensionality as the input CT image, with each voxel assigned a binary value of one to indicate MPM tumour and zero elsewhere. M was calculated by evaluating the probability map ($P(x,y,z)$) with respect to the optimal threshold, t. This binary segmentation was then converted into a measurement of tumour volume (V) [25]:

$$V = S_x S_y S_z \sum_{x=0}^{X} \sum_{y=0}^{Y} \sum_{z=0}^{Z} M(x,y,z), \tag{3.2}$$

where S_x, S_y and S_z denote the image voxel sizes in x,y and z respectively.

3.5.4 False Positive Rate Estimation

NLST Study Data:

The National Lung Screening Trial (NLST) study enrolled 53,454 persons at high risk for lung cancer between 2002 and 2004 from 33 centres in the United States. The study had two arms, comparing chest X-rays and CT imaging for detecting lung cancer. 26,722 participants were enrolled in the CT arm of the study. Of these, 14,965 subjects are used to provide a further testing set for the automated mesothelioma detector. The subset of NLST images was selected to include subjects with reported lung abnormalities and lung nodules. The NLST study was not focused on mesothelioma, and it is unlikely that many images in the study contain mesothelioma (it was not indicated as an incidental finding for any images in the study). Hence this dataset is used to analyse the specificity of the automated detector across a large and independent cohort. Since imaging alone cannot give a definite diagnosis of mesothelioma — the appearance of the tumour in CT images is similar to many other findings — biopsy is often the only definitive test. For this reason, it is possible that images acquired for the NLST study contain one or more subjects with mesothelioma.

Time-points: The CT images acquired for the NLST study spanned three annual time-points. Participation was terminated upon either: a) completion of the third time point, b) subject drop out, or c) a significant finding impeding the ability to complete the study. In this analysis, images from all the available time points were included in the analysis.

Study Findings: As a part of the NLST study, a variety of findings of relevance were recorded. For the purposes of this analysis, hyperdense pathologies which have a bright appearance in CT images are of relevant — such findings are most likely to be confused with MPM by the automated detector. The specific NLST findings of interest are listed in table 3.1. Note that since the NLST study recorded findings by *subject* rather than by image, not all of the images from a subject with a positive finding will necessarily contain evidence of the specific finding(s).

Reconstruction Kernel: The CT images acquired for the NLST study were reconstructed using "hard" (sharp) kernels, "soft" kernels. For some subjects images

Table 3.1: *List of NLST study findings considered positive in the false positive detection rate analysis.*

NLST findings of interest
Pleural thickening or effusion, Non-calcified hilar/mediastinal adenopathy or mass, Chest wall abnormality, Consolidation, Emphysema.

were reconstructed from the same acquisition data using both types of kernel. CT manufacturers offer a variety of different reconstruction kernels, described by their own naming conventions. Table 3.2 lists the kernels by CT manufacturer that were considered hard for the purposes of this study. In total, this resulted 20,139 hard image reconstructions and 26,474 soft image reconstructions.

Table 3.2: *List of CT reconstruction kernels considered "hard" and "soft" in this study. Kernels names used to reconstruct two or more images are listed.*

Manufacturer	Hard kernel names	Soft Kernel Names
GE	LUNG and BONE	STANDARD, BODY FILTER / STANDARD and BODY FILTER / BONE
Philips	D and B	A, C and EC
Siemens	B50f, B60f, B80f and B45f	B30f, B20f, B31f, B30s, B50s, B70f, B31s, B40f, B60s and B35f
Toshiba/Canon	FC51, FC53	FC01, FC30, FC50, FC02, FC10, FC82 and FC11

3.5.5 Experiments

The convolutional neural network was trained seven times on seven folds of the training dataset, as described in 3.5.2. The seven resulting CNN models were combined into an ensemble to generate the final volume measurement result, by calculating the mean of the volumes from the seven models. The results obtained by this method were:

1. Compared with those obtained from the individual seven models for all subjects with histologically confirmed MPM (DIAPHRAM and PRISM),
2. Stratified by whether hyperdense pathology is present (NLST),
3. Stratified by hard/soft kernel reconstructions (NLST).

The 100 cases where the algorithm finds the largest volumes of tumour were qualitatively analysed.

3.6 Results

Manual annotation time varied between subjects, taking approximately 2.5 hours per image. Automated measurements required approximately 60 seconds per image, using an Nvidia 1080Ti graphics processing unit (GPU), 32 GB of RAM and a 12-core Intel Xeon CPU (3.40 GHz).

3.6.1 Inter-slice consistency processing

Three-dimensional binary closing was proposed to increase between-slice manual segmentation consistency (c.f. 3.5.1). This processing increased detected pleural volume from 301.1 cm^3 (standard deviation 263.9 cm^3) to 514.7 cm^3 (standard deviation 336.1 cm^3) over the cohort. Figure 3.10 shows a typical binary closing result, and highlights the additional voxels added by the closing operation. Visually, the closed version appears more contiguous and physically plausible.

3.6.2 Volumetric agreement

The cohort mean predicted volume was 547.2 cm^3 (standard deviation 290.9 cm^3) across seven-folds of analysis.

Raw manual annotations:

The mean tumour volume in the raw manual segmentations is 405.1 cm^3 (standard deviation 271.5 cm^3), which is significantly lower than the automatically detected volume. The Bland-Altman plot in figure 3.11 shows a minor, though statistically significant, trend where the volume error increases slightly with tumour volume ($p < 0.001$). This indicates that on average, the algorithm over-segments the tumour compared with the raw ground truth (here the manual measurement is *without* the binary closing operation to increase consistency between slices).

Closed manual annotations:

Binary closing increased the mean tumour volume of the manual segmentations to 574.4 cm^3 (standard deviation 327.1 cm^3). The Bland-Altman plot in figure 3.12 shows that using closed manual annotations gives a mean difference of -27.2 cm^3, which is not significantly different from zero mean difference ($p = 0.225$). To facilitate

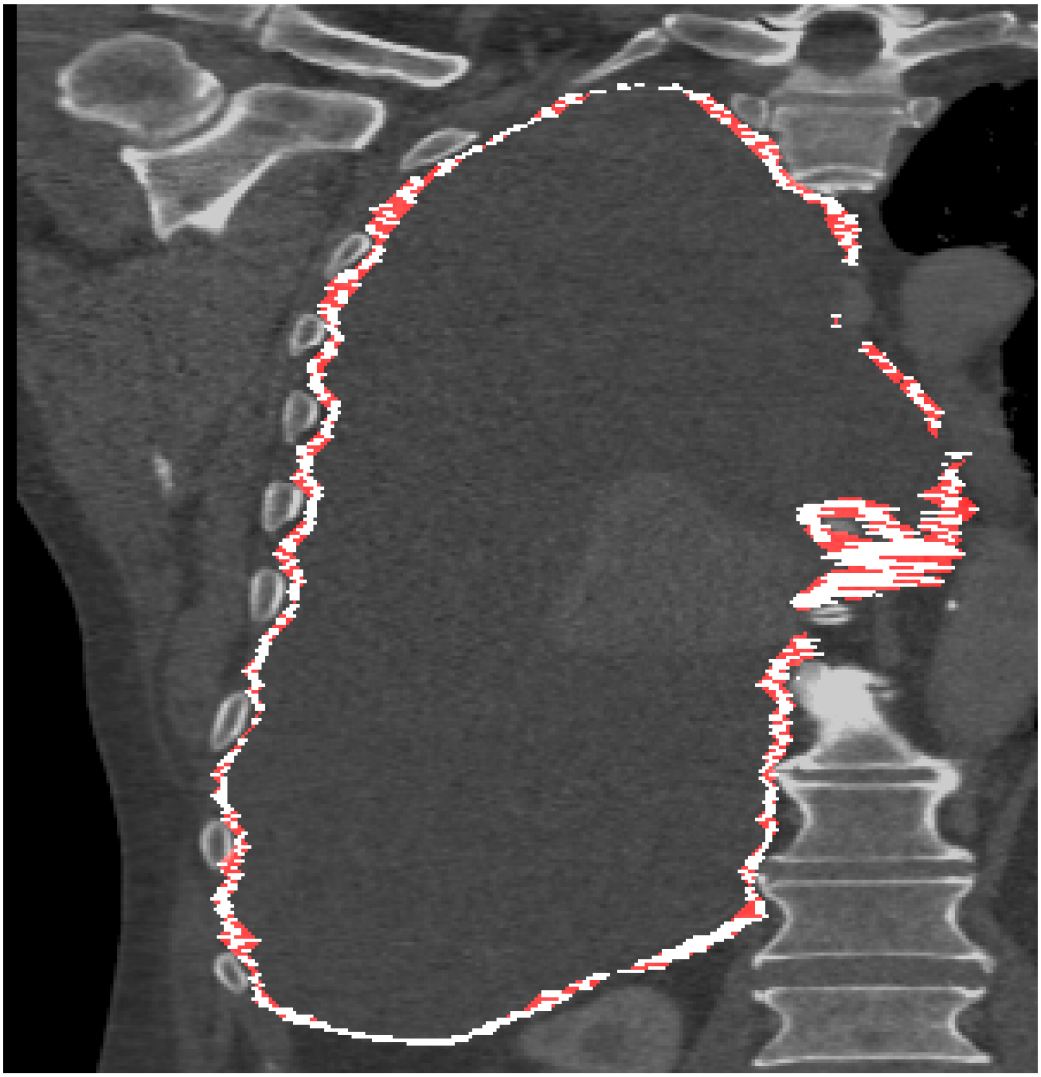


Figure 3.10: *A CT coronal view of a subject with MPM, showing the right lung. The white annotation indicates the location of tumour, as drawn by an expert annotator in the axial plane, which follows the bounds of the pleural cavity, surrounding a region of pleural effusion. Red shows the regions which are closed by a binary closing operation. Figure from [25].*

comparison to other methods, the results are equivalent to 95% limits of agreement which span 129.2% of the total tumour volume.

Four measurement differences in 3.12 are outliers (outside of the 95% limits of agreement): three of these are where the algorithm predicts a higher volume of tumour than recorded by the observer. Inspection of these cases showed extremely narrow tumour in these images. The algorithm often identifies the bulk of the tumour mass (where it is thicker and more visible), but does not propagate the tumour into the rind-like surface which, although narrow, encloses a significant proportion of the lung surface area. This is potentially where the slice-based nature of the approach

limits performance. A fully 3D CNN approach may offer higher accuracy in such cases. Inspection of the remaining outlier (under-segmentation by the algorithm) showed tumour which was unusually thick compared with the other images in the training cohort. For this case, it is likely the algorithm failed to generalise to this degree of tumour thickness, unseen during training.

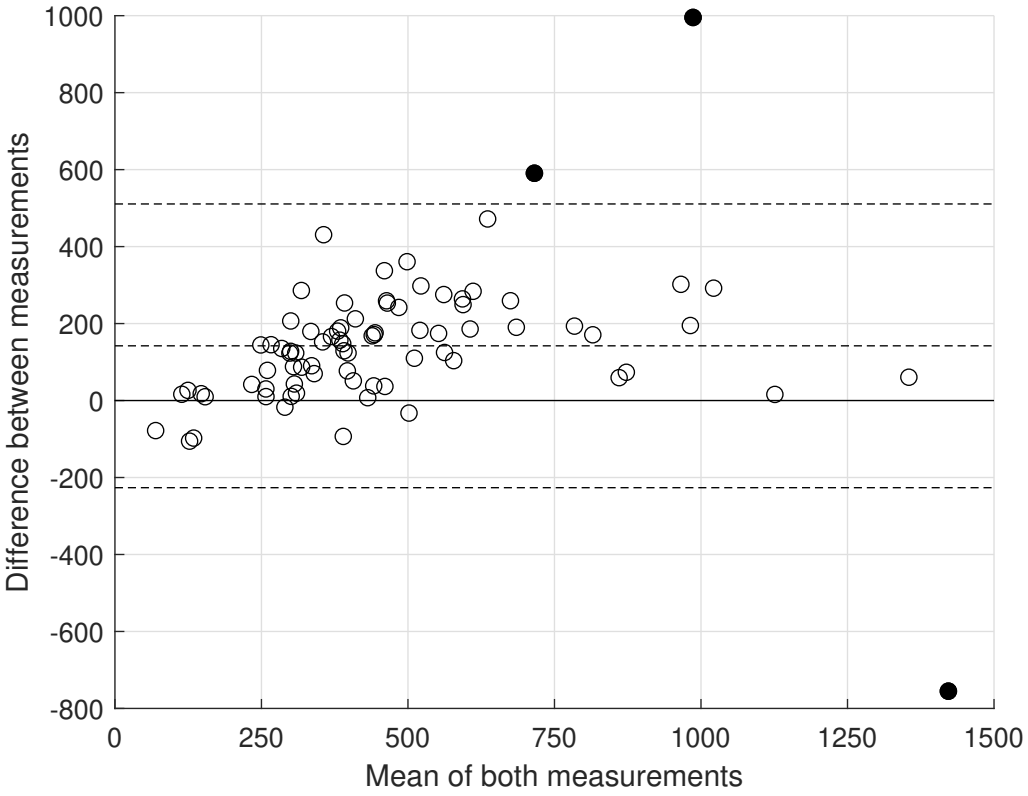


Figure 3.11: *Bland-Altman plot of the algorithm-annotator agreement for tumour volume measurements, across* 80 *subjects. The central dashed line indicates a mean difference of* 142.2 cm^3 *over-segmentation by the algorithm. Outer dashed lines indicate upper and lower 95% limits of agreement of* $[-224.1, +508.5]\,cm^3$ *respectively. Figure from [25].*

3.6.3 Region overlap (Dice score)

The mean overall Dice coefficient was 0.64 (standard deviation 0.12) using the binary closed ground truth. In comparison, the Dice score was 0.55 (standard deviation also 0.12) using the raw ground truth, confirming higher voxel-wise correspondence following binary closing to improve inter-slice consistency. Dice coefficients varied between subjects and between analysis folds. Due to the wide range of tumour shapes and volumes in this dataset (c.f. section 3.5.1), some test sets simply contained more

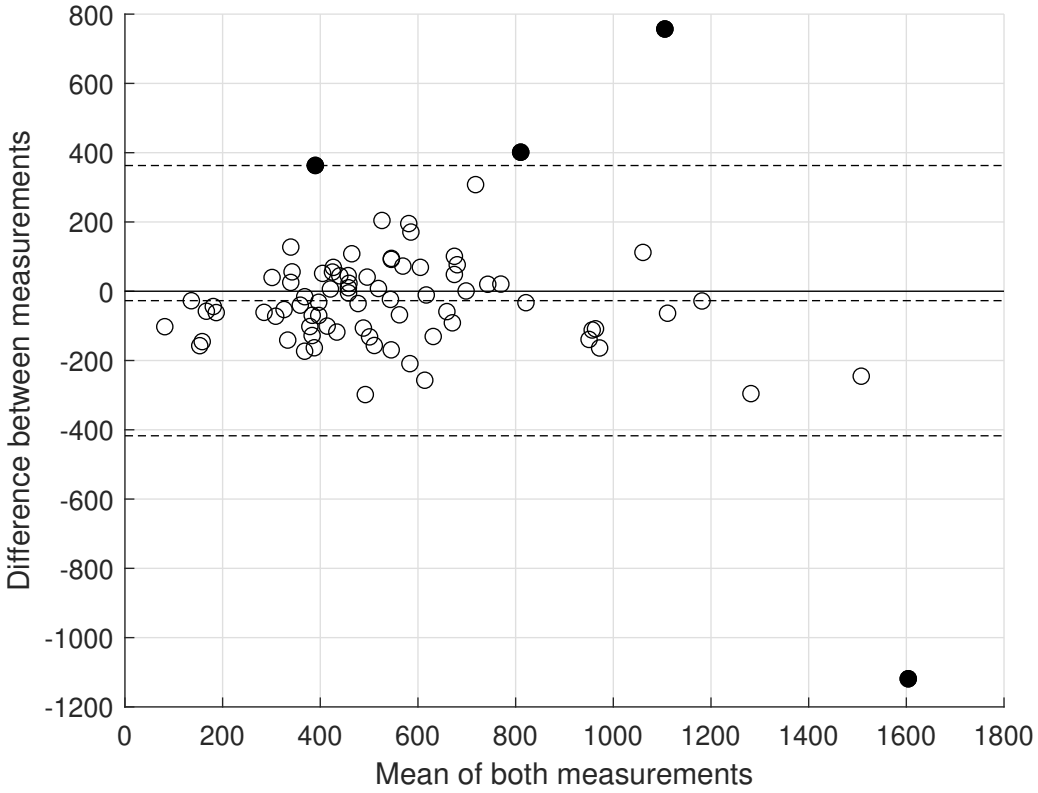


Figure 3.12: *Bland-Altman plot of the algorithm-annotator agreement for tumour volume measurements across* 80 *subjects, using cleaned ground truth. The central dashed line indicates a mean difference of* $-27.2\,cm^3$ *under-segmentation by the algorithm. Outer dashed lines indicate upper and lower 95% limits of agreement of* $[-414.2, +360.5]\,cm^3$ *respectively. Figure from [25].*

difficult cases. Figure 3.13 shows the ground truth and predicted tumour for a subject from the PRISM sub-cohort.

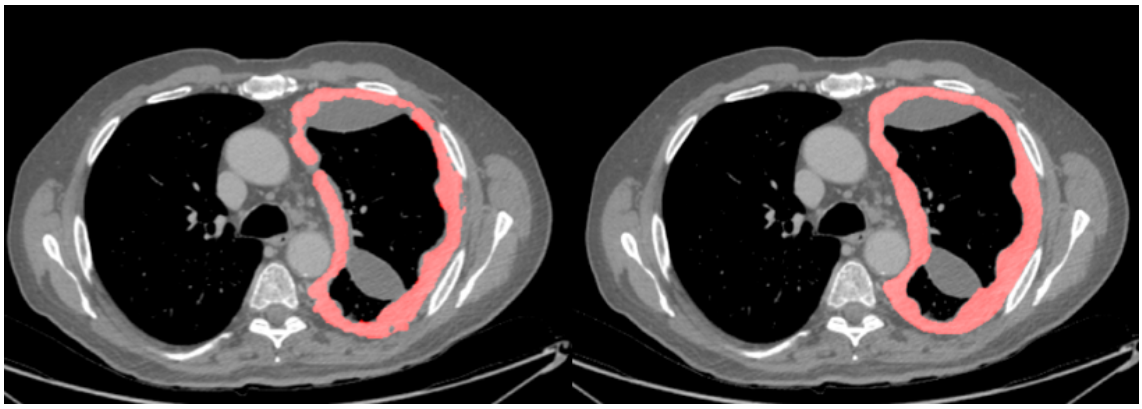


Figure 3.13: *A CT slice from a subject positive for MPM. Top: Image overlaid with the ground truth segmentation (in red). Bottom: The corresponding predicted segmentation from one of the seven-fold models. Figure from [25].*

3.6.4 False Positive Rate Estimation

Using the ensembled algorithm, prediction time increased to around 120 seconds per image, using an Nvidia 1080Ti graphics processing unit (GPU), 32 GB of RAM and a 12-core Intel Xeon CPU (3.40 GHz).

Comparison to MPM positive images:

For the NLST dataset, which should contain little or no mesothelioma, in the vast majority of images the automated detector segmented very little. Figure 3.14 shows the predicted volumes for the NLST hard kernel images, together with the mesothelioma positive volumes from the DIAPHRAM and PRISM studies. The average volume measurement from the hard kernel NLST images is 3.6 cm^3 (standard deviation 6.5 cm^3). In contrast the mean automated volume measurement in the DIAPHRAGM and PRISM datasets was 547.2 cm^3 (standard deviation 290.9 cm^3).

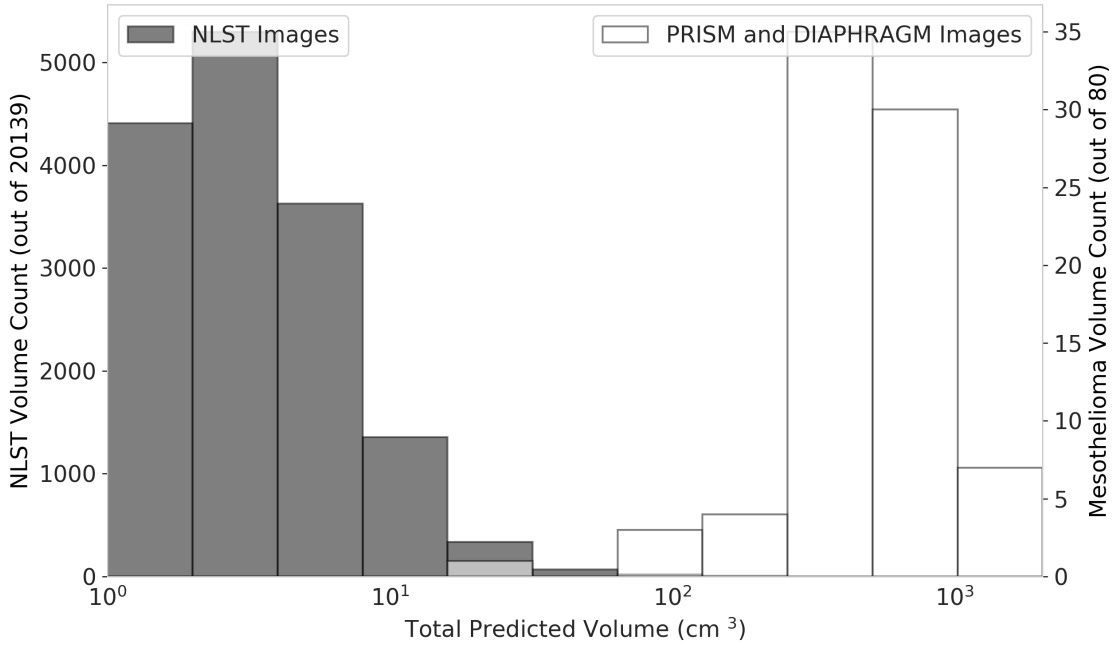


Figure 3.14: *A histogram of predicted MPM volumes across CT images from the NLST study with reference to the volume results from the mulit-fold analysis across images from the PRISM and DIAPHRAM studies. The NLST images are reconstructed using hard kernels. For the volume measurements, a logarithmic scale is used.*

Stratification by NLST finding:

Of the hard kernel images, 11,157 were finding positive and 8,982 were finding negative (see table 3.1 for details of the positive and negative groups).

Across the hard kernel image reconstructed images there was a small but significant ($p < 0.001$) difference between the algorithm predictions for finding positive and finding negative images. The mean segmented volume in finding negative images was 2.9 cm^3 (s.d. 3.4 cm^3, median 2.0 cm^3). For the finding positive images this increased to 4.1 cm^3 (s.d. 8.2 cm^3, median 2.2 cm^3). Given that the finding positive subjects may have up to two time-points where no pathology was present, as the pathology finding is per subject rather than per image, an overlap of the groups is to be expected. Figure 3.15a shows predicted volumes for the NLST datasets, stratified by whether the image was graded as finding negative or finding positive.

Effect of reconstruction kernel:

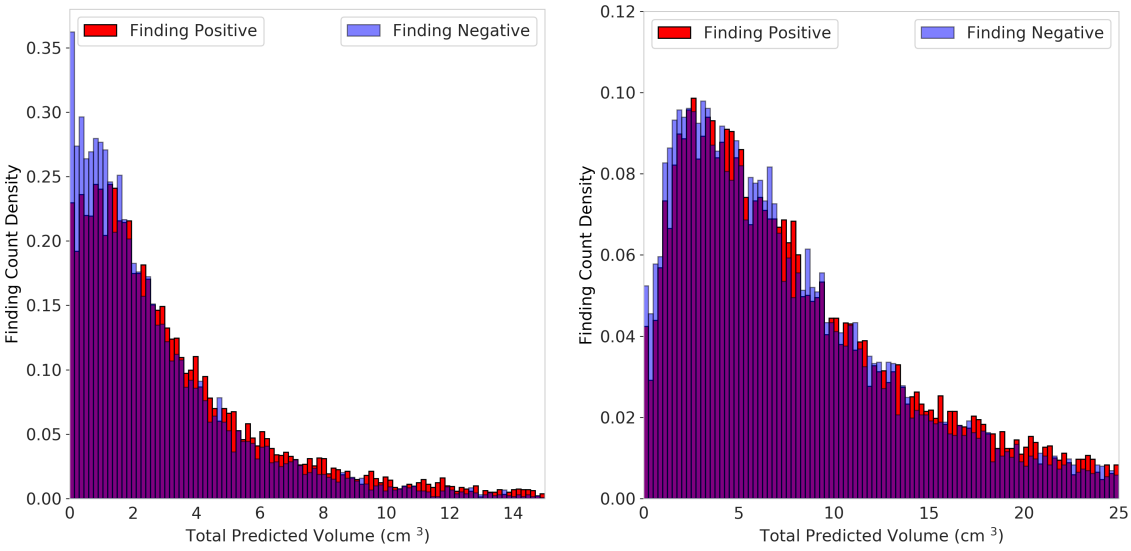
Of the soft kernel images, 14,763 were finding positive and 11,711 were finding negative.

For the soft kernel reconstructions the mean detected volume was 10.1 cm^3 (s.d. 13.8 cm^3), an increase compared with the mean volume of 3.6 cm^3 for the hard kernels. Figure 3.16 shows the distribution of detected volumes for all the images reconstructed with the soft and hard kernels. Figure 3.17 shows a direct comparison between the hard and soft kernel image segmentations, where the softer kernel results in a thicker segmentation. The figure also provides an example of how the appearance of the images may change between the reconstruction kernel used.

Using the soft kernel images, differentiation by bright pathology finding is less clear. The mean segmented volume in finding negative images was 9.0 cm^3 (s.d. 8.8 cm^3, median 6.5 cm^3). For the finding positive images this increased to 11.0 cm^3 (s.d. 16.7 cm^3, median 7.1 cm^3). Although remaining statistically significant ($p < 0.001$), this difference is less apparent than for hard kernel images (figure 3.15b).

In general across the images, a softer kernel results in a thicker segmentation. Due to the nature of the segmented regions, any volume measurements are extremely sensitive to this thickness change. In some cases (and as shown in figure 3.16), a difference in volume arises because new regions were segmented — sometimes regions which are segmented in hard kernel images extend further in the equivalent soft kernel images. This may be due to an increased ambiguity — areas which the algorithm could differentiate in hard kernel images may be less distinguishable in soft kernel images. This may also arise because of a minimum tumour thickness

which the algorithm can segment (this is discussed further in section 3.8.1).



(a) *Hard Kernel* **(b)** *Soft Kernel*

Figure 3.15: *Comparison of predicted MPM volumes reconstructed by hard 3.15a and soft 3.15b kernels. Subjects are stratified into finding positive and finding negative. Note that different axis limits are used for the hard and soft kernel subplots.*

Observation of outliers:

Using both hard and soft kernel reconstructed images, 94/100 upper outliers were for subjects reported to have a bright pathology finding. Many images show evidence of pleural thickening and considerable pleural effusion. Examples of 9/100 outliers are provided in figure 3.18. The training data only contains unilateral examples of MPM, however it is likely the algorithm has not fit to this aspect of the data. Several of the upper outliers in figure 3.18 show subjects with pathologies in both lungs which have been identified by the algorithm. By design, the algorithm had sufficient receptive field to encompass the entire image, and had the capability to use information in one lung to guide any tumour delineation in the other, however the unilateral nature of the disease in the training data appears not to have been learned. It is likely the algorithm would generalise to measurement of bilateral examples of MPM, although such cases are exceptionally rare. The remaining 6/100 images with no reported bright pathology finding associated were abnormal. For one, the automated method segmented volume in the liver. Generally, the algorithm segmented more around the diaphragm, and this region was where false positives were most frequent. This

is a region where axial slices are particularly difficult to interpret, and where more extensive 3-D information could help disambiguate the images.

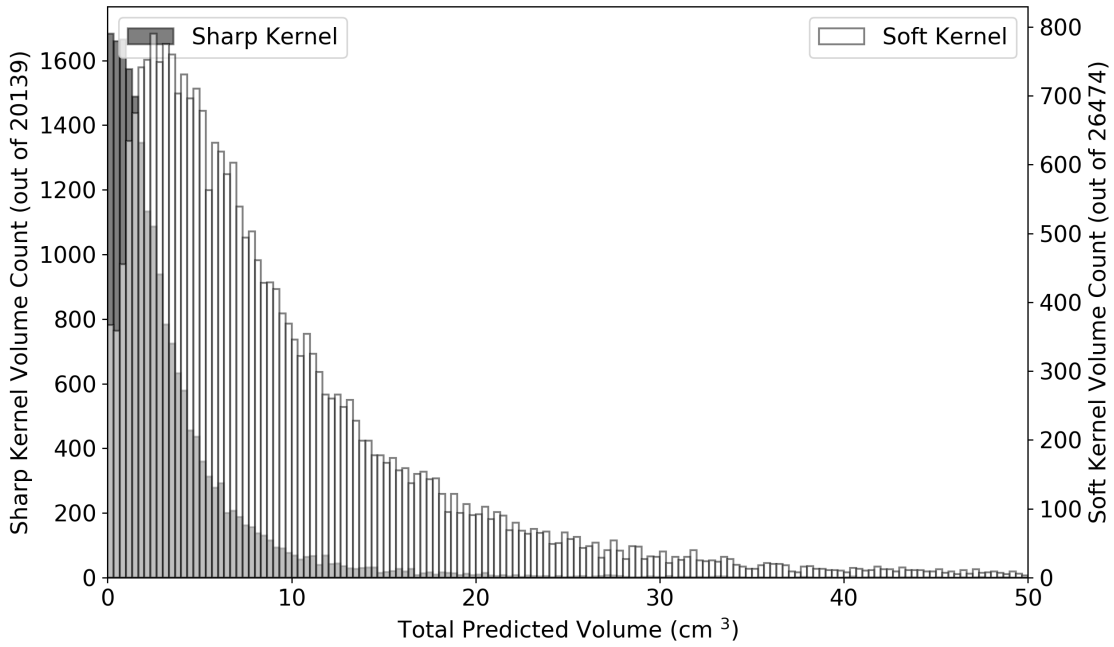


Figure 3.16: *A histogram of MPM volume predictions across the NLST dataset, stratified by hard or soft image reconstruction kernel.*

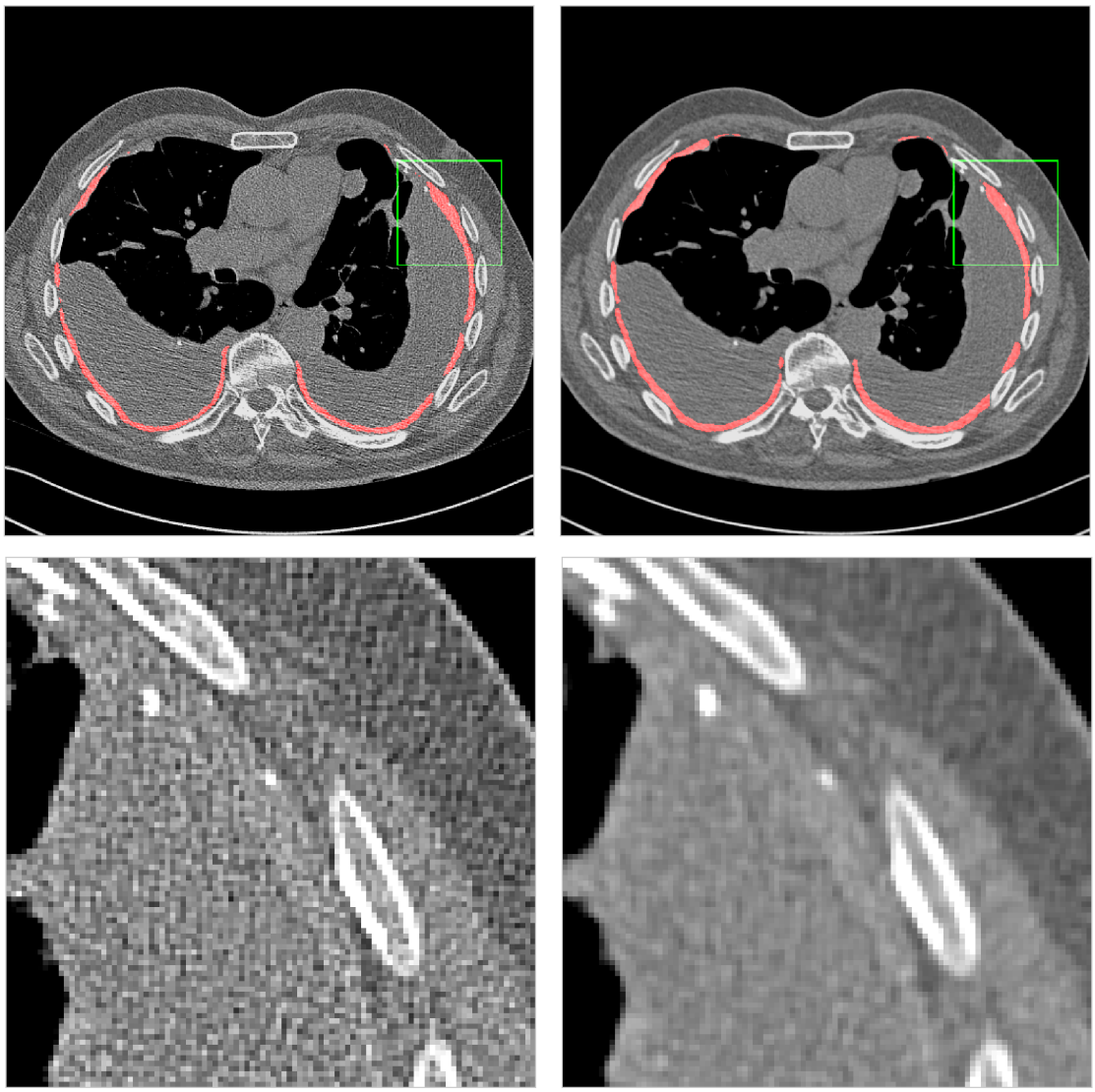


Figure 3.17: *Top row: a comparison of corresponding hard-kernel (left) and soft-kernel (right) reconstructed images from the NLST study, with an overlay the segmentation produced by one of the 7-fold models. Bottom row: A cropped region corresponding to the green box in the top row, showing the smoother appearance of the soft kernel reconstructed images.*

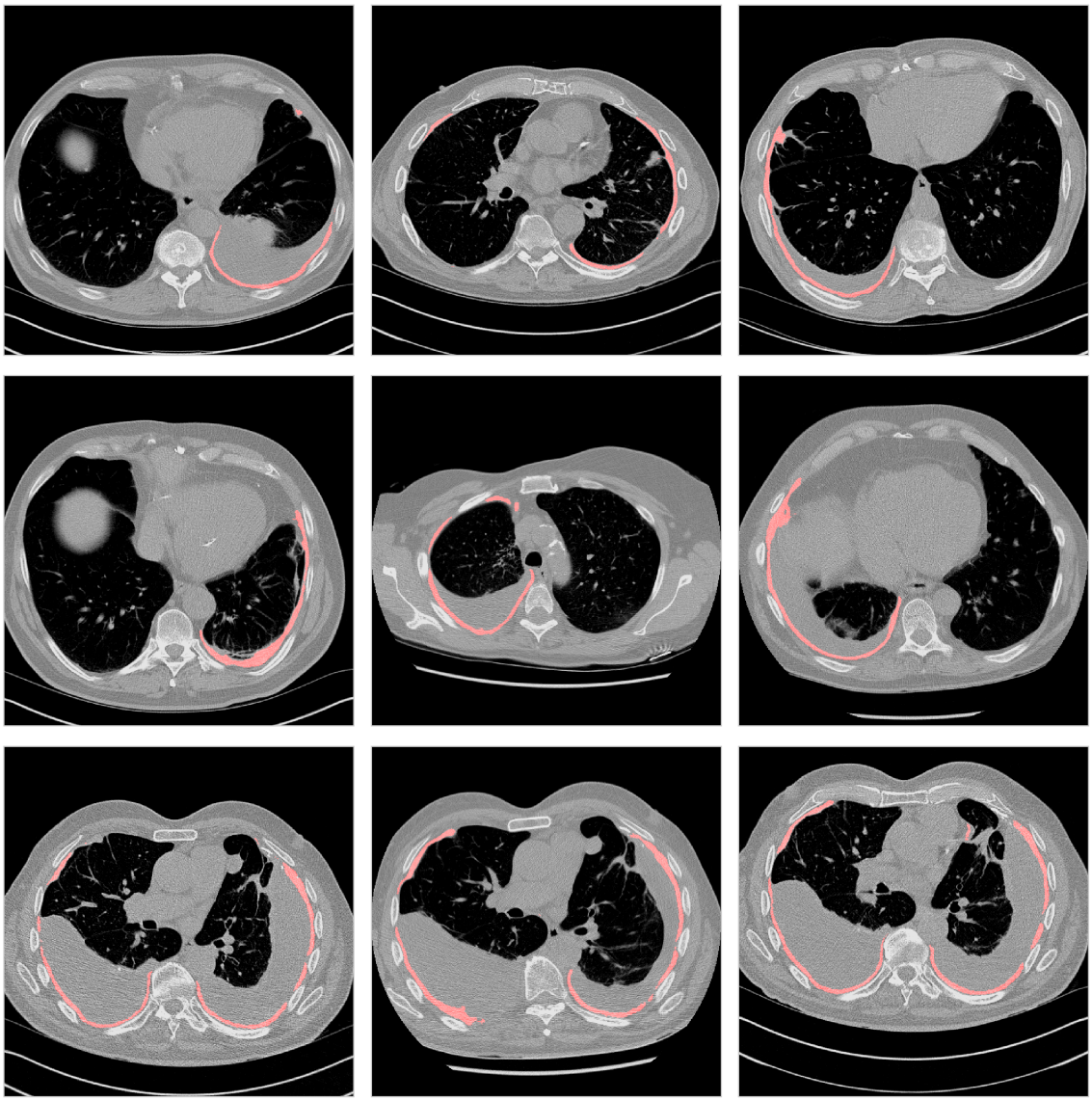


Figure 3.18: *A selection of images from the NLST study for which the algorithm predicted a relatively high volume of MPM tumour. The images are overlaid with segmentations by a random selection of the 7-fold models.*

3.7 External Validation

In this section, material is presented which is based on "Fully automated volumetric measurement of malignant pleural mesothelioma by deep learning AI: validation and comparison with modified RECIST response criteria" which was published in the journal Thorax [27]. This work was completed as a component of a multi-centre collaboration between Mesothelioma specialists based at the Queen Elizabeth University Hospital (Glasgow, Scotland) and Canon Medical Research Europe (Edinburgh, Scotland). This group of clinical collaborators examined the performance of the algorithm described in Section 3.5 on a novel, multi-centre validation set.

3.7.1 Methods

Data: The external validation set consisted of 60 CT images (pre- and post-treatment) of a total of 30 patients with MPM from Glasgow (n=10), Wythenshawe (n=10) and Leicester (n=10). These cases were selected from two multi-centre MPM boimarker studies (DIAPHRAGM, Diagnostic and Prognostic Biomarkers in the Rational Assessment of Mesothelioma [88] and PRISM, Prediction of Resistance to chemotherapy using Somatic Copy Number Variation in Mesothelioma [89]). All images were annotated using Myrian Intrasense software by two respiratory physicians with doctoral training in MPM.

Algorithm: The algorithm (previously described in Section 3.5) was developed on seven folds of analysis, and for prediction on the novel test set the results from these seven-fold algorithms were ensembled by averaging the MPM tumour volume predictions. DICE coefficients were computed for each of the seven-fold algorithms individually, and the average of these coefficients is reported.

Volumetric Response Classification: For the external validation set, tumour volume change was computed for the manual and automated MPM measurements. This was calculated as:

$$\Delta V = \frac{V_f - V_i}{V_i} \times 100, \tag{3.3}$$

where V_i and V_f are the pre- and post-treatment volume measurements respectively. Tumour volume change was grouped into three classes: Partial Response (PR), indicating a $\geq 30\%$ reduction; Progressive Disease, indicating a $\geq 20\%$ reduction; and Stable Disease (SD), for those cases which did not meet the PR and SD thresholds. These criteria are based on the mathematical modelling of Oxnard *et al.* [90], which show that the grouping by these thresholds accurately reflect those disease classifications achieved by the mRECIST scoring system. The modelling assumes tumours which are approximately crescent-shaped.

Statistical Analysis: Depending on the distribution, the median (IQR) or mean (SD) were used to summarise data. Mostly, the data was non-normally distributed, so non-parametric tests were used. For paired volume data (e.g. comparing automated and manual measurements, or pre-treatment and post-treatment measurements)

the Wilcoxon matched-pairs signed rank test was used. Correlation was assessed using the Spearman's rho test and Bland-Altman analysis was used to evaluate agreement between the automated and manual assessments. For voxel-wise tumour segmentation comparisons, region overlap was scored using the Dice coefficient (which is equivalent to the F1 score). Comparison between response classification by mRECIST scoring, manual volumetry and automated volumetry was assessed by the Cohen's kappa statistic. The Kruskal-Wallis was used to compare the differences in disease classification by the Kruskal-Wallis test, and the Dunn's test was used for multiple comparisons. Survival analysis was performed using the Kaplan-Meier methodology.

Inter-observer and intra-observer variability for manually derived tumour volume measurements was assessed by the intraclass correlation coefficient (ICC). For the interobserver comparisons, 10 randomly selected CT images from the DIAPHRAGM study were re-annotated by an independent observer. For the intraobserver comparisons, 10 images (averaging 225 slices per image) were re-annotated by the same observer after at least three weeks had passed since the primary annotation.

For the statistical tests, SPSS (V.24.0, Chicago, USA), GraphPad (V.9.1.0, San Diego, USA) and MATLAB (V.9.10, MathWorks, Natick, USA) were used.

3.7.2 Results

Fidelity to reference human annotations by region overlap Comparing manual and automated segmentation over the test set, the mean Dice coefficient was 0.55 (SD 0.12). A mean Dice coefficient of 0.54 (0.08) and 0.54 (0.16) was achieved for the two sub-cohorts of 10 CT scans used to assess interobserver and intraobserver agreement. In comparison for these CT datasets, this was higher than the agreement with a second human reader (mean DICE 0.36 (0.1), p=0.002), but lower than the agreement achieved by repeat annotation by the same reader (mean DICE 0.61 (0.09), p=0.014).

Human versus AI volumes: Manual and automated measurements of tumour volume were strongly correlated (r=0.851, p$\leq$0.0001) (shown in Figure 3.19), and Bland-Altman analysis showed a mean bias of +31 cm^3, which was not significantly different to zero (p=0.182), and 95% limits of -345 to +407 cm^3 (shown in Figure 3.19). These results were similar when computed on the sub-groups of pre- and

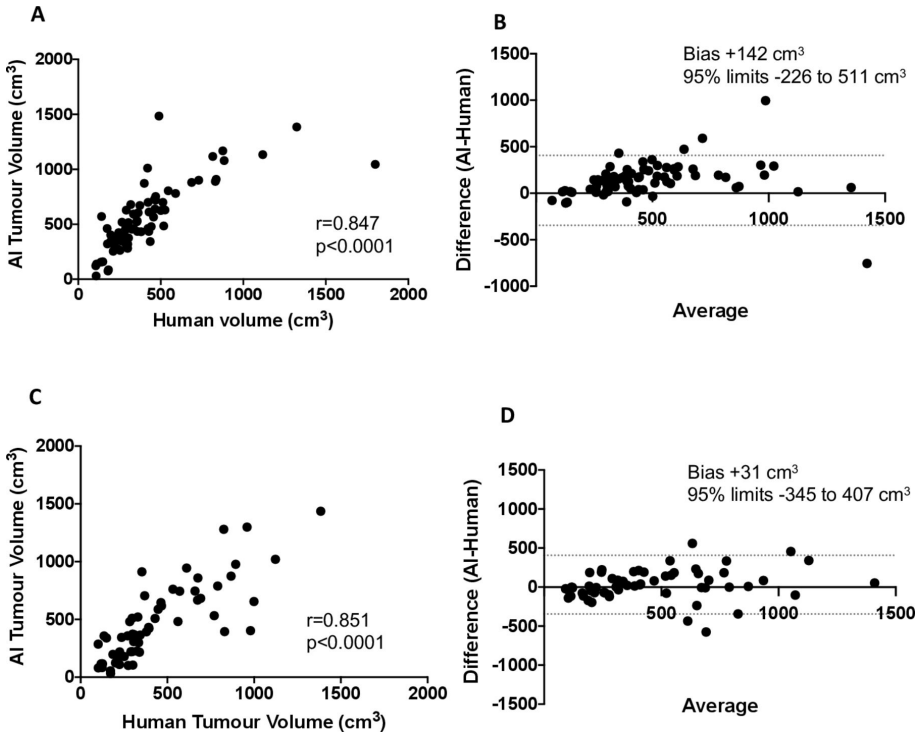


Figure 3.19: *Correlations (panels A and C) and Bland-Altman analysis (panels B and D) comparing manual and automated MPM volume measurement. Panels A and B show the results on 80 scans from the cross-validation set (which have been provided in Section 3.6). Panels C and D show the results on the 60 scans in the unseen validation set.*

post-treatment volumes. For the Bland-Altman analysis, there were four outliers. Two oversegmented scans showed, when inspected by MPM experts, one case of a mistaken inclusion (by the algorithm) of atelectatic lung overlying the hemidiaphragm (shown in Figure 3.20 B) and a second case where the algorithm segmented a region of benign pleural thickening in the tumour-free lung of a subject with unilateral disease (shown in Figure 3.20 C). Those under-segmented outliers showed failures to include fissural tumour (shown in Figure 3.20 A).

Volumetric change following chemotherapy: For volumetric change, the pre- and post- treatment volumes change was not statistically significant (by manual measurements: 366 cm^3 (244 to 656) vs 328 cm^3 (225 to 663), p=0.196; by automated measurements : 427 cm^3 (220 to 682) vs 371 cm^3 (122 to 689), p=0.081). Manual and automated assessment of volume change were closely correlated (r=0.611, p=0.0003) (shown in Figure 4A). A mean bias (automated minus manual) of +2.1% which was not significantly different to zero (p=0.425), 95% limits of agreement -59.6% to 55.5% (see figure 4B). For the volume change classification into the PR, SD and

PD categories, there was agreement between the manual and automated methods in 20/30 (67%) cases, with κ=0.439 (0.178 to 0.700) (shown in Figure 3.21 C). When response was simplified to non-PD versus PD (by the combination of the SD and PR categories), agreement increased to 26/30 (87%), κ=0.586 (0.227 to 0.945) (shown in Figure 3.21 D).

mRECIST versus AI volumetric response: Moderate agreement was achieved between volume change classification by mRECIST and the automated approach. For 16/30 (55%) cases there was agreement, κ=0.284 (0.026 to 0.543) (shown in Figure 3.22 A). When response was simplified to non-PD versus PD (by the combination of the SD and PR categories), agreement increased to 20/30 (67%), κ=0.223 (-0.128 to 0.574) (shown in Figure 3.22 B)

Survival analyses: The median survival duration for the 30 subjects in the validation cohort was 377 days (with a median follow-up of 4.7 years). There were no significant trends between survival and tumour volume change classification by the mRECIST, manual volumetry or automated volumetry approaches. However, baseline tumour volume was significantly associated with survival by both manual and automated measurements (HR 4.01 (1.67 to 9.64) p=0.0019 for manual volumetry, and HR 2.45 (1.08-5.55) p=0.010 for automated volumetry).

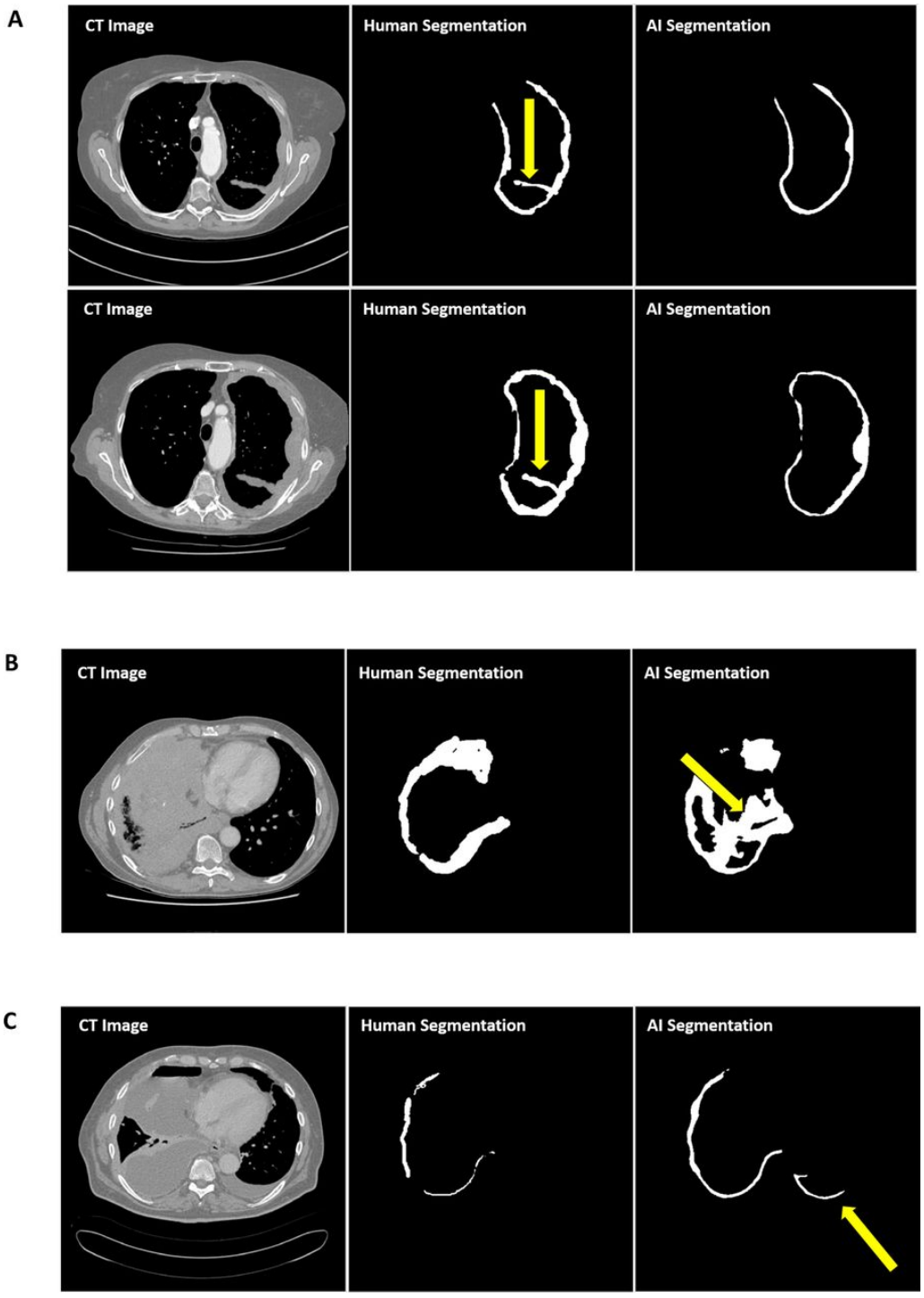


Figure 3.20: *Outliers from the Bland-Altman analysis shown in Figure 3.19. For two cases in panel A (a pre- and post- treatment image from the same subject) the algorithm undersegments a region of fissural tumour (arrow). In panel B, a case where tumour has been oversegmented by the automated approach is shown, where an area of atelectatic lung overlying the right hemidiaphragm is erroneously included (arrow). Panel C shows a case where a region of benign pleural thickening is included in the automated tumour segmentation (arrow).*

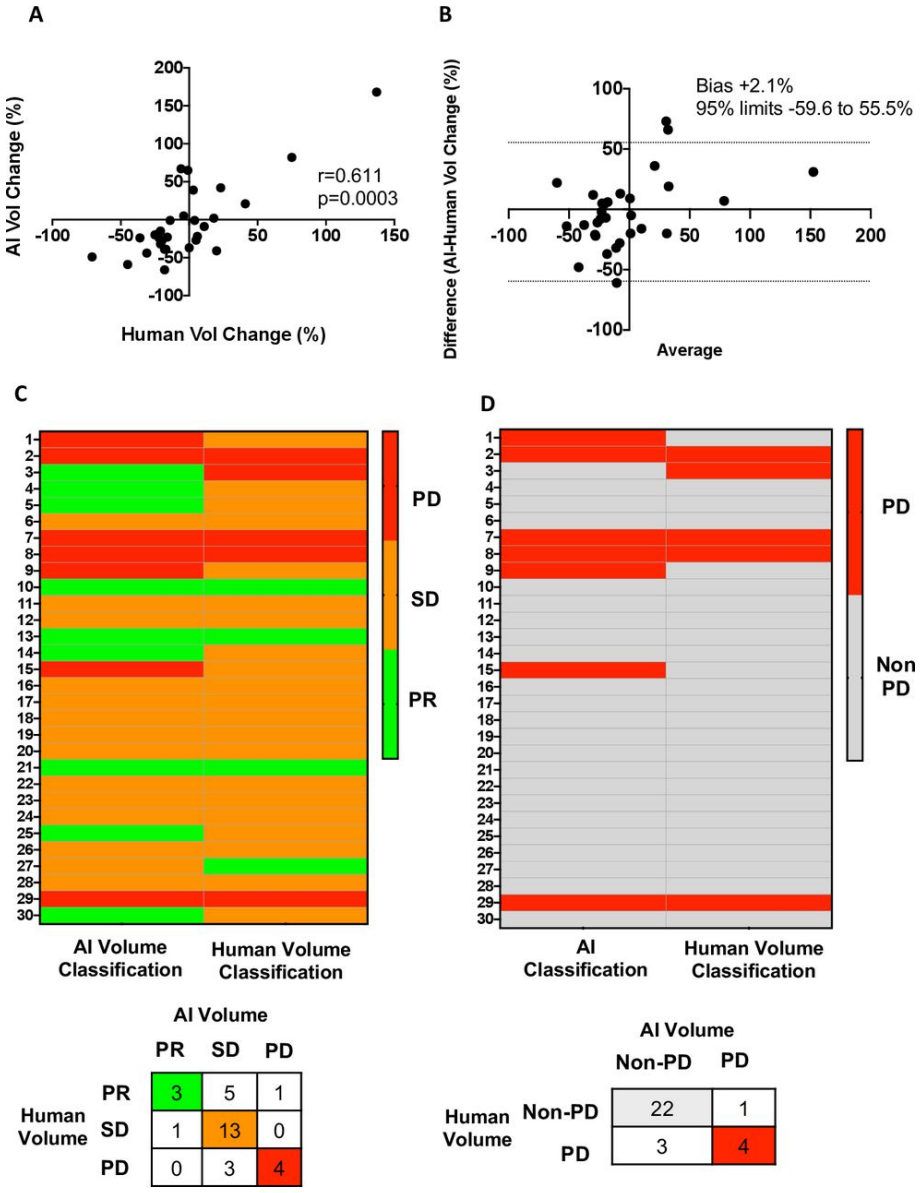


Figure 3.21: *Panel A shows the Spearman's correlation for volumetric change between the automated (AI) and manual (human) derived measurements. Panel B shows the corresponding Bland-Altman analysis. Panel C shows a confusion matrix between the automated and manual classifications, as dichotomised into Partial Response (PR), Stable Disease (SD) and Progressive Disease (PD) categories. Panel D shows a confusion matrix where the SD and PR categories have been combined.*

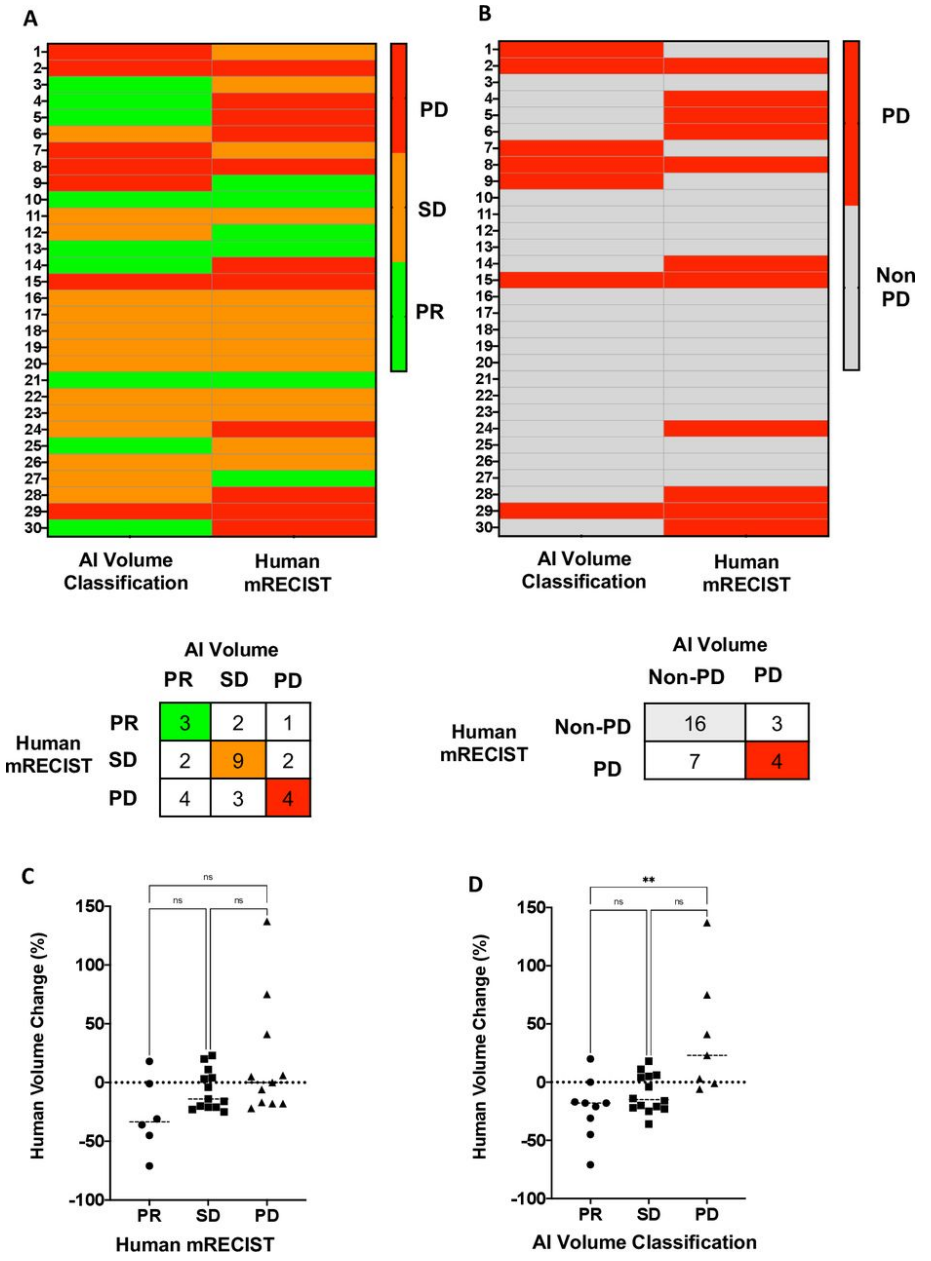


Figure 3.22: *Panel A shows a confusion matrix of tumour change classification agreement by mRECIST and the automated approach, as dichotomised into Partial Response (PR), Stable Disease (SD) and Progressive Disease (PD) categories. Panel B chows a corresponding analysis where the SD and PR categories have been combined. Panels C and D show both the mRECIST and automated volume change classifications against the gold standard measurement - manual measurements of volumetric change.*

3.8 Discussion

Although there is no curative treatment for MPM, tumour volume measurements would support clinicians to find the most effective care for each patient, and could enable more powerful clinical trials. Manual measurements of volume are too time-consuming to be routine, and still suffer from uncertainty. Some of this uncertainty arises from ambiguous features in the images — many structures appear very similar to MPM tumour in CT images. Manual measurements require significant clinical expertise to disambiguate the images, the expert uses an understanding of anatomy and experience of how the tumour develops. The distillation of such complex domain knowledge makes the application of traditional image analysis techniques complex. Such tasks, however, are where deep learning is readily applied.

Distilling expertise does not overcome the inherent uncertainty in annotating a tumour of this shape, with an unusually high surface-to-volume ratio. The large proportion of edge voxels means that any volume measurement is highly sensitive to the edge dilatation of the tumour segmentations - changing the boundary by half a voxel can change volume measurements by up to 60% (based on the analysis of tumour shapes from the DIAPHRAGM study). This poses many technical challenges — for the automated method, we have shown that in the regime of narrow segmented regions in MPM negative subjects, the choice of reconstruction kernel consistently impacts measurements.

3.8.1 Critical analysis

Generally, the literature shows significant variability in MPM tumour measurements. Sensakovic *et al.* [76] found an inter-observer mean Dice coefficient of 0.68 across slices from 31 subjects. Gudmundsson *et al.* [82] achieve a mean Dice coefficient of 0.690 on slices which are selected to contain pleural effusion. This mean Dice coefficient increases to 0.780 on a second test set, containing different disease characteristics. Over full volume images from 15 subjects, Chen *et al.* [75] achieve a Dice score of 0.825. Our mean volumetric Dice coefficients of 0.64 (by multi-fold analysis) and 0.55 (by independent testing) is lower than that achieved by Chen *et al.* Some of the difference may arise from the semi-automated nature of their approach, however on some images we achieve similarly high Dice coefficients. Across our cohort, higher Dice scores were achieved for images where the tumour was thicker — these are

images which are inherently easier to annotate, both manually and automatically, and a higher Dice coefficient is more easily achieved. Although these comparisons provide interesting context, we can only draw limited conclusions without a like-for-like comparison between methods on the same cohort.

3.8.2 False Positive Rate Estimation

Further large scale analysis across data from an independent study indicates that the algorithm is robust to the majority of negative cases. This is a one-sided analysis, and does not provide a measure of sensitivity, however analysis of outliers shows the algorithm is providing plausible output. Where the predicted volumes are highest, the algorithm confounds other bright pathologies with MPM tumour — most of the outliers are unhealthy, and many have images similar in appearance to those from MPM positive subjects. We would not expect the algorithm to be capable of distinctions between many pathologies and MPM tumour based on the images alone.

The analysis suggests that the choice of CT reconstruction kernel is significant where there is little or no MPM present. Smoother images may increase the ambiguity in delineation at the edges of the tumour, and given the algorithm has only been trained on positive cases, it is likely to be biased towards inclusion of these ambiguous regions. As mentioned in section 3.6.4, for some cases using a softer image reconstruction results in additional segmented areas, which could be due to an increased image ambiguity in these regions, or because a larger spatial extent is more likely to be detected by the algorithm. By its design, the CNN outputs smooth and continuous probability maps. After thresholding, it is unlikely that segmented regions will be narrower than a few voxels. Generally, for measurements of MPM tumour this is not a problem, however for the NLST cohort images that show pleural thickening, it is possible that a thickened pleura is thinner than the algorithm can segment. In CT images, a healthy pleura is invisible, and a thickness of even one or two voxels may be significant. Expanding a region of bright pleural thickening (or other pathological regions) in the images by using a softer kernel may slightly increase the thickness of these regions, allowing them to be detected by the algorithm. We note that inspection of several outliers in the cross-validation on MPM positive subjects did show undetected, thin tumour regions (section 3.6.2). It is possible that for these outliers, the choice of reconstruction kernel would also impact

any automated measurements. To overcome this, an algorithm which segments the images at an increased resolution may be more appropriate.

Of the images from the PRISM and DIAPRAGM studies, 107/123 were reconstructed using soft kernels. This leaves 16/123 hard kernel images, and meaningful statistics could not be derived to measure how the manual annotations were impacted by reconstruction kernel. It may be that the algorithm is biased by reconstruction kernel imbalance in the training data — it is possible that segmenting greater volumes would be measured as higher segmentation performance for subjects with known MPM. This cannot be determined by analysis of cases with no known MPM.

For the task of MPM segmentation on histologically confirmed cases, where the disease characteristics can vary dramatically between subjects, as well as between time-points and observers, performance of an algorithm depends heavily on the training and testing cohort. An increased variance between subjects means that a large and diverse test set is required to truly establish whether any automated method can generalise to unseen cases. A potential limitation of this work is that we have demonstrated the performance of the algorithm on 80 subjects which have not undergone treatment for the disease, all from imaging centres based in Glasgow, all annotated by a single observer. Images from a further 14,965 subjects from 33 different centres have provided an insight into some aspects of algorithm performance on independent images. However, to truly understand performance, more images containing MPM tumour (with ground truth segmentations) are required. We have used an unusually large cohort with full volume annotation of MPM tumour, however a large, independent and varied test set by multiple observers is still necessary to truly determine the performance of this algorithm.

3.8.3 External Validation

For the independent validation set (60 CT datasets), the mean difference between AI and human volumes was not significantly different to zero with 95% limits of -345 to +407 cm^3 (shown in Figure 3.19 C and D). Segmentation errors exceeding this interval were observed in 4/60 cases, and were related to fissural tumour, contralateral pleural thickening and adjacent lung atelectasis. These are important features of MPM, and it is possible that further expansion of the training set to include more similar cases may improve algorithm performance.

The mean bias decreased from +142 cm^3 on the internal cross-validation set to 31 cm^3 for the independent set. This is likely to reflect the inclusion of subjects from the DIAPHGRAM study in the training set — these were subjects with early stage disease, and often extremely thin tumour. It is from cases from the DIAPHRAGM set that the interobserver DICE coefficient was found to be 0.36, showing again that these cases were extremely challenging to annotate. The automated method tended towards over segmentation of the tumour for these cases. The predicted segmentations by the algorithm are extremely sensitive to the choice of the model threshold, and it is possible that an optimal threshold for one cohort (or, in this case, disease stage) is sub-optimal for another. This manifested in the necessity to denoise the ground truth for the early stage cohort, whilst the test set ground truth was more contiguous between slices due to a thicker tumour region.

Only moderate agreement was achieved for disease classification by the automated approach and the mRECIST score. This may be due to the improper selection of cut-points to classify the volumetric tumour change, as low agreement was also found between manual volumetry and the mRECIST scoring system. For all the disease classification approaches, no significant relationships were found between classification and outcome, however an association was found between the initial tumour volume and survival as assessed by manual and automated volumetry.

The Dice coefficient between observers on a sub-cohort of 10 cases was 0.36. This was a particularly challenging sub-cohort to annotate, with predominantly early stage disease. It would be beneficial to repeat this analysis on a larger number of cases, with more varied disease characteristics. Whilst over a small cohort, this relatively low agreement score has implications for the algorithm. Where varied interpretations between observers exist, it would be beneficial to evaluate the proposed approach on a set annotated by multiple observers. This would allow the extent to which the algorithm is over fit to one observers interpretation of the disease to be determined.

Chapter 4

Conclusions

I have developed and evaluated two algorithms to quantify cancers which grow in and around the lungs. The first algorithm is for the detection of lung nodules, which may precede lung cancer. The second cancer is mesothelioma.

4.1 Lung Nodule Detection by Deep Learning

I developed a novel algorithm for lung nodule detection which was of a two-stage design, and benchmarked it against two publicly available high-performing algorithms (DeepLung DPN and DeepLung ResNet-18). The evaluation was conducted in three parts: an LIDC-IDRI multi-fold analysis, and LIDC-IDRI held-out test set analysis, and an independent NLST analysis. LUNA CPM scores of 0.784, 0.807 and 0.684 were achieved for these analysis respectively and peak nodule sensitivities ranged from 84% to 91% across these analysis. The novel algorithm was found to be equivalent in performance to the benchmarks, and also comparable in performance to other works cited in the literature. Critically, the extent to which the testing data impacts performance was shown, and how the most important component of assessing performance is the data (both in terms of image qualities and ground truth) on which analysis is conducted. It is likely that the assessment on NLST is a better approximation to the real-world performance of these methods. Following this, I formulated a novel invention (published as a U.S. Patent), whereby hierarchical relationships (e.g. those that exist between lung and lung nodules) can be leveraged to increase DL algorithm performance. Analysis showed the approach qualitatively improved resultant lung segmentation, and quantitatively improved lung nodule detection performance for the first stage lung nodule detector.

Future Work

An invention to impose hierarchical relationships was presented and demonstrated on the first stage "In-House" nodule detector, however its benefit on the entire detection pipeline was not evaluated. Future work would involve a full integration and assessment of this invention within the pipeline as a whole, to ascertain a more complete assessment of both the impact to lung nodule detection and lung segmentation.

As hardware capabilities continue to improve, the maximum size of image volume presented to DL algorithms increases. It is possible that further improvements could be leveraged on the candidate proposal section of the "In-House" approach providing a full 3-D volume to the CNN. This would improve its capability to leverage further 3-D information to disambiguate confounding structures.

From a clinical perspective, some false positives are more tolerable for a CAD system to flag than others. For example, a false positive which is detected in the stomach is a more glaring error than a false positive in pathological regions of the lung, and would not promote confidence in an automated tool. The extent to which this problem exists may not be apparent in the numerical performance of the algorithm, and thus further development would be usefully informed by a review of the tool by an expert, and detailed feedback on the types of errors observed.

4.2 Mesothelioma Measurement by Deep Learning

A novel algorithm was developed to fully-automate the process of MPM measurement based on CT images. An algorithm was developed using 123 CT images containing MPM, and multi-fold analysis reported across 80 CT images showed a mean Dice coefficient of 0.64 for segmentation accuracy. Volumetric measurements were not significantly different from zero, and 95% LOAs between -417 and +363 cm^3 were achieved using de-noised ground truth. A qualitative evaluation over 14,695 subjects was conducted to ascertain algorithm robustness, and images which were identified by the algorithm showed evidence of hyperdense pathology, pleural thickening, and significant pleural effusion.

The algorithm was subsiquently tested on an independent dataset of 60 CT volumes from 30 patients, and the resultant tumour segmentation maps were reviewed

and analysed by MPM specialists. The algorithm bias for volumetric measurements remained low (+31 cm^3), and 95% LOAs of -345 to +407 cm^3 were achieved. These results were highly consistent to those achieved by multi-fold analysis. The mean Dice coefficient for segmentation accuracy was 0.55. Additional analysis showed that for a subset of 10 cases, the algorithm Dice score was superior than that of an inter-observer comparison between two human readers (0.54 versus 0.36), but inferior to repeat annotation by the same observer (0.54 versus 0.61), demonstrating that the algorithm is performing within the level of agreement that exists between independent observers for this subset of the data. Volumetric change classification was assessed, and the algorithm was in agreement with the gold-standard manually derived classification for 67% of the 30 cases.

The task of manual MPM tumour segmentation is too time consuming to be routine, and the standard mRECIST scoring system is often incapable of accurately characterising tumour development. This work shows a proof-of-concept algorithm which automates the process of full volumetric segmentation. The results are highly encouraging, showing similar agreement between the algorithm and a human expert as exists between human observers. Whilst encouraging, an independent test set of 30 cases is too small to draw any definitive conclusions on real-world clinical utility for volume change assessment, due to the high degrees of variability in disease characteristics between subjects. This work represents an important first step towards an improved ability to routinely accurately discern progression of MPM.

Future Work

Automated image analysis techniques continue to rapidly progress in performance, and a number of new techniques have been developed in the field. For MPM prediction, the necessity to marry local high-resolution prediction accuracy with a more global understanding of any present pathology is an important factor to consider. For the presented works, the algorithm consumed contextual neighbouring slices. However, in areas such as the apex of the lung, distinction between healthy and unheathly lung was not always possible based on those provided image slices, resulting in low accuracy for these regions. Simple post-processing (e.g. limiting the predicted segmentation to the largest predicted tumour regions) did not improve the scores. It is likely that a more sophisticated approach would yield improved

performance. Specifically, multi-resolution approach, or one which could consume the entire patient volume as input would be of benefit. This would allow the strong prior of uni-lateral disease to be fully leveraged, and the full 3-D shape prior of MPM to be better incorporated by the model.

The presented algorithm was developed and validated using segmentations by one expert annotator. For a disease such as MPM, where agreement between observers is low, it is likely that a level of over-fitting has occurred which has not been measured in the results — the algorithm may be strongly fit to an individual interpretation of the disease. Thus, the work would benefit from validating (and training) the algorithm on data annotated by different experts, so that the extent of this over-fitting may be determined.

This work is being continued under the PREDICT-Meso network, comprised of MPM experts from across Europe. The presented algorithm will be utilised, tested and further developed on novel data.

4.3 Overall Conclusions

Automatic approaches for the measurement of two types of cancers have been presented: lung cancer and mesothelioma. For lung cancer, I developed an algorithm for computer assisted detection of lung nodules from CT images. At a setting of 8 false positives per scan, nodule sensitivities ranged from 84% to 91% across three analyses on different data. This was competitive with other works. Following this, I developed a novel approach to leverage the related task of lung segmentation. This was shown to improve nodule sensitivity for the first stage detector, and qualitatively improved lung segmentation in pathological regions. For mesothelioma, I developed the first fully automatic segmenter of disease based on CT images. Previously, only semi-automated approaches had been conceived. On an independent test set, bias for volumetric measurements was low (+31 cm^3), and 95% LOAs of -345 to +407 cm^3 were achieved. This algorithm has the potential to greatly reduce the time and cost to accurately measure the disease, and may enable wider study into new treatments.

4.4 Future Work

Regarding computer assisted detection of lung nodules, the developed approach would require to be extensively clinically evaluated before clinical deployment. There are several ways the algorithm could be integrated into a reader assist tool, for example: candidate nodules could be presented to the radiologist for subsequent review, or those missed nodules could be flagged to the reader after their annotation has been completed. The specifics of this integration would have an impact on the sensitivity and specificity of the annotation pipeline as a whole. For the clinical evaluation to accurately capture this performance, the next step would be to integrate the algorithm into a tool to assist radiologists.

The development of a fully automatic mesothelioma segmentation algorithm opens up several avenues of further work. To date, research into novel treatments has been limited. This is partially due to the expense in determining treatment efficacy accurately. In the future, the presented algorithm may be used to provide proxy endpoints for clinical trials into novel treatments, or to facilitate larger scale retrospective analysis of individuals who have had the disease. There is also the potential for this tool to be employed in routine care, to estimate tumour volume

or tumour volume change. These quantities are related to to the outcome of the disease, and could be used for more accurate staging of patients. Importantly, the algorithm sets a benchmark for future development of automatic approaches. This work is being undertaken as a part of the PREDICT-Meso network.

Bibliography

[1] CAITLYN ALLBURY et al. *Advanced Anatomy*. Second Edition. British Columbia/Yukon Open Authoring Platform, 2018.

[2] Milena Sant et al. "EUROCARE-4. Survival of cancer patients diagnosed in 1995–1999. Results and commentary". In: *European Journal of Cancer* 45.6 (2009), pp. 931–991. ISSN: 0959-8049. DOI: https://doi.org/10.1016/j.ejca.2008.11.018. URL: https://www.sciencedirect.com/science/article/pii/S095980490800926X.

[3] Matthew B. Schabath and Michele L. Cote. "Cancer Progress and Priorities: Lung Cancer". In: *Cancer Epidemiology, Biomarkers & Prevention* 28.10 (Oct. 2019). _eprint: https://aacrjournals.org/cebp/article-pdf/28/10/1563/2285802/1563.pdf, pp. 1563–1579. ISSN: 1055-9965. DOI: 10.1158/1055-9965.EPI-19-0221. URL: https://doi.org/10.1158/1055-9965.EPI-19-0221.

[4] S. Tsim et al. *Staging of non-small cell lung cancer (NSCLC): A review*. Dec. 2010. DOI: 10.1016/j.rmed.2010.08.005.

[5] Cesare Gridelli et al. "Non-small-cell lung cancer". In: *Nature Reviews Disease Primers* 1 (1 2015), p. 15009. ISSN: 2056-676X. DOI: 10.1038/nrdp.2015.9. URL: https://doi.org/10.1038/nrdp.2015.9.

[6] Brett W. Carter et al. "Revisions to the TNM staging of lung cancer: Rationale, significance, and clinical application". In: *Radiographics* 38 (2 Mar. 2018), pp. 374–391. ISSN: 15271323. DOI: 10.1148/rg.2018170081.

[7] E A Eisenhauer et al. "New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1)". In: *European Journal of Cancer* 45.2 (2009), pp. 228–247. ISSN: 09598049. DOI: 10.1016/j.ejca.2008.10.026. arXiv: NIHMS150003. URL: http://dx.doi.org/10.1016/j.ejca.2008.10.026.

[8] Lu Peng et al. “A Multiscale Mathematical Model of Tumour Invasive Growth”. In: *Bulletin of Mathematical Biology* 79 (3 Mar. 2017), pp. 389–429. ISSN: 15229602. DOI: 10.1007/s11538-016-0237-2.

[9] Konstantinos Zarogoulidis et al. *Treatment of non-small cell lung cancer (NSCLC)*. 2013. DOI: 10.3978/j.issn.2072-1439.2013.07.10.

[10] *Systemic anti-cancer therapy for advanced non-small-cell lung cancer: treatment options*. National Institute for Heath Care and Excellence (NICE), 2022.

[11] Boudjelal Abdelwahhab, Z Messali, and Abderrahim Elmoataz. “A Novel Kernel-Based RegularizationTechnique for PET Image Reconstruction”. In: *Technologies* 5 (Dec. 2017), p. 37. DOI: 10.3390/technologies5020037.

[12] Christoph Wald et al. *ACR Data Science Institute AI Central*. 2022. URL: https://aicentral.acrdsi.org/.

[13] Matthew Diamond. *Proposed Regulatory Framework for Modifications to Artificial Intelligence / Machine Learning (AI/ML)-based Software as a Medical Device (SAMD)*. Food and Drug Administration (FDA), Feb. 2020.

[14] Bjarne L. Nørgaard et al. “Diagnostic performance of noninvasive fractional flow reserve derived from coronary computed tomography angiography in suspected coronary artery disease: The NXT trial (Analysis of Coronary Blood Flow Using CT Angiography: Next Steps)”. In: *Journal of the American College of Cardiology* 63 (12 Apr. 2014), pp. 1145–1155. ISSN: 15583597. DOI: 10.1016/j.jacc.2013.11.043.

[15] Neil Hewitt, Paul Dimmock, and Jae Long. *HeartFlow FFRCT for estimating fractional flow reserve from coronary CT angiography*. 2021.

[16] Y Le Cun et al. “Handwritten Digit Recognition with a Back-Propagation Network”. In: *Advances in Neural Information Processing Systems 2*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1990, pp. 396–404. ISBN: 1558601007.

[17] Stan Benjamens, Pranavsingh Dhunnoo, and Bertalan Meskó. “The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database”. In: *npj Digital Medicine* 3.1 (Dec. 2020). ISSN: 23986352. DOI: 10.1038/s41746-020-00324-0.

[18] William J. Dally et al. “Hardware-Enabled Artificial Intelligence”. In: *2018 IEEE Symposium on VLSI Circuits*. 2018, pp. 3–6. DOI: 10.1109/VLSIC.2018.8502368.

[19] F Rosenblatt. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Cornell Aeronautical Laboratory. Report no. VG-1196-G-8. Spartan Books, 1962. URL: https://books.google.ca/books?id=7FhRAAAAMAAJ.

[20] Michael A. Nielsen. *Neural Networks and Deep Learning*. Determination Press, 2015. URL: http://neuralnetworksanddeeplearning.com.

[21] Lu Lu et al. “Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators”. In: *Nature Machine Intelligence* 3.3 (Mar. 2021), pp. 218–229. ISSN: 2522-5839. DOI: 10.1038/s42256-021-00302-5. URL: https://doi.org/10.1038/s42256-021-00302-5.

[22] Yann LeCun et al. “Gradient-Based Learning Applied to Document Recognition”. en. In: (1998).

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: (2015), pp. 1–8. ISSN: 16113349. DOI: 10.1007/978-3-319-24574-4_28. arXiv: 1505.04597. URL: http://arxiv.org/abs/1505.04597.

[24] Owen Anderson et al. “MODEL TRAINING APPARATUS AND METHOD”. 2022.

[25] Owen Anderson et al. “Fully automated volumetric measurement of malignant pleural mesothelioma from computed tomography images by deep learning: Preliminary results of an internal validation”. In: *BIOIMAGING 2020 - 7th International Conference on Bioimaging, Proceedings; Part of 13th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2020*. 2020. ISBN: 9789897583988. DOI: 10.5220/0008976100640073.

[26] Owen Anderson et al. “Estimating the False Positive Prediction Rate in Automated Volumetric Measurements of Malignant Pleural Mesothelioma”. In: ed. by Filipe et al. Springer International Publishing, 2021, pp. 116–139. ISBN: 978-3-030-72379-8.

[27] Andrew C Kidd et al. “Fully automated volumetric measurement of malignant pleural mesothelioma by deep learning AI: validation and comparison with modified RECIST response criteria”. In: *Thorax* 77 (12 2022), pp. 1251–1259. ISSN: 0040-6376. DOI: 10.1136/thoraxjnl-2021-217808. URL: https://thorax.bmj.com/content/77/12/1251.

[28] *What is a Lung Nodule?* American Thoracic Society, 2016. URL: https://www.nlm.nih.gov/medlineplus/ency/article/000071.htm.

[29] Hamid Mithoowani and Michela Febbraro. *Non-Small-Cell Lung Cancer in 2022: A Review for General Practitioners in Oncology*. Mar. 2022. DOI: 10.3390/curroncol29030150.

[30] Carola A. Van Iersel et al. “Risk-based selection from the general population in a screening trial: Selection criteria, recruitment and power for the Dutch-Belgian randomised lung cancer multi-slice CT screening trial (NELSON)”. In: *International Journal of Cancer* 120 (4 Feb. 2007), pp. 868–874. ISSN: 00207136. DOI: 10.1002/ijc.22134.

[31] Simran Randhawa et al. “Lung Cancer Screening in the Community Setting: Challenges for Adoption”. In: *The American Surgeon* 84 (9 2018), pp. 1415–1421. DOI: 10.1177/000313481808400942. URL: https://doi.org/10.1177/000313481808400942.

[32] Samuel G Armato et al. “The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans”. In: *Medical Physics* 38 (2 Jan. 2011), pp. 915–931. ISSN: 00942405. DOI: 10.1118/1.3528204. URL: http://www.ncbi.nlm.nih.gov/pubmed/21452728%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3041807%20http://doi.wiley.com/10.1118/1.3528204.

[33] Bram van Ginneken et al. “Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: The ANODE09 study”. In: *Medical Image Analysis* 14 (6 2010), pp. 707–722. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2010.05.005. URL: https://www.sciencedirect.com/science/article/pii/S1361841510000587.

[34] Arnaud Arindra Adiyoso Setio et al. "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge". In: *Medical Image Analysis* 42 (2017), pp. 1–13. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2017.06.015. URL: https://www.sciencedirect.com/science/article/pii/S1361841517301020.

[35] DR Aberle et al. "Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening". In: *New England Journal of Medicine* 365 (5 Aug. 2011), pp. 395–409. ISSN: 0028-4793. DOI: 10.1056/NEJMoa1102873. URL: http://www.nejm.org/doi/10.1056/NEJMoa1102873.

[36] Lea Marie Pehrson, Michael Bachmann Nielsen, and Carsten Ammitzbøl Lauridsen. *Automatic pulmonary nodule detection applying deep learning or machine learning algorithms to the LIDC-IDRI database: A systematic review.* Mar. 2019. DOI: 10.3390/diagnostics9010029.

[37] Nikos Sourlos et al. *Possible Bias in Supervised Deep Learning Algorithms for CT Lung Nodule Detection and Classification.* Aug. 2022. DOI: 10.3390/cancers14163867.

[38] Nasrullah Nasrullah et al. "Automated lung nodule detection and classification using deep learning combined with multiple strategies". In: *Sensors (Switzerland)* 19 (17 Sept. 2019). ISSN: 14248220. DOI: 10.3390/s19173722.

[39] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *ICLR* (Dec. 2015).

[40] Jia Deng et al. "ImageNet: A large-scale hierarchical image database". In: 2009. ISBN: 978-1-4244-3992-8. DOI: 10.1109/CVPRW.2009.5206848.

[41] Yarin Gal and Zoubin Ghahramani. "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning". In: ed. by Maria Florina Balcan and Kilian Q Weinberger. Vol. 48. PMLR, Dec. 2016, pp. 1050–1059. URL: https://proceedings.mlr.press/v48/gal16.html.

[42] Sergey Ioffe and Christian Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *ArXiv* abs/1502.03167 (2015).

[43] Tsung-Yi Lin et al. “Focal Loss for Dense Object Detection”. In: *International Conference on Computer Vision (ICCV)* (2017).

[44] Guy Hacohen and Daphna Weinshall. “On The Power of Curriculum Learning in Training Deep Networks”. In: (Apr. 2019). URL: http://arxiv.org/abs/1904.03626.

[45] Wentao Zhu et al. “DeepLung: Deep 3D Dual Path Nets for Automated Pulmonary Nodule Detection and Classification”. In: . *IEEE Winter Conference on Applications of Computer Vision (WACV)* (Jan. 2018). URL: http://arxiv.org/abs/1801.09555.

[46] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016). URL: http://image-net.org/challenges/LSVRC/2015/.

[47] Shaoqing Ren et al. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. In: *Int. Conf. Neural Information Processing Systems (NIPS)* (June 2015). URL: http://arxiv.org/abs/1506.01497.

[48] Yunpeng Chen et al. “Dual Path Networks”. In: (July 2017). URL: http://arxiv.org/abs/1707.01629.

[49] Ruigang Fu et al. “CNN with coarse-to-fine layer for hierarchical classification”. In: *IET Computer Vision* 12.6 (2018). _eprint: https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/iet-cvi.2017.0636, pp. 892–899. DOI: https://doi.org/10.1049/iet-cvi.2017.0636. URL: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-cvi.2017.0636.

[50] Zhicheng Yan et al. “HD-CNN: Hierarchical Deep Convolutional Neural Networks for Large Scale Visual Recognition”. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 2740–2748. DOI: 10.1109/ICCV.2015.314.

[51] Yu Gu et al. “Automatic lung nodule detection using a 3D deep convolutional neural network combined with a multi-scale prediction strategy in chest CTs”. In: *Computers in Biology and Medicine* 103 (2018), pp. 220–231. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2018.10.011. URL: https://www.sciencedirect.com/science/article/pii/S001048251830310X.

[52] Jing Gong et al. “Automatic detection of pulmonary nodules in CT images by incorporating 3D tensor filtering with local image feature analysis”. In: *Physica Medica: European Journal of Medical Physics* 46 (Feb. 2018). doi: 10.1016/j.ejmp.2018.01.019, pp. 124–133. ISSN: 1120-1797. DOI: 10.1016/j.ejmp.2018.01.019. URL: https://doi.org/10.1016/j.ejmp.2018.01.019.

[53] Bin Wang et al. *Automated Pulmonary Nodule Detection: High Sensitivity with Few Candidates: 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part II*. Sept. 2018. DOI: 10.1007/978-3-030-00934-2_84.

[54] W E Cooke. “FIBROSIS OF THE LUNGS DUE TO THE INHALATION OF ASBESTOS DUST”. In: *BMJ* 2.3317 (1924), pp. 140–147. ISSN: 0007-1447. DOI: 10.1136/bmj.2.3317.147. URL: https://www.bmj.com/content/2/3317/147.

[55] M A Riva et al. “Mesothelioma and asbestos, fifty years of evidence: Chris Wagner and the contribution of the Italian occupational medicine community”. In: *La Medicina del lavoro* 101.6 (2010), pp. 409–415. ISSN: 0025-7818. URL: http://europepmc.org/abstract/MED/21141345.

[56] Simone Pollastri et al. “The crystal structure of mineral fibres. 3. Actinolite asbestos”. In: *Periodico Di Mineralogia* 86 (2017), pp. 89–98.

[57] GOV.UK. *Asbestos: general information*. URL: https://www.gov.uk/government/publications/asbestos-properties-incident-management-and-toxicology/asbestos-general-information.

[58] Frances Perraudin. *700 English schools reported over asbestos safety concerns*. 2019. URL: https://www.theguardian.com/education/2019/jul/04/700-english-schools-reported-over-asbestos-safety-concerns.

[59] Lucy Darnton. *Asbestosis, mesothelioma,asbestos related lung cancer and non-malignant pleural disease in Great Britain 2021*. 2021.

[60] Oluf Dimitri Røe and Giulia Maria Stella. “Malignant Pleural Mesothelioma: History, Controversy, and Future of a Man-Made Epidemic”. In: *Asbestos and Mesothelioma*. Ed. by Joseph R Testa. Cham: Springer International Publishing, 2017, pp. 73–101. ISBN: 978-3-319-53560-9. DOI: 10.1007/978-3-319-53560-9_4. URL: https://doi.org/10.1007/978-3-319-53560-9_4.

[61] S Luo et al. “Asbestos related diseases from environmental exposure to crocidolite in Da-yao, China. I. Review of exposure and epidemiological data”. In: *Occupational and Environmental Medicine* 60.1 (2003), pp. 35–42. ISSN: 1351-0711. DOI: 10.1136/oem.60.1.35. URL: https://oem.bmj.com/content/60/1/35.

[62] TUSHAR KANT JOSHI, UTTPAL B BHUVA, and PRIYANKA KATOCH. “Asbestos Ban in India”. In: *Annals of the New York Academy of Sciences* 1076.1 (2006), pp. 292–308. DOI: https://doi.org/10.1196/annals.1371.072. URL: https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1196/annals.1371.072.

[63] Robert Virta. *Mineral Commodity Summaries 2002*. ENGLISH. Tech. rep. 2002. DOI: 10.3133/mineral2002. URL: http://pubs.er.usgs.gov/publication/mineral2002.

[64] Giovanni Gaudino, Jiaming Xue, and Haining Yang. “How asbestos and other fibers cause mesothelioma”. In: *Translational Lung Cancer Research* 9.Suppl 1 (Feb. 2020), S39–S46. ISSN: 2218-6751. DOI: 10.21037/tlcr.2020.02.01. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7082251/ (visited on 09/13/2023).

[65] Jennifer Faig et al. “Changing pattern in malignant mesothelioma survival”. In: *Translational Oncology* 8 (1 2015), pp. 35–39. ISSN: 19365233. DOI: 10.1016/j.tranon.2014.12.002.

[66] David Michael Abbott et al. “Malignant Pleural Mesothelioma: Genetic and Microenviromental Heterogeneity as an Unexpected Reading Frame and Therapeutic Challenge”. In: *Cancers* 12.5 (2020). ISSN: 2072-6694. DOI: 10.3390/cancers12051186. URL: https://www.mdpi.com/2072-6694/12/5/1186.

[67] Ranjit K Goudar. “Review of pemetrexed in combination with cisplatin for the treatment of malignant pleural mesothelioma”. In: *Therapeutics and Clinical Risk Management* 4 (1 2008), pp. 205–211.

[68] M J Byrne and A K Nowak. “Modified RECIST criteria for assessment of response in malignant pleural mesothelioma”. In: *Annals of Oncology* (2004). ISSN: 09237534. DOI: 10.1093/annonc/mdh059.

[69] Rob J. Van Klaveren et al. “Inadequacy of the RECIST criteria for response evaluation in patients with malignant pleural mesothelioma”. In: *Lung Cancer* (2004). ISSN: 01695002. DOI: 10.1016/S0169-5002(03)00292-7.

[70] Soon Ho Yoon et al. “Observer variability in RECIST-based tumour burden measurements: A meta-analysis”. In: *European Journal of Cancer* 53 (2016), pp. 5–15. ISSN: 18790852. DOI: 10.1016/j.ejca.2015.10.014. URL: http://dx.doi.org/10.1016/j.ejca.2015.10.014.

[71] T Frauenfelder et al. “Volumetry: An alternative to assess therapy response for malignant pleural mesothelioma?” In: *European Respiratory Journal* 38.1 (2011), pp. 162–168. ISSN: 09031936. DOI: 10.1183/09031936.00146110.

[72] Zacariah E. Labby et al. “Variability of tumor area measurements for response assessment in malignant pleural mesothelioma”. In: *Medical Physics* (2013). ISSN: 00942405. DOI: 10.1118/1.4810940.

[73] Eyjolfur Gudmundsson, Christopher M. Straus, and Samuel G. Armato. “Deep convolutional neural networks for the automated segmentation of malignant pleural mesothelioma on computed tomography scans”. In: *Journal of Medical Imaging* (2018). ISSN: 2329-4310. DOI: 10.1117/1.jmi.5.3.034503.

[74] E. Gudmundsson et al. “P1.06-04 Deep Learning-Based Segmentation of Mesothelioma on CT Scans: Application to Patient Scans Exhibiting Pleural Effusion”. In: *Journal of Thoracic Oncology* (2019). ISSN: 15560864. DOI: 10.1016/j.jtho.2019.08.991.

[75] Mitchell Chen et al. “Computer-aided volumetric assessment of malignant pleural mesothelioma on CT using a random walk-based method”. In: *International Journal of Computer Assisted Radiology and Surgery* 12.4 (2017), pp. 529–538. ISSN: 18616429. DOI: 10.1007/s11548-016-1511-3.

[76] William F Sensakovic et al. “Computerized segmentation and measurement of malignant pleural mesothelioma”. In: *Medical Physics* 38.1 (2011), pp. 238–244. ISSN: 00942405. DOI: 10.1118/1.3525836.

[77] Wael Brahim et al. “Malignant pleural mesothelioma segmentation for photodynamic therapy planning”. In: *Computerized Medical Imaging and Graphics* (2018). ISSN: 18790771. DOI: 10.1016/j.compmedimag.2017.05.006.

[78] K.G. Blyth et al. “An update regarding the Prediction of ResIstance to chemotherapy using Somatic copy number variation in Mesothelioma (PRISM) study”. In: *Lung Cancer* (2018). ISSN: 01695002. DOI: 10.1016/s0169-5002(18)30090-4.

[79] Selina Tsim et al. “Diagnostic and Prognostic Biomarkers in the Rational Assessment of Mesothelioma (DIAPHRAGM) study: Protocol of a prospective, multicentre, observational study”. In: *BMJ Open* (2016). ISSN: 20446055. DOI: 10.1136/bmjopen-2016-013324.

[80] National Lung Screening Trial Research Team. “The national lung screening trial: Overview and study design”. In: *Radiology* (2011). ISSN: 15271315. DOI: 10.1148/radiol.10091808.

[81] J. Martin Bland and Douglas G. Altman. “Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement”. In: *The Lancet* (1986). ISSN: 01406736. DOI: 10.1016/S0140-6736(86)90837-8.

[82] Eyjólfur Gudmundsson et al. “Deep learning-based segmentation of malignant pleural mesothelioma tumor on computed tomography scans: application to scans demonstrating pleural effusion”. In: *Journal of Medical Imaging* (2020). ISSN: 2329-4310. DOI: 10.1117/1.jmi.7.1.012705.

[83] Jia Deng et al. “ImageNet: A large-scale hierarchical image database”. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009. ISBN: 978-1-4244-3992-8. DOI: 10.1109/CVPRW.2009.5206848.

[84] Nitish Srivastava et al. “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”. In: *Journal of Machine Learning Research* 15 (2014), pp. 1929–1958. ISSN: 15337928. DOI: 10.1214/12-AOS1000. arXiv: 1102.4807.

[85] S Ioffe and C Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift”. In: *Proc. Int. Conf on Machine Learning (ICML)*. Vol. 37. Proceedings of Machine Learning Research. 2015, pp. 448–456. URL: http://proceedings.mlr.press/v37/ioffe15.html.

[86] Francois Chollet. *Keras*. 2015. URL: https://keras.io/.

[87] Leslie N. Smith. “Cyclical learning rates for training neural networks”. In: *IEEE Winter Conference on Applications of Computer Vision*. 2017. ISBN: 9781509048229. DOI: 10.1109/WACV.2017.58.

[88] Selina Tsim et al. “Serum Proteomics and Plasma Fibulin-3 in Differentiation of Mesothelioma From Asbestos-Exposed Controls and Patients With Other Pleural Diseases”. In: *Journal of Thoracic Oncology* 16 (10 Oct. 2021). doi: 10.1016/j.jtho.2021.05.018, pp. 1705–1717. ISSN: 1556-0864. DOI: 10.1016/j.jtho.2021.05.018. URL: https://doi.org/10.1016/j.jtho.2021.05.018.

[89] K G Blyth et al. “6An update regarding the Prediction of ResIstance to chemotherapy using Somatic copy number variation in Mesothelioma (PRISM) study”. In: *Lung Cancer* 115 (2018), S26–S27. ISSN: 0169-5002. DOI: https://doi.org/10.1016/S0169-5002(18)30090-4. URL: https://www.sciencedirect.com/science/article/pii/S0169500218300904.

[90] Samuel G Armato III et al. “Measurement of mesothelioma on thoracic CT scans: A comparison of manual and computer-assisted techniques”. In: *Medical Physics* 31 (5 May 2004). https://doi.org/10.1118/1.1688211, pp. 1105–1115. ISSN: 0094-2405. DOI: https://doi.org/10.1118/1.1688211. URL: https://doi.org/10.1118/1.1688211.