Leveraging Machine Learning for Automated Abdominal Ultrasound Scanning

Alistair Lawley

Department of Electronic and Electrical Engineering

University of Strathclyde

A thesis submitted for the degree of

Doctor of Engineering

September 2023

Copyright

Signed: 

Date: 25/September/2023

Acknowledgements

I would like to express my deepest gratitude to my Supervisors Gordon Dobie and Keven Worrall, as well as Rory Hampson for sharing in the many frustrations of this project, easing my transition from clinical worker to engineer and researcher. Canon Medical Systems for their sponsorship of the engineering doctorate, James Mathews, Donald Stewart, and Marco Razeto for their support throughout my time with Canon. Many thanks to FUSE CDT, Glasgow University and University of Strathclyde: Department of Biological Engineering for their use of medical ultrasound equipment and support rendered. I also offer my thanks to the Theil Cadaver Centre at the University of Dundee.

Abstract

Diagnostic imaging is the gold standard for differential diagnosis of disease, with ultrasound being the second most requested scan after X-ray with more than 8 million ultrasounds performed by NHS England in 2021 accounting for over 20% of all imaging performed. Ultrasound cross sectional imagery is used every day to make critical decisions that could drastically affect patient outcome. While diagnostic ultrasound cross sections are clearly defined within a clinical protocol, the clinician is solely responsible for acquisition and interpretation of ultrasound imagery, with few safeguards against human error. This Canon sponsored EngD looked at the potential of machine learning to standardise processes, reduce burden on users by automating the adherence to protocols, reduce the time required by streamlining workflows, and lower the skill requirement of the clinical user. The initial study was the first to characterise the response of neural networks for the classification of cross sections specified by the Japanese abdominal scanning protocol. This protocol, one of the largest ultrasound protocols ever studied, consists of 16 overlapping cross sectional views of the abdomen, and achieved a classification accuracy of 79.9%. This provided a baseline for a transfer learning study, utilising pre-trained neural networks to increase training efficiency and lead to an increase in accuracy to 83.9%. Small mobile networks were shown to be just as effective at classification of ultrasound at a fraction of the system resources, achieving

comparable accuracies of 84.5%. Novel methods of cost reduction were explored to lower the burden of production of datasets for machine learning using power theory and active learning, providing a novel cost-effective framework for data collection and labelling. In order to overcome the limitations of image-based classification, a novel approach of augmenting neural network classification with positional data from lab based positional tracking systems was proposed. Ultrasound and positional data were collected from an abdominal phantom which allowed for the classification of six overlapping and hard to recognise abdominal cross sections with accuracies above 98%. A novel pilot study on 11 soft body Thiel cadavers, further refined this technique by exploring normalisation as a method to reduce the variability of coordinates produced when scanning the abdominal cavity and achieved an accuracy 96.8% using 3 points of normalisation. This work has demonstrated the efficacy of classification of abdominal ultrasound cross sections using neural networks and overcome the accuracy limitations of image-only classification of common ultrasound edge cases using a novel positional tracking approach, that achieved results far exceeding the current industry classification standards of abdominal cross sections.

Contents

List of Figures

List of Tables

Abbreviations

3D	Three dimensional
AAA	Abdominal aortic aneurism
AFP	Alpha Fetoprotein
ASIC	Application specific integrated circuit
CT	Computer Tomography
CPU	Central processing unit
CNN	Convolutional neural network
DL	Deep Learning
EKG	Electrocardiogram
FIR	Finite impulse response filtering
GFR	Glomerular filtration rate
GPU	Graphics processing unit
IR	Infrared
IVC	Inferior vena cava
ML	Machine learning
MRI	Nuclear magnetic resonance imaging
RBM	Restricted Boltzmann machine
ROI	Region of interest
SVM	Support vector machine
UL	Ultrasound
Xray	Radiography

Chapter 1

Introduction

1.1. Project introduction

Medical imaging scans such as those produced by Ultrasound, Xray, CT and MRI, are an increasingly essential part of diagnostic process, clinicians rely on imaging modalities such as ultrasound to provide evidential proof for confirmation of differential diagnosis. Medical ultrasound is one of the most used diagnostic imaging modalities in the world, more than 8M ultrasound scans were performed in the NHS in England in 2021 accounting for 20% of all medical imaging activity [1]. Ultrasound offers a safe, portable method for producing real time scans of the human body but relies heavily on the skills and experience of the operator limiting further uptake and advancement of the modality. Currently, clinical diagnostics depends heavily on the experience and vigilance of the increasingly busy clinician to produce images of sufficient quality by manually pressing the ultrasound probe against the patient. There is currently no system to assist operators in adhering to these guidelines, nor assist in the triaging where operator fatigue may play a role in detection of important clinical details vital to correct diagnosis. The main goal of this EngD project was to develop AI solutions to assist in reducing the burden of these vital diagnostic procedures, with the ultimate goal of

automating many common ultrasound tasks, such as capturing cross sectional planes and organ scans for diagnostic analysis. This would reduce the skill level required by sonographers to product high quality imagery and even open up potential for laymen lead scans performed by patients at home.

1.1.1. Background

This Future Ultrasonic Engineering EngD project was a collaboration between the University of Strathclyde department of Electronic and Electrical Engineering, University of Glasgow, department of Engineering and Canon Medical Research with the goal of applying machine learning to the collection of abdominal ultrasound cross sections.

Canon Medical Systems Corporation are a leading worldwide manufacturer of medical imaging devices, such as CT, MRI, and advanced ultrasound scanners, and have been actively involved in the development of medical ultrasound technology since the 1970s. The company is committed to innovation and invests heavily in research and development to advance medical imaging technology both through in house development and through its partnerships with leading academic and research institutions to develop new imaging techniques and applications. Canon Medical Systems Edinburgh, since its founding as Voxar in 1995 subsequent acquisition first by Toshiba Medical and then by Canon Corporation in 2016, has a strong focus on the development of ground-breaking medical AI systems.

1.1.2. Motivation

This Canon Medical Research proposed project sought to apply machine learning to the process of abdominal ultrasound diagnostic scanning. The main goals of this project were to ensure uniformity of the capture, reduce the time taken to perform the scans, assist with the triaging, and ultimately automate the whole screening process.

This project has substantial strategic value to both Canon Medical Systems and Strathclyde University. The development of automated and guided ultrasound scanning comes at a time where there is not only an increased dependency on imaging modalities as the primary focus of confirming a differential diagnosis, but also a worldwide shortage of Sonographers capable of performing advanced ultrasound scans. The collection of standardised cross sections is of particular interest in Japan due to the annual standard screening programs mandated by corporate insurance. Once productised this technology will assist operators with lower levels of training and experience in the collection of high-quality ultrasound scans. Reducing the skill floor for this workflow, allowing it to be undertaken more cost effectively and also decreasing the burden on senior clinicians. This will also ensure greater adherence to protocol, vital for ensuring adherence to insurance. Automating the collection processes of medical ultrasound using machine learning and robotics would allow for increased uptake throughout diagnostic medicine as quality and accuracy would no longer be operator dependant reducing the burden on clinicians potentially increasing the number of diseases where ultrasound could serve as a first line diagnostic modality.

1.2. Research Questions

Primarily this project was designed to explore the potential to automate various aspect of the clinical and acquisition workflow in diagnostic abdominal ultrasound. Canon medical research specified the following research questions as part of the project specification:

1. How can AI be utilized to automatically annotate ultrasound images to identify and mark specific planes in accordance with clinical guidelines?
2. What methods can be developed to simplify the screening procedure in ultrasound imaging, particularly in automating the identification of correct planes during a continuous sweep?
3. How can the accuracy of classifying edge cases in ultrasound imaging, especially in the context of the Japanese abdominal protocol focusing on kidneys, be improved through AI?

1.3. Knowledge Contribution

This thesis made the following contributions:

Section 2.2: Provided a literature review of the foundational principles of ultrasound as well as the reasonings behind its use as a medical diagnostic modality. This was contextualised with a discussion of the positives and negatives of alternatives diagnostic medical imaging modalities.

Section 2.3: An analysis of the Japanese abdominal ultrasound screening protocol, providing the clinical reasoning behind the collection of the specified cross sections and common diagnosis.

Section 2.4: A literature review of neural networks covering the history of its development, the transition of deep learning methods, this review includes an overview of computer imaging and analysis used during pre-processing of the dataset and discussion of the foundational techniques and methods used in training a neural network.

Section 3.4: For the first time a baseline accuracy of neural network classification of the Japanese abdominal screening protocol was produced using commonly available neural networks. This was then compared to that of networks pre-trained using transfer learning from the ImageNet challenge dataset. There is a number of consistent classification error due to similarity of the ultrasound images and intersecting regions of interest.

Industry Prototype: The image-only neural networks described in Section 3.4 were integrated into an industrial prototype by engineers in Canon Japan with the assistance of the author and are currently undergoing testing.

Section 3.5.1: The accuracy results of neural networks designed for mobile applications. Producing accuracy results comparable to that of larger networks, suggesting that network size and depth provided limited training improvement due to the restricted visual information available in ultrasound scans.

Section 4.3: Development of a cost focused framework, using statistical methods commonly used in healthcare for collection and labelling of ultrasound data for the production of lower cost datasets for pilot studies. Active learning was applied to significantly reduce the cost of manual annotation of data.

Section 5.4: A proof of concept infrared positional sensor system to augment machine learning classification was tested on an abdominal phantom to improve the image-based classification of difficult to identify and edge case abdominal ultrasound cross sections. Optical Infrared proved to be a highly accurate method of probe tracking but required an expensive camera setup. The IR sensor prototype which was based off a VR body tracking setup provided increased classification accuracy using a much lower cost setup. This prototype sensor system was then subsequently used in a real-life scenario for the collection of cadaver cross sections.

Section 6.4: The efficacy of Thiel cadaver for ultrasound machine learning datasets was examined. Common physiological difficulties that may cause the region of interest to look different were explored.

Section 6.5: The positional tracking system from section 5.4 was testing in a pilot cadaver study. This study examined the real-world use of sensor information to improve edge case classification of abdominal ultrasound cross sections. Even where there was significant deviation from image norms the positional information successfully improved classification of those cross sections even when an image only approach failed due to the complexity of the imagery.

1.4. Publications

Conference Proceedings:

A. Lawley, R. Hampson, K. Worrall, and G. Dobie, “Prescriptive method for optimising cost of data collection and annotation in machine learning of clinical ultrasound,” in *45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2023)*, 2023. Based on chapter four of thesis

R. Hampson, **A. Lawley**, and G. Dobie, "Phantom study of arterial localisation using tactile sensor array and a normal vs. shear pulse pressure propagation method," in *45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2023)*, 2023.

T. Vedran, **A. Lawley**, S. McKnight, E. Mohseni, G. Dobie, T. O'Hare, C. MacLeod, and G. Pierce. "Automated Bounding Box Annotation for NDT Ultrasound Defect Detection." in *IOP Physics Enhancing Machine Learning in Applied Solid Mechanics*, 2022.

Journal Publications:

A. Lawley, R. Hampson, K. Worrall, and G. Dobie, "Analysis of neural networks for routine classification of sixteen ultrasound upper abdominal cross sections," in Abdominal Radiology, pp. 1-11, 2024. Based on chapter three of this thesis.

A. Lawley, R. Hampson, K. Worrall, and G. Dobie, "A cost focused framework for optimizing collection and annotation of ultrasound datasets," in Biomedical Signal Processing and Control, vol. 92, p. 106048, 2024. Based on chapter four of this thesis.

A. Lawley, R. Hampson, K. Worrall, and G. Dobie, “Using Positional Tracking to Improve Ultrasound Machine Learning Classification” in *Machine Learning: Science and Technology*. Based on chapter five of this thesis.

1.5. Thesis structure

The thesis is structures as follows:

Chapter 2 provides a contextual literature review of the positives and negatives of medical ultrasound and alternative modalities. An analysis of the Japanese abdominal ultrasound screening protocol with special context to the anatomical and physiological concerns of these organs. Following this is a literature review of machine learning, its history, and the transition towards more deep learning techniques. A foundational overview of computer imaging techniques used in pre-processing of the data and machine learning training methods is provided. This is then contextualised with both current and future trends in medical machine learning.

Chapter 3 documents the creation of a baseline response to the Japanese abdominal ultrasound protocol. This chapter examines the effectiveness of 9 neural networks using transfer learning for 16 abdominal ultrasound cross sections from 64 patient sets to establish a baseline response. The highest validation accuracy was attained by both GoogLeNet and InceptionV3 is 83.9% using transfer learning and the large sample set of

26,294 images. When the first and second highest predictions (top-2) is considered an accuracy of 95.1% was achieved using InceptionV3. Alexnet attained the highest accuracy of 79.5% (top-2 of 91.5%) for the smaller sample set of 800 images. The neural networks evaluated during this study were also successfully able to identify problematic individual cross sections such as between kidneys, with right and left kidney being accurately identified 78.6% and 89.7% respectively. A further case study of mobile and small sized networks confirmed that small efficient network could be highly effective for Ultrasound classification. This chapter builds upon existing studies, demonstrating the potential accuracy of multiple neural network architectures when classifying standard abdominal cross sections. Neural network depth provides only limited improvement to classification accuracy with a difference of just 2.2% between the top results of the nine networks tested. Dataset size proved a more important factor with more complex neural networks providing higher accuracy as dataset size increases and simpler linear neural networks providing better results where the dataset is small.

Chapter 4 seeks to reduce the burden of primary data collection of medical ultrasound images, which presents a notable hurdle in the form of the high costs associated with clinical data generation and annotation. The challenge of balancing costs against dataset size is a concept well-recognised within the realm of clinical trials. Consequently, the strategies employed in this domain can be adapted to streamline the data collection and annotation procedures, thereby mitigating expenses and timelines in the context of machine learning-driven feasibility studies. This chapter introduces a biphasic framework designed to evaluate the cost of data collection via iterative predictions of

accuracy in relation to sample size. The framework also incorporates active learning techniques to guide and optimise comprehensive human annotation specifically for machine learning applications within the domain of medical ultrasound imaging. The chapter showcases the potential reduction in costs through the utilisation of publicly available breast, foetal, and lung ultrasound datasets, as well as presenting a practical case study centred around the breast ultrasound dataset. The findings underline the ability to predict the correlation between dataset size and ultimate accuracy, echoing a pattern akin to that seen in clinical trials. Substantial enhancements in accuracy are observed with the utilisation of just 40-50% of the data, contingent on the applied tolerance metric. The employment of active learning further reduces the necessity for manual annotation, resulting in a marked cost reduction of approximately 66%, while maintaining a permissible accuracy deviation of around 4% of theoretical maximums. The significance of this work lies in its ability to quantify how much additional data and annotation will be required to achieve a specific research objective. These methods are already well understood by clinical funders and so provide a valuable and effective framework for feasibility and pilot studies where machine learning will be applied within a fixed budget to maximise predictive gains, informing resourcing and further clinical study.

Chapter 5 explores the use of positional data to augment machine learning classification of six otherwise difficult to identify cross sections. For large protocols like those commonly performed on the abdomen, traditional image only machine learning classification can provide only limited functionality, for example it can be difficult to

differentiate between multiple liver cross sections or those of the left and right kidney from image alone. In this proof of concept, positional tracking information was added to the image as an additional input to a neural network to provide the additional context required to recognise these optical and sensor based infrared tracking (IR) was used to track the position of an ultrasound probe during the collection of clinical cross sections on an abdominal phantom. Convolutional neural networks were then trained using both image-only and image with positional data, the classification accuracy results were then compared. The addition of positional information significantly improved average classification results from ~90% for image-only to 95% for optical IR position tracking and 93% for Sensor-based IR in six common abdominal cross sections. The addition of low-cost positional tracking to machine learning ultrasound classification will allow for significantly increased accuracy for identifying important diagnostic cross sections, with the potential to not only provide validation of adherence to protocol but also could provide navigation prompts to assist the user in capturing cross sections in future.

Chapter 6 explores the suitability of Thiel cadavers for producing abdominal ultrasound data for training neural networks. This cadaver data is then used to test positional input for improving machine learning classification of difficult to identify abdominal cross section. The Thiel embalming method preserves cadavers effectively, maintaining tissue properties similar to that of fresh cadavers, making them valuable for medical research and teaching. While there have been studies of image quality for organ recognition prior to this study, these did not consider the unique requirements of machine learning. Cross sectional ultrasound images were taken from 11 Thiel cadavers, aorta, gallbladder, bile

duct, portal vein, as well as the left and right kidneys. The chosen cross-sections include regions with overlapping structures or visual similarities, simulating the complexity of clinical scenarios. The study demonstrates that relying solely on image-based training for machine learning models would likely encounter many challenges due to physiological variations and incidental findings common in cadavers. These challenges arise from morbidity and disease processes, making it crucial to select training data with relatively normal anatomy. The previous phantom study showed that infrared positional sensor information was shown to improve machine learning classification accuracy to ~93% for six common difficult to differentiate ultrasound abdominal cross sections, but the limitations of this study meant that there was significant overfitting due to the use of a single subject and did not allow the calibration algorithm to be fully tested. The variation in anatomy visibility coupled with the variability of abdominal cavity size of a cadaver allows for additional validation of the positional tracking and calibration system for machine learning classification. 6 common abdominal scans and 3 calibration point scans were collected from 11 cadavers using an ultrasound probe with an infrared sensor attached. Neural networks were trained using image-only and position augmented datasets using transfer learning. While an image-only approach using transfer learning from the previous phantom trained models failed due to the large variation within the cadaver image sample set. The addition of positional inferred sensor data allowed for the networks to achieve average classification accuracies of 92.8% for three-point calibration. This result suggests that positional tracking could therefore substantially improve recognition of edge case and difficult to identify diagnostic ultrasound cross sections. The use of machine learning to assist in the collection of ultrasound diagnostic

cross sections could not only improve clinical workflows by automatically collecting the best image and supporting decision making, but it also provides a route towards automating the collection process.

Overall, this research provides valuable insights into the complexities of using machine learning for abdominal ultrasound diagnosis and highlights the importance of incorporating positional data for improved accuracy.

1.6. Covid-19 Impact Statement

The COVID-19 pandemic had a significant impact on both university and industrial partner operation during the time of this EngD. The initial project specification was a software-based machine learning EngD project, supported closely by data from the industrial partner. Canon Medical Research was severely affected by COVID-19 and was unable to provide required clinical data to move forward with the original project plan. In order to mitigate this disruption, it was necessary to switch to a primary collection model, using phantom and cadaver studies. While necessitating a much smaller scale level of activities, it provided a small-scale ultrasound image dataset for traditional machine learning analysis and also provided the opportunity for the development of a sensor-based position recognition proof of concept for machine learning classification but there are notable limitations due to the sample size of the datasets that could be generated using available resources.

Chapter 2

Literature Review

2.1. Introduction

This literature review is designed to provide an overview of diagnostic medical imaging and the difficulties surrounding its collection and analysis. It will provide an overview of ultrasound as a modality and outlines the fundamental scientific principles of ultrasound imaging and provides a brief discussion on alternative medical imaging modalities. There is then an in-depth overview of the Japanese abdominal protocol, looking at the anatomical and physiological reasoning behind the collection of these medical cross sections and examining the role ultrasound plays in diagnosis. Finally, computer vision and machine learning principles are discussed with regards to medical imaging and devices. This section provides a general overview of the history of machine learning as well as an initial introduction to the foundational techniques underpinning this project. Each subsequent chapter has its own literature review focused on its specific specialist area.

2.2. Diagnostic Medical Imaging

The utilisation of medical imaging has progressively gained widespread acceptance within clinical protocols, as it is able to provide a consistent empirical means of confirming differential diagnosis. This shift towards medical imaging stems from a collective effort to establish safer and more consistent empirical diagnostic procedures, moving away from relying solely on clinical judgment, where individual clinician skills could significantly impact the diagnostic outcome [2]. While dissenting opinions exist within clinical practice [3, 4], the growing reliance on medical imaging has consistently yielded a reduction in diagnostic errors [5-7]. It is estimated that Radiographers make an error in around 3-5% [8] of all cases with 60-80% of those error being perception based [9], this is where an anomaly is not noticed by the radiographer during reporting. In an era characterised by escalating litigation, with 25% of all law suits against clinicians and hospitals due to diagnostic error [10], the assurance of confirmatory diagnosis and treatment effectiveness stands as a paramount concern.

Between 1990 and 2000, medical imaging contributed to an estimated 150% surge in the average cost of diagnoses [11-13]. This upward trend is anticipated to continue due to mounting demand throughout the healthcare system [14]. Notably, from 2014 to 2019, the NHS observed an average annual increase of demand of ~5% for ultrasound and ~7% for CT and MRI [15]. Almost 10 million ultrasound scans were performed by the NHS in 2022, with cancer screening and diagnosis being cited as a significant factor [16]. However, investment in scanning technology has not kept up with the increasing demand leading to extended wait times for diagnosis confirmation, likely resulting in

delayed and overlooked diagnoses. This, in turn, exacerbates healthcare costs within an already strained system [17]. Alongside these factors, there is also the time and effort required to perform each scan. Ultrasound procedures necessitate clinicians to directly apply pressure with the probe against the patient. As patient numbers grow and protocols become more involved, the risk of workplace injuries such as repetitive strain injuries is increased [18, 19], it is suggested that this may play a significant role in occupational burnout [20].

These factors have culminated in a chronic shortage of skilled sonographers [21, 22]. Although clinical personnel are capable of conducting these scans, time constraints limit their capacity. Unlike other modalities, such as radiology, in these cases ultrasound often lacks the safety net of double reporting [23, 24], where both an expert radiologist and diagnosing clinician review the images and approve scans, leaving the operator solely responsible for on-the-spot decision-making during the procedure. Clinicians are grappling with escalating patient loads, often involving complex cases, resulting in fatigue and an increased likelihood of overlooking crucial details in medical image analysis, consequently contributing to diagnostic errors [25-27]. This has led to a difficulty in both the training and retention of ultrasound skills within the clinical community [28-30].

Outside of the acute setting, diagnostic screening and monitoring programs are widely acknowledged as a useful tool for the early identification of disease [31, 32]. The success of screening programmes hinges on both the adherence and uptake of often asymptomatic individuals to the specific methodology employed and the nature of the

disease under examination [33]. Diagnostics in medical screening is characterised by noticing intricate nuances and sifting through a multitude of variables to converge on a potential cause using the process of making an early differential diagnosis [34]. The early diagnosis and treatment not only has consistently yielded a higher prevalence of favourable outcomes, it is also shown to have a lower financial cost associated with intervention [35, 36], however, creates a significant workload for those tasked with performing the diagnosis [37].

2.2.1. Medical Ultrasound Imaging

Medical ultrasound, also known as diagnostic medical sonography or ultrasonography, is a non-invasive imaging technique that uses high-frequency sound waves to create images of the internal structures of the human body [38]. It is widely used in the medical field to visualise organs, tissues, and various anatomical structures in real time. Ultrasound is safe, painless, and does not involve ionising radiation, making it suitable for use in a variety of medical settings.

Medical ultrasound devices are typically small enough to be mounted in a cart (as seen in Figure 2.1(a)) allowing for procedures to be performed at the patient's bedside or clinical office. A handheld device called a transducer (as seen in Figure 2.1(b)) emits sound waves into the body. These sound waves are then reflected back as echoes when they encounter different tissues and structures with varying densities [39, 40]. The transducer also receives these echoes and sends them to a computer, which processes the information to create real-time images on a monitor. While there are a wide variety of the transducers on the market today, the most common transducers used for abdominal

ultrasound are curved convex arrays commonly used for capturing standard abdominal cross sections and flat, linear probes, more commonly used for visualising musculoskeletal (MSK) or vascular structures. The resulting images provide valuable information about the size, shape, texture, and movement of the organs and tissues being examined (Figure 2.1(c)).

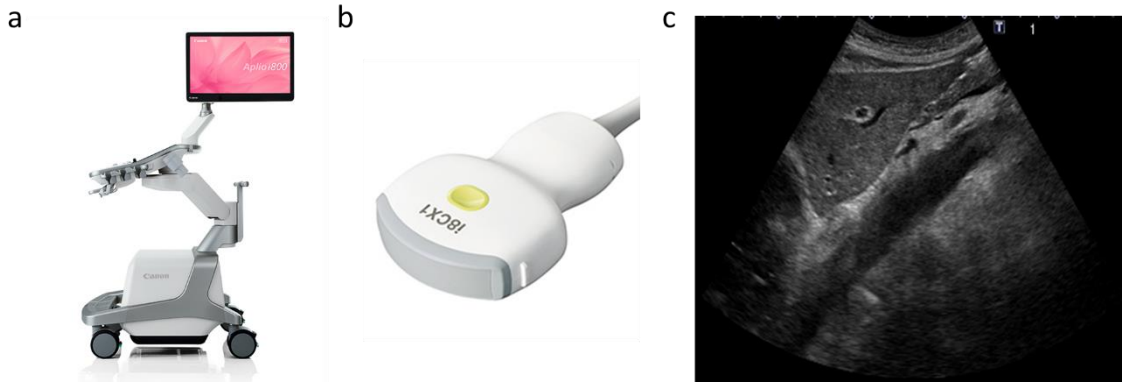


Figure 2.1 - a) example of a typical Canon medical ultrasound device [41]. b) example of a typical convex ultrasound transducer [42]. c) example of b-mode medical ultrasound image of the liver.

The market for medical ultrasound devices is expanding rapidly, it is currently estimated to be worth around US$6.5billion per annum (p/a) with that number set to grow to around US$11.5billion p/a by 2030 [43]. While there are hundreds of manufacturers of medical ultrasound devices, the five largest manufacturers are: GE, Siemens, Phillips, Hitachi and Canon Medical (previously Toshiba Medical) [43] Typically a cart mounted medical ultrasound devices cost around £30,000-100,000 with a transducer costing between £6,000 and £20,000. New technologies such as wireless ultrasound scanners costing as little as £1,700 are currently being trialled by the NHS with encouraging

results [44] but it is important to highlight that there is always a trade-off among portability, cost, and image quality.

2.2.1.1. Fundamentals of Ultrasound

Ultrasound imaging works by transmitting and detecting high-frequency sound waves and analysing how the sound frequency changes as it travels through a scanned material. This process allows us to create images of internal structures [45]. The creation and measurement of ultrasound is performed using a piezoelectric transducer, built upon the discoveries made by the Curie Brothers in 1880 [46]. This transducer leverages the ability of piezoelectric materials to convert electrical stimuli into mechanical vibrations and vice versa, enabling the precise manipulation and measurement of mechanical sound waves via electrical current. The optimal frequency for effective propagation hinges on the intended application. Diagnostic medical ultrasound typically operates within the frequency range of 2 MHz to 15 MHz [38] with lower typically frequencies used for wide area scan of anatomical areas deep within the body. Higher ultrasound frequencies enhance spatial resolution for imaging, yet they are hampered by wave attenuation, reducing their effective detection distance. This phenomenon of acoustic impedance and attenuation significantly influences ultrasound theory. As material density or sound wave speed rises, so does the impedance against ultrasound propagation. Attenuation occurs as the wave traverses an object, contingent upon its frequency and density [47]. Consequently, a trade-off between resolution, wave speed, and scan depth emerges during ultrasound scans [48].

Medical ultrasound devices are composed of transducer arrays arranged for specific beam shapes to generate images. B-mode ultrasound (as seen in Figure 2.1(c)) is one of the most common ultrasound imaging modes and employs pulse echoes to construct cross-sectional images. The amplitude of the returning echo modulates image brightness of the returning elements that make up the image. Organ complexity necessitates image construction based on transducer angles and echo amplitudes. However, image quality is restricted by array type, frequency, and scanning depth. Ultrasound data can further be distorted by artifacts and noise during image formation, such as in-filling between array elements, leading to interpolation issues [49].

Ultrasound artifacts are errors in the visual composition of the image and arises from various causes, including technique errors and disease processes. They often cause a reduction in ultrasound image quality, crating false landmarks and can obscure important visual information in noise and shadow [50]. The three primary types: reflection, shadowing, and speckle, have a distinct impact on image quality [51, 52]. Reflection artifacts (Figure 2.2(a)) stem from reverberated pulse data, resulting in artefacts such as bright lines or trailing, and mirror imaging transposing and obscuring anatomical details [53]. Reflection artifacts are known to cause difficulties in machine learning as it creates false landmarks and obscure true landmarks [54].

Shadowing (Figure 2.2(b)) is due to pulse echo attenuation caused by dense materials like bone, calcifications, tumours, gas pockets, and tissues [51]. Speckle, or acoustic noise (Figure 2.2(c)), caused by nearby transducer elements, consistently hampers image quality, affecting contrast and spatial resolution. This can cause significant difficulties in

machine learning [55] especially for segmentation and boundary tasks [56]. Ongoing research explores techniques like higher frequencies [57], statistical analyses [58-60], and filtering [61, 62] is being explored in order to mitigate speckle that reduces image quality. Noise artifacts are known to cause significant reduction in ultrasound image quality which could reduce machine learning accuracies [63].

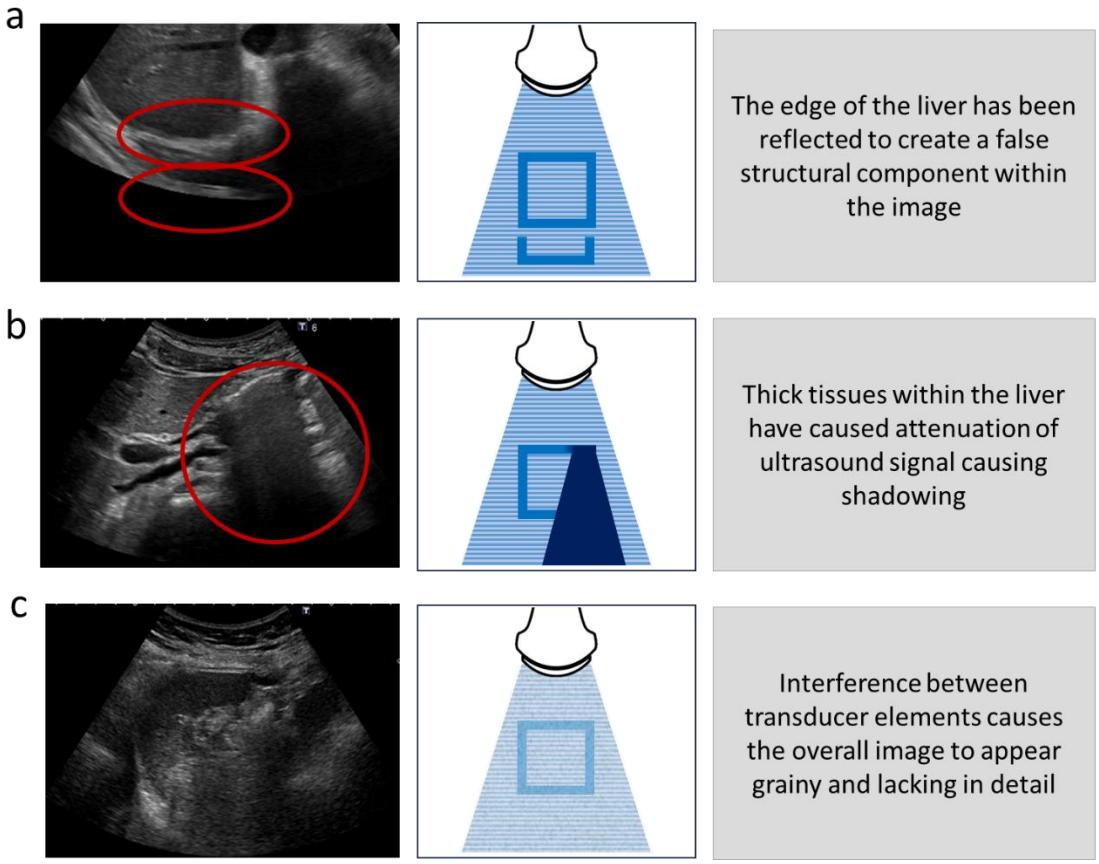


Figure 2.2- Diagrammatic representation of ultrasound artifacts that can be seen within liver scans: a) reflection, b) shadowing, c) speckle.

Ultrasound is non-invasive and boasts an excellent safety track record with no recorded injury due to diagnostic ultrasound in the last 50 years of use although there is an accepted non-zero risk of injury [64]. Soundwave transmission from the transducer into

the object necessitates direct contact or a coupling medium. With each ultrasound pulse, energy is lost as heat, this is often insignificant over short durations, but can build during high-frequency operations, or with stationary beams potentially leading to potential burns [65]. The accumulation of mechanical energy as cavitation, where high .frequency ultrasound generates destructive super-heated bubbles, also poses a limited safety concern [66], with cavitation extremely difficult to induce at diagnostic frequencies even under lab conditions [67]. Clinical and scientific regulatory bodies tightly govern ultrasound's usage. Ultrasound vibrates the tissues of the body which causes a build-up of heat energy which can cause tissue damage, this heat energy is measured as a thermal index. Vibrations can form into waves that cause the formation of bubbles, which can burst causing damage in a process known as cavitation, the measurement of these bioeffects is done via a mechanical index. The Safety Group of the British Medical Ultrasound Society formulates guidelines dictating its clinical application [68]. These guidelines mandate trained operators, strict operational protocols, and recommend a thermal index below 1.0 for non-obstetric uses and a mechanical index of below 0.7 where contrast agents are in use. Minimising mechanical stress by limiting the amount of time the probe is pressed against the skin and using the lowest required amplitude and frequency to help mitigate the risk of heat damage and cavitation.

Ultrasound's role in diagnostics is growing, particularly in acute settings where there is a necessity for rapid access to imaging data, such as in emergency and surgical interventions [69]. Its high response rate also benefits real-time application and procedures such as guided biopsies and aspirations, where real time positional awareness

is critical to the procedures success [70-72]. The use of more advanced imaging techniques such as Doppler (Figure 2.3(a)) which measures the soundwaves that have been scattered and reflected off of red blood cells to visualise the movements of blood around the body, make it ideal for detecting blood clots and vascular constrictions [73]. As well as elastography (Figure 2.3(b)) which creates images of tissue stiffness, using either a strain method whereby the tissue displacement is measured in response to pressure or through shear wave elastography where the speed of shear waves traversing tissue (often generated by a burst of high frequency ultrasound) is measured and used to create an image of tissue stiffness [74, 75]. While medical ultrasound affords real-time accessibility, safety, and diagnostic imaging potential, it remains constrained by the physical limitations inherent of sound waves such as those caused by the frequency and amplitude of the ultrasound.

Diagnostic imaging encompasses a spectrum of modalities, each offering distinct data visualisation methods rooted in their underlying physics and mechanical processes. Among the most prevalent imaging techniques are X-ray, CT, and MRI, each characterised by its own set of advantages and limitations [76], while the use of hybrid methods that combine multiple modalities have the potential to enhance imaging [75, 77], the choice of modality requires careful consideration of the imaging requirements in order to make a suitable choice for differential diagnosis and therefore provide context for the use of ultrasound.

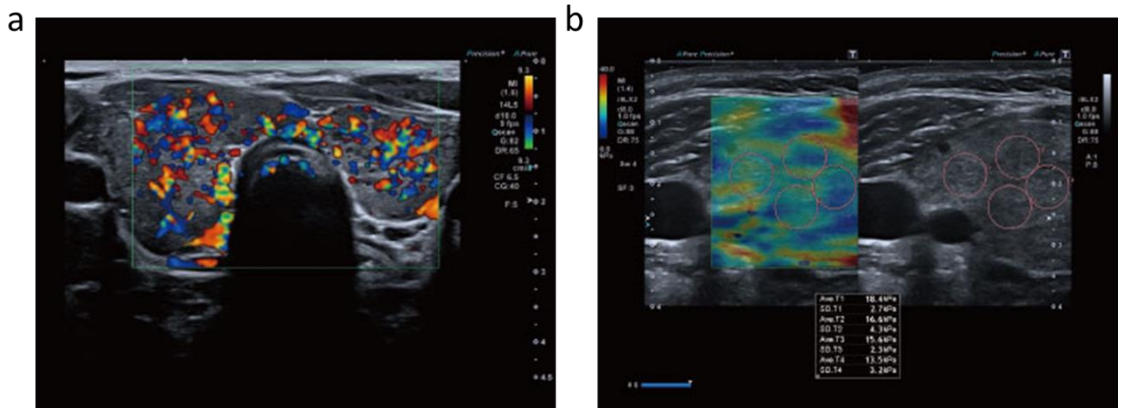


Figure 2.3 - a) example of ultrasound Doppler scan showing blood flow within the liver b) example of Elastography ultrasound scan showing different tissue stiffnesses with the tissues of the liver [78]

2.2.2. Alternative modality – Diagnostic Radiography

Radiography, commonly referred to as X-ray imaging, stands as the cornerstone of diagnostic procedures in clinical practice and is the most commonly used modality in the world [1]. X-ray scanners (an example of which can be seen in Figure 2.4) finds widespread application in clinical scenarios ranging from the diagnosis of fractures and mammography to angiography [79]. It employs high-energy ionising radiation capable of traversing through the human body. As X-rays pass through the body, they encounter varying absorption rates based on the type and density of the tissue they encounter. The resulting radiation is then captured by a specially treated plate/film or modern electronic sensors designed to detect radiographic wavelengths within the correct spectrum [80]. This differential in absorption generates detectable variations that are then translated into an image [76, 81].

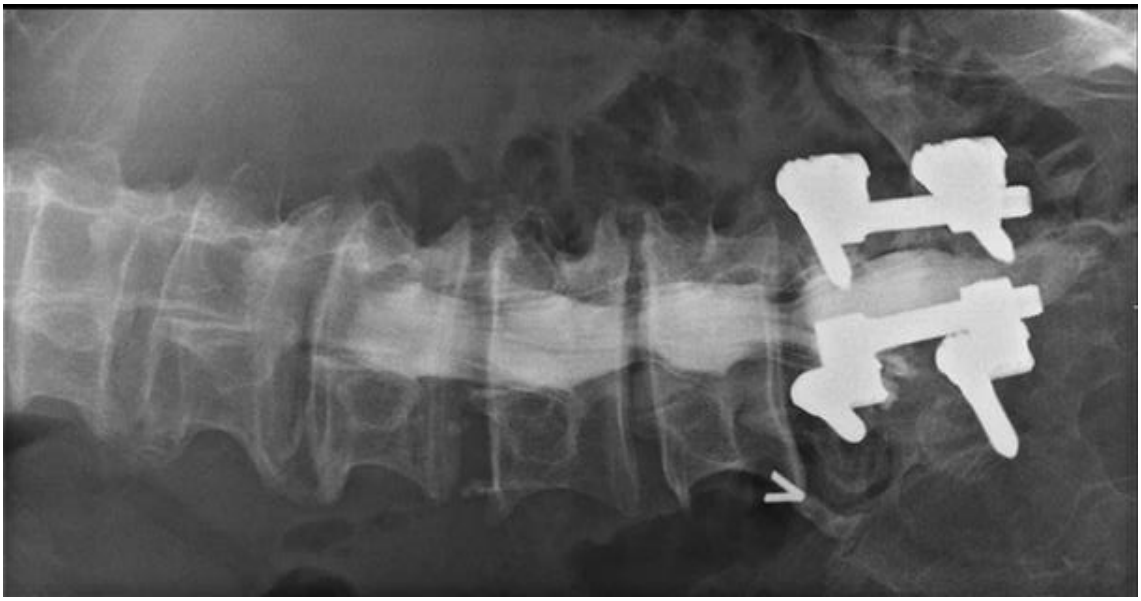


Figure 2.4 - example of an X-ray image showing spinal fusion with metal fixation [82]

Specialised techniques, such as Fluoroscopy, enable the creation of dynamic images by triggering multiple captures in rapid succession. However, it's important to note that each frame in Fluoroscopy demands a full X-ray emission, which can elevate both patient and operator exposure with each capture [83]. The introduction of contrast agents has extended the utility of X-ray imaging by enhancing visibility of anatomical structures like veins within the circulatory system that are typically not discernible on conventional X-rays [84]. Ongoing research explores how contrast agents might further expand the applications of X-ray imaging [85] .

In a review article, the British Institute of Radiology evaluated the radiation risks associated with standard diagnostic X-ray procedures [86]. Their conclusion emphasised that the benefits offered by this imaging modality outweigh the proportional risks, given that the risks are aligned with the level of exposure and individual risk factors. Standard X-ray procedures entail relatively low dose of around 0.02 millisieverts (mSv)for a

typical chest X-ray [87] which is well within the average tolerances of naturally occurring background radiation (around 2.23 millisieverts (mSv) in the UK) exposure over the course of a year. However, the increased utilisation of advanced imaging techniques has been linked to as much as a six-fold rise in ionising radiation exposure, sparking concerns about potential long-term effects of such cumulative exposure [88, 89].

2.2.3. Alternative modality – Computer tomography (CT)

Computed Tomography (CT) as seen in Figure 2.5 is a cutting-edge medical imaging technique that employs computer-controlled X-ray emitters to construct a composite image through numerous scans. This process involves the rapid rotation of the radiation emitter and detector on a gantry along a circular track. During each cycle, which takes approximately 0.5 seconds, a sequence of 2D slices is generated. These slices are then compiled to form a 3D cross-sectional view, allowing for the creation of intricate and multifaceted images that far surpass the capabilities of traditional X-ray imaging [76].

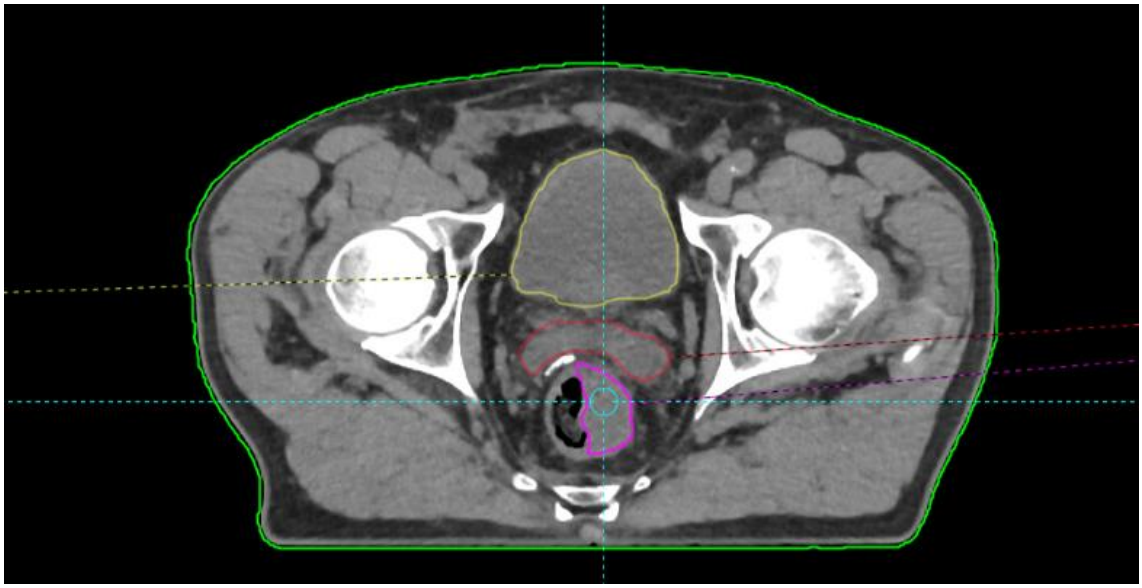


Figure 2.5 - example of a cross sectional CT scan of the abdomen with coloured segmentation [90]

In contrast to conventional X-rays, the quality of a CT scan hinges on factors like the pitch and speed of the scan [91]. High-speed, high-pitch scans might exhibit increased noise and reduced spatial resolution. Any slight movement, even breathing, can cause misalignment and blurring in the resulting images, often necessitating a rescan. The radiologist will often instruct the patient momentarily hold their breath and be as still as possible to minimise movement artifacts [92]. CT's sophisticated approach enables the generation of highly detailed imagery, to the extent that scan data could potentially be harnessed to construct intricate 3D models of specific anatomical structures of interest [93]. The capabilities of CT are further amplified through supplementary techniques such as Positron Emission Tomography (PET). This technique combines a radionucleotide tracer material with the CT scanner to capture functional and metabolic scans. These scans are instrumental in detecting structures like cysts and tumours, adding an extra dimension of diagnostic insight [94].

CT's inception revolutionised medical imaging, leading it to be considered the gold standard for diagnostic imagery; however, the associated risks of radiation exposure must be acknowledged. The radiation exposure from a single CT scan is much higher than that of a basic X-ray, according to Public Health England [87] the average population of the UK is exposed to 2.23mSv of background radiation a year, with the average chest Xray dose being just 0.02mSv, in comparison the dose for the average CT abdominal scan can be as high as 10mSv, with subsequent scans further amplify the potential for adverse events, particularly an increased risk of cancer [95]. There has been a concerted effort to explore alternative modalities, minimise scan size to the smallest area of interest, and restrict exposure to the lowest levels necessary for effective imaging. Even in relatively modest numbers, the incidence of cancer is discernibly increased due to cumulative radiation exposure [96, 97] although recent studies using modern constraints suggest repeat low dose CT poses much less risk than previously thought [98]. There has been a marked increase in awareness within the medical community regarding the risk factors associated with CT scans. While CT remains a safe imaging technology, it has been noted in clinical practise manuals and protocols that CT should be used sparingly in cases where its ability to provide highly accurate imaging can be best utilised and avoided for routine scans or where other modalities may be sufficient [99, 100].

2.2.4. Alternative modality – Nuclear Magnetic Resonance Imaging (MRI)

Nuclear Magnetic Resonance Imaging (MRI) as exampled in Figure 2.6, employs non-ionising magnetic radiation to create high-contrast images by exciting water molecules within the human body. The technique capitalises on the behaviour of hydrogen nuclei within water cells. When placed within a strong magnetic field, these hydrogen nuclei align to a lower energy state. A radiofrequency pulse then energises some of these nuclei, prompting them to transition to a higher energy state. As the excited nuclei relax back to their lower energy state, data is acquired by the MRI's coil receiver, capturing the decay. This data is subsequently converted into image data. By applying gradients selectively to the magnetic field, only thin slices of the body are influenced at a time. Leveraging this principle, MRI machines utilise sophisticated pulse sequences to reconstruct both 2D and 3D images [76, 101].

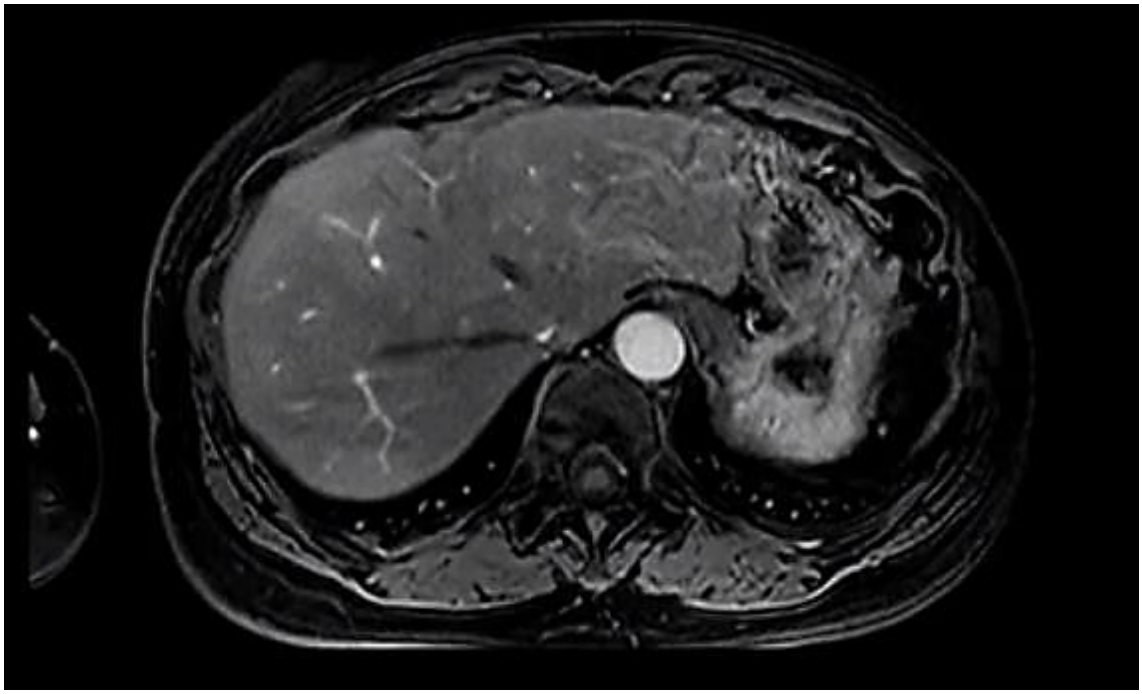


Figure 2.6 - example of a cross sectional MRI scan of the abdomen [102]

MRI delivers valuable clinical insights without resorting to ionising radiation, but the technology has a number of limitations. The quality of an MRI image is tied to the number of slices or scans acquired as well as the stability of the magnetic field. Consequently, higher-quality images necessitate more scans, extending the scan duration. This can become impractically lengthy, prompting the utilisation of techniques like sparse data processing to reduce scan times [103]. Research is currently ongoing into the development of high magnetic field MRI devices that have shown to substantially improve image quality but are not in wide clinical use [104, 105]. The field stability can have a marked effect on the signal and image quality. MRI scans are performed within an artificial magnetic field which is prone to drift due to a number of factors such as temperature and vibration with as much as a 5% deviation in signal strength reported

within a 15 minute scan, not only causing noise and interference but also potentially make scans less clear [106].

While offering an alternative to ionising radiation for medical imagine, one of the major drawbacks of MRI is that the system emits a strong electromagnetic radiation, necessitating comprehensive shielding. The strong magnetic field requires cautious handling near ferromagnetic materials to prevent accident or damage. This contraindication is becoming increasingly problematic due to an increase in the usage of medical implantable devices within an aging population [107]. To ensure optimal image quality, the electromagnets within MRI machines must be maintained at extremely low temperatures. This refrigeration process involves substances like liquid helium and nitrogen, which can be expensive to obtain and are susceptible to shortages, potentially posing challenges for future scaling of MRI technology [108-111].

2.2.5. Modality Comparison

Ultrasound offers a low-cost medical imaging solution, uses high-frequency sound waves to create real-time images, which can be performed in any clinical setting without the need for ionising radiation or costly coolants making it ideal for repeated examinations such as abdominal health screening and during pregnancy. Ultrasound excels in visualising soft tissues, organs, and blood flow but faces a number of challenges in its use:

- It is highly dependent on operator skill for both collection of data and interpretation.
- It offers only a limited field of view compared to other imaging modalities.
- There are fundamental limitations in its ability to penetrate deep into the body.
- Image artifacts can cause interference and reduce image quality.
- There is substantial difficulty scanning areas surrounded by bone.

Diagnostic radiology is the most common diagnostic modality in the world today. It is invaluable for visualising dense structures like bones and detecting fractures or abnormalities such as pneumonia and cancerous tumours. However, X-rays are limited in their ability to visualise soft tissues, like muscles and organs, without contrast material. C-arm fluoroscopy can provide real-time imaging, but both patient and clinician are exposed to repeated doses of ionising radiation during the procedure leading to potential increase in cancer risk. Ultrasound has increasingly been seen as a replacement for fluoroscopy-based procedures as it is much safer for the operator, who potentially performs multiple procedures every day.

CT is considered one of the highest quality medical imaging modalities currently available today, it is the gold standard in diagnosis, producing detailed high resolution cross-sectional images throughout the body. CT is particularly valuable for detecting and characterising abnormalities, but it is generally more expensive, less portable, and requires more patient preparation compared to ultrasound. There are also concerns at the amount of ionising radiation required to produce high quality imagery, leading to other

modalities such as Ultrasound being prioritised for first line diagnosis, with CT being used for follow-up work.

While MRI offers superb image quality compared to ultrasound and is particularly adept at visualising soft tissues, the brain, spinal cord, and musculoskeletal structures throughout the entire body, but these MRI facilities are expensive to set up and maintain, with coolant shortages likely to curtail future growth as demand outstrips supply. The use of a strong magnetic field is contraindicated in patients with implants, as even where these implants are not ferromagnetic, the agitation of molecules can cause heating within the implant and potential burns. All of these factors will likely make ultrasound a more attractive modality over MRI in all but the most specialist use cases.

When comparing diagnostic medical imaging modalities (as in Table 2.1), it is important to understand the landscape with which these devices operate. While individual modalities may be considered gold standard for diagnosis, often, their individual strengths are used to complement one another within a diagnostic protocol to provide a comprehensive evaluation. An initial first differential diagnosis could be performed using ultrasound by a general practitioner with a referral for a CT made based on those initial findings. The choice between these imaging modalities depends on the clinical context, the information needed, and considerations such as radiation exposure and cost.

Table 2.1 - Comparison of typical strengths and weaknesses of medical imaging modalities for abdominal scans

	Ultrasound	**X-ray**	**MRI**	**CT**
Cost	£30,000+	£40,000+	£500,000+	£1 Million+
Time Till Complete Scan	Real time	5-10 second delay	15-90 minutes dependant on resolution	2-5 minutes ~20 seconds for low resolution
Operator	Limited training	Trained professional	Specialist Radiologist	Specialist Radiologist
Diagnostic Image Quality	Low	Low	Moderate dependant on resolution	High dependant on resolution
Contraindications	Safe within standard operational parameters	Discouraged in pregnancy	Do not use in proximity of ferromagnetic materials	Limit exposure due to high radiation
Radiation Safety	Ultrasound radiation	Low level ionising radiation	Magnetic radiation	Potentially high ionising radiation exposure
Running Costs	Low very few additional costs on purchase	Low running costs after initial purchase	Requires expensive coolant	Expensive to setup and run

2.3. The Japanese Abdominal Ultrasound Cross Sections

The Japanese abdominal ultrasound protocol as defined by the Japanese Society of Sonographers consists of 16 cross sectional views of the abdomen [112]. The human abdominal cavity is located inferior to the thoracic cavity beneath the diaphragm. This cavity contains many of the human bodies vital organs such as those of the hepatic, renal, lymphatic, endocrine, digestive, and biliary systems, as well as many major circulatory structures such as those of the aorta and inferior vena cava that circulate blood throughout the body. This screening protocol is used extensively within the Japanese health insurance industry as part of a yearly physical. Recipients of this scanning protocol are largely working age Japanese males with health insurance provided by their employer. Although it is also extensively used for detection and monitoring of abdominal diseases and in cancer detection protocols.

Anatomically, the posterior demarcation of the abdominal cavity is established by the vertebral column, extending from the fifth thoracic vertebra (T5) to the first sacral vertebra (S1). The anterolateral margin is defined by ribs five through ten and the oblique and transverse abdominal musculature, which collectively constitute the external abdominal wall. The organs themselves are contained within the Peritoneum, a serous membrane lining the cavity creating a highly contained homeostatic environment [113]. The concentration of anatomical structures and organs within the abdominal cavity necessitates a technique to examine these structures in-situ without causing disruption to homeostasis and risks of infection that would occur during exploratory surgery, as such

the abdomen has already been a site for major research into diagnostic imaging techniques for many years. The 16 cross sections are as follows:

1. Epigastric sagittal scan: Liver/aorta
2. Epigastric horizontal scan to right subcostal scan: Hepatic vein
3. Right Epigastric oblique scan: Horizontal portal vein
4. Right Subcostal scan: Gallbladder
5. Right hypochondrium vertical scan: Gallbladder
6. Right hypochondrium vertical to oblique scan: Extrahepatic bile duct
7. Right subcostal scan: Liver
8. Right intercostal upper scan: Liver
9. Right intercostal mid scan: Liver
10. Right intercostal lower scan: Liver
11. Right intercostal scan: Right kidney
12. Epigastric vertical scan: Extrahepatic bile duct/pancreas
13. Epigastric horizontal scan: Pancreas
14. Epigastric oblique scan: Pancreas
15. Left intercostal scan: Spleen.
16. Left intercostal scan: Left kidney.

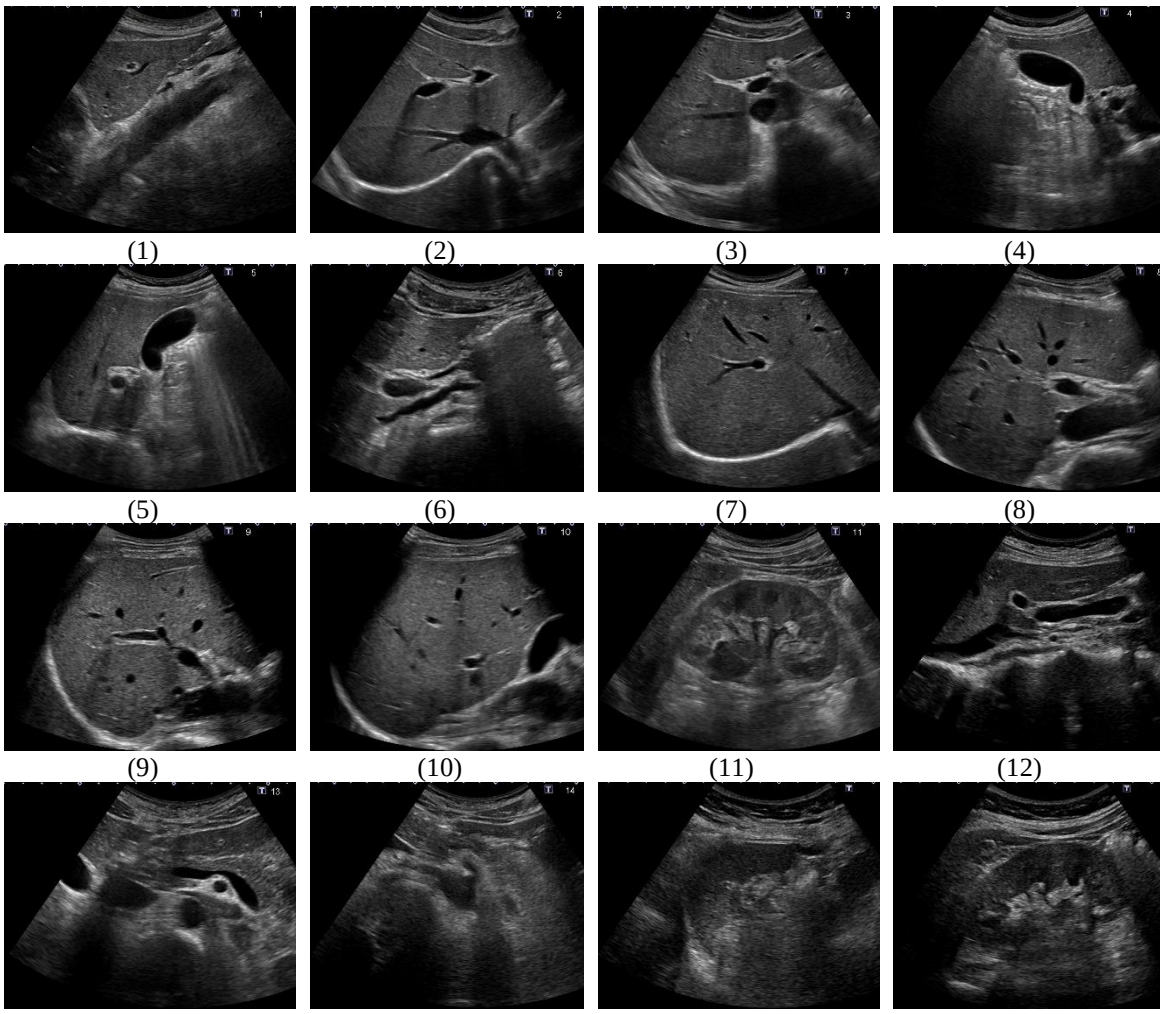


Figure 2.7 – example of the 16 upper abdominal cross sections outlined within the Japanese abdominal ultrasound screening protocol.

In order to provide context to why these scans are so important to the abdominal diagnostic protocol, the anatomical structures and organs featured within the 16 cross sections will be discussed as well as any prominent physiological anomalies relevant to ultrasound scanning. Abdominopelvic quadrants are used to partition the abdominal cavity and provide accurate positional landmarking for organ and structures of interest to this protocol. As the probe position overlaps for many cross sections, please refer to the protocol for an accurate representation of probe position [112].

2.3.1. Aorta and Inferior Vena Cava

The Aorta (as seen in Figure 2.8) and Inferior Vena Cava (IVC) stand as the most significant conduits within the arterial and venous networks, both are fundamental in the systemic circulation of blood throughout the human body. Any compromise, injury, or occlusion affecting these conduits can precipitate catastrophic injury and mortality. Situated medially along the entire abdominal length and anterior to the spine, they establish vital connections to major organs via a network of arterial and venous branches. Consequently, they often reside posterior to a range of vital anatomical structures and therefore can be difficult to visualise. While the aorta is the focus of cross section one, the circulatory system can be seen within multiple cross sections within the dataset. The role of the circulatory system is so vital to the human body that even the shortest disruption can potentially cause irreversible damage from cell ischemia, as such non-invasive diagnostic methods and anomaly detection that maintain the bodies homeostasis remains a major priority for medical imaging research.

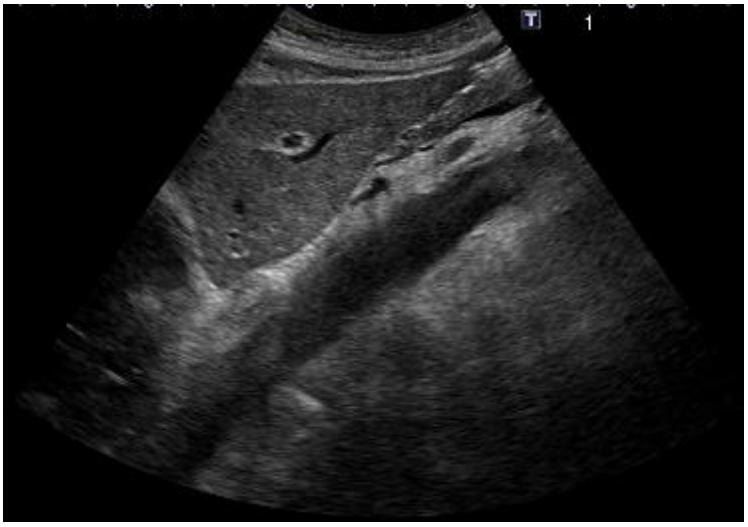


Figure 2.8 – example of an ultrasound scan of aorta.

The Aorta is the largest artery in the body and as such is exposed to some of the highest haemodynamic forces in the body, with each heartbeat, it experiences the highest changes in blood pressure outside of the heart itself. These pressure changes render it particularly susceptible to vulnerability and disease. In instances where weakening occurs due to congenital defects or pathological conditions, there is a greater potential risk of aortic rupture leading to an aneurysm. Treatment of an abdominal aortic aneurism (AAA) is best achieved prior to any rupture taking place. The lack of clear access promotes the need for imaging based diagnosis, for example ultrasound often looks for abnormalities to the thickness, displacement and flow rate in order to detect and diagnose [114]. Methods such as Doppler and contrast ultrasound allow for the visualisation and estimation of flow within the circulatory system, highlighting problem areas such as sites of arteriosclerosis.

Ultrasound has demonstrated effectiveness in identifying and monitoring risk factors associated with aortic aneurysms, such as changes in wall thickness and deviations in flow [115]. It has been shown to be a reliable method of detecting these anomalies both during screening [116, 117] and in an emergency diagnosis [118, 119]. Ultrasound operates in real time allowing for the capture of a segment of time alongside three dimensions of array data, it is therefore possible to perform a complex assessment of the strain occurring as the aorta undergoes a cardiac cycle, this information can be used to assess wall health and identify invisible weak points [120]. Blood cells are typically invisible to ultrasound but Doppler ultrasound analyses changes in the frequency of the ultrasound echo to determine the relative motion of blood using the Doppler effect,

which can reliably measure blood flow and potentially indicate a site of calcification or other form of blockage that may eventually cause an ischemic event [121].

2.3.2. Liver and Hepatic System

Situated within the upper right quadrant of the abdomen, the liver is a major organ in the hepatic system, specialising in the metabolisation of key enzymes and proteins, detoxification, and serves as a repository for glucagon, a key component of maintaining the bodies levels glucose levels. As seen in Figure 2.7 and Figure 2.9 the liver features prominently throughout the cross sections of the dataset either as the main featured organ or within the background as a landmark feature. Traditionally, the liver is segmented using the Couinaud method [122] which anatomically divides the liver into eight segments based on the vascular and biliary anatomy. Each of these segments is supplied independently by branches of the hepatic artery, portal vein, and bile ducts. This segmentation technique while still the gold standard in surgical planning and understanding the liver's functional anatomy is currently under review as several studies suggest that this technique of using indirect landmarks may not be accurate or reliable enough for modern imaging modalities [123].

Figure 2.9 – example of an ultrasound scan of the liver.

Fast paced contemporary lifestyles are seen as a contributing factor to the increase in hepatic diseases such as fatty liver disease with approximately 15-30% of adults in first world nations expected to develop a form of hepatic lipid accumulation throughout their lifetime [124]. While adopting healthier dietary habits and incorporating exercise routines can prevent long-term damage in many instances, a failure to address these concerns can culminate in chronic conditions. Ultrasound has been shown to be an effective methodology for detecting lipid build-up with both the detection of steatosis and the increase of attenuation of the liver as liver health declines [125, 126]. Ultrasound has also been shown to be effective for early detection of liver cancer in a number of studies [127], both in conjunction with liver function blood tests (such as AFP) [128], and under contrast [129]. Elastography is increasingly used to detect liver cirrhosis, with many commercial devices, such as Fibroscan already in clinical use [130]. The global incidence of chronic liver disease is on the rise, propelled by risk factors such as increased alcohol consumption, diabetes, fatty diets, obesity, and genetics playing a

role in the increase in cases of liver failure which can often lead to requiring transplantation [131]. The incidence of hepatic carcinomas (liver cancer) has also seen a significant increase [132] throughout the world.

Research is ongoing to develop quantifiable medical imaging techniques, such as grading the liver to assess liver health during long term monitoring. The attenuation properties of ultrasound have shown promise as a method of detecting liver diseases but has so far struggled with early detection [125], with research looking to use contrast-enhanced imagery [133, 134] and elastography [135] as potential solutions.

2.3.3. Kidneys and Renal System

Positioned within the lateral confines of the upper quadrant, the kidneys fulfil a pivotal role in filtration, the regulation of diverse electrolyte concentrations, acid-base equilibrium, osmolarity, and the expulsion of toxins. These processes culminate in the formation of urine, which is subsequently conveyed through the urinary tract to the bladder for eventual excretion. Renal function and the expulsion of toxins is a key attribute for overall body health, nephritic anomalies, exemplified by conditions such as renal carcinoma, kidney injury, renal calculi, and urinary tract obstructions, all have the potential to inflict prolonged physiological damage leading to increased morbidity. As seen in Figure 2.10, the kidneys are major features in cross sections 11 and 16 but feature in a number of sweeps as incidental landmarks particularly in that of the spleen and some right sided subcostal liver scans.

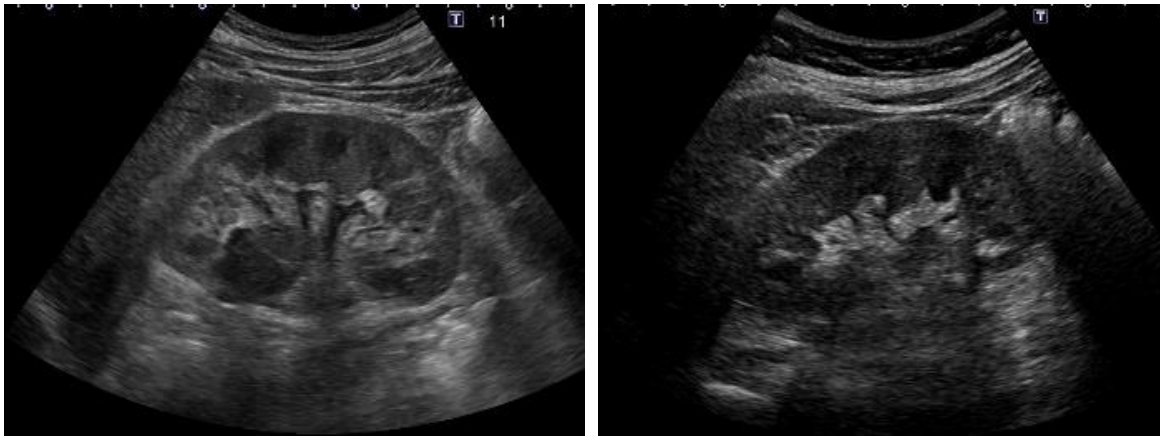


Figure 2.10 – example of ultrasound scans of the left kidney (left) and right kidney (right)

Structurally, the kidney comprises of a large renal cortex housing multiple lobes, each containing nephrons responsible for blood filtration and urine generation. While there are accurate haematological tests to measure renal function in the form of glomerular filtration rate (GFR) and measuring Serum Creatinine levels, systematic dysfunction such as cirrhosis can potentially cause sudden loss without significant haematological derangement [136, 137]. This is not ideal as these tests rely on extensive clinical knowledge of the patient's history to make an accurate differential, leading to potentially missed or delayed diagnosis. As the incidence of kidney diseases continues to rise, ultrasound has shown to be an effective imaging modality for assessing a range of renal conditions. It has shown effectiveness in the detection of renal masses [138], renal calculi [139], and the evaluation of kidney health through vascularisation studies [140]. Notably, these approaches are progressively garnering attention within clinical realms, particularly as microbubble contrast agents demonstrate the potential to enhance Doppler ultrasound's efficiency. Notably, this augmentation occurs without a substantial elevation in the risk of air emboli [141]. Due to the highly vascular nature of the kidney, contrast-enhanced techniques have proven to be highly effective in pinpointing lesions

and masses within these organs [142-144], even in cases complicated with chronic kidney disease where there may be difficulties with perfusion [145].

The incidence of renal cell carcinomas has stabilised or even decreased in certain regions, early identification remains imperative for successful treatment outcomes [146]. Ongoing efforts are dedicated to refining segmentation techniques suitable for imaging modalities like ultrasound [147]. Kidney transplantation has the potential to cause difficulty when considering machine learning, failed kidneys are routinely left in place unless there are suspected complications from the organ, meaning that multiple kidneys (or a single organ in the case of the donor) may be located within the abdominal cavity which would contraindicate the use of machine learning, as the dataset would need to label to detect such anomalies.

Despite its promise, ultrasound adoption has been slow in comparison to other contrast-based modalities like contrast X-ray and CT, which have largely dominated the field. However, ultrasound holds distinct advantages, including its affordability, safety, and portability, all achieved without recourse to ionising radiation [148].

2.3.4. Biliary System and Gallbladder

The gallbladder is a small ‘pear’ shaped hollow organ located medially beneath the right lobe of the liver. The gallbladder and bile ducts (as seen in Figure 2.11) are major features in cross sections 4, 5 and 12, but also feature as landmarks in a number of other cross sections of the liver. Functionally, it serves as a reservoir for bile, a digestive fluid secreted by the liver, which is subsequently dispensed via the bile duct and biliary

system into the stomach to facilitate digestion within the stomach. Common aetiology of the biliary system is often due to difficulties in drainage, which can cause discomfort. Cholecystitis, characterised by inflammation caused by obstruction by a ‘stone’ of calcium salts is a prevalent example which has many potential complications such as jaundice and pancreatitis [149]. Gallbladder Carcinomas are fairly rare, often progressing asymptomatically until later stages, making screening the best option for diagnosis [150].

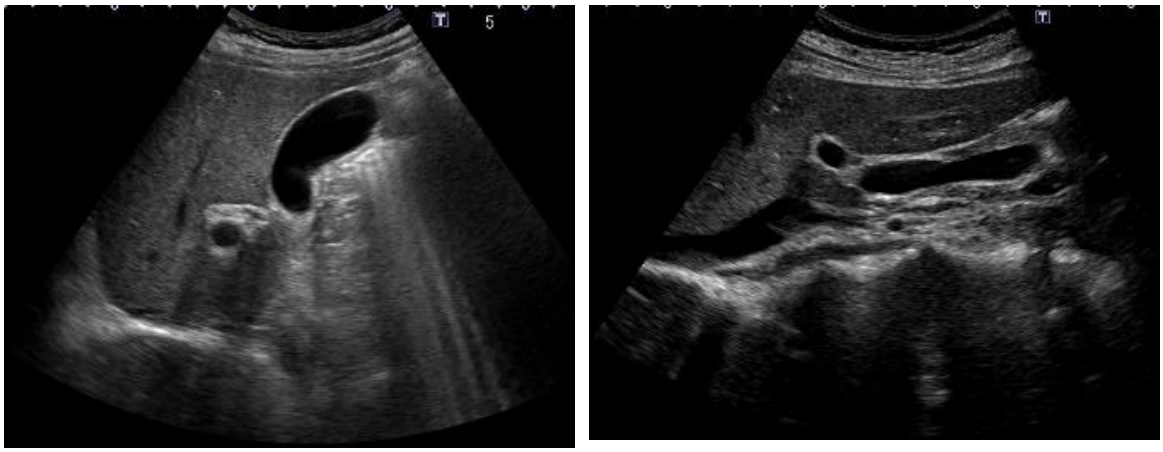


Figure 2.11 – example of ultrasound scans of gallbladder (left) and bile duct (right).

Ultrasound is the dominant modality during initial investigation of diseases of the gallbladder and highly effective at detecting gall stones, small crystallised calcium salt stones within the Gallbladder, as well as more complex conditioning such as wall thickening and inflammation [151, 152]. There is a notably high learning curve when assessing the Gallbladder with Ultrasound, with inexperienced clinicians misdiagnosing or simply unable to interpret what they were looking at in comparison to sonographers and consultant level experienced users [153].

The Gallbladder can be difficult to image as its size is dependent on the levels of bile contained within it at the time of imaging, segmentation techniques and imaging can be very limited, with ultrasound considered the gold standard for Gallbladder assessment [154]. The partial or complete removal of the gallbladder is one of the most common abdominal surgical procedures performed by the NHS with over 60,000 performed every year [155], while still a relatively limited occurrence the absence of the gallbladder as a landmark should be carefully considered when applying machine learning, as it would potentially impact multiple cross sectional views.

2.3.5. Spleen

Situated within the upper left quadrant, this organ can be categorised based on function into two distinct sections. The red pulp functions as a reservoir and filter for blood cells, effectively clearing cellular detritus such as pathogens. In contrast, the white pulp predominantly comprises lymph cells that play a pivotal role in orchestrating immune responses. This dualistic arrangement facilitates the entry of lymphocytes and macrophages into the bloodstream [156]. Cross section 15 (Figure 2.12) represents a sweep scan of the spleen, with it also serving as a landmark when localising the left kidney in cross section 16.

Figure 2.12 – example of an ultrasound scan of the spleen.

Abnormalities in Spleen physiology have an extremely wide range of aetiologies from infections such as malaria, syphilis, endocarditis and HIV, to haematologic disorders such as Cirrhosis, leukaemia and lymphoma [157]. While primary spleen carcinomas are relatively rare [158], it is often the site of secondary metastases in lung, colorectal, ovarian, skin and breast cancer [159]. While segmentation techniques can be complex with varying results, modern computer based segmentation procedures are starting to see results especially where comparisons can be made with local organs for diagnostic procedures [160]. During high trauma events the spleen may rupture causing massive internal bleeding due to its high perfusion and therefore it is safer to remove the organ entirely.

Enlargement of the Spleen can be an indication of serious infection, disease or even some cancers that may not show any other symptoms, Ultrasound has shown to be just as effective at measuring the spleen as a 3D CT scan [161], while being safer, cost effective and more accessible. Ultrasound can also be used to detect splenic injury such

as small ruptures, cysts and infarction [162]. Doppler can be used as part of a differential diagnosis for liver disease although its accuracy as a single attribute is limited [163]. Ultrasound has shown to be an effective modality for guided biopsy as it allows for real-time imaging during a spleen biopsy that is both accurate and low risk [164].

2.3.6. Pancreas

The pancreas occupies a position in the upper left quadrant, located posteriorly to the stomach. Its roles encompass the secretion of digestive enzymes into the stomach and the release of vital hormones, such as insulin, into the bloodstream to regulate metabolic processes [165]. It is featured in cross sections 13 and 14 (seen in Figure 2.13). Imaging of the pancreas presents a significant challenge due to its location and structure. Uniform uptake of contrast agents may not be achieved, and proximity to adjacent organs can introduce interference. Signs of disease can be diffuse and difficult to detect even on biopsy, CT, and MRI. This is a pressing problem in clinical diagnosis as pancreatic cancer is extremely deadly due to the operation of the pancreas and sparse nature of the organ and difficulty to deliver treatment to affected cells [166]. At present, assessment is primarily conducted using CT (computed tomography), although a burgeoning body of research suggests the potential efficacy of ultrasound for evaluation [167, 168].

Figure 2.13 – example of an ultrasound scan of the pancreas.

Ultrasound has long been considered a viable modality for scanning the pancreas [169], although interference from the stomach such as fluids and gasses have been known to cause difficulties such as shadows and artifacts that may obscure important anatomical details [170] but recent advances in ultrasound contrast have shown positive results in improving the accuracy of ultrasound imagery of the pancreas, allowing for better visualisation of cysts and tumours within the pancreas [171].

2.4. Machine Learning with Medical Application

Having established the clinical rationale that underpins the use of medical ultrasound and examined the use case behind the Japanese diagnostic screening protocol, it is important to provide a contextual overview of the machine learning methodologies and techniques that are central to this thesis.

Machine learning is a sub-field of artificial intelligence research (as seen in Figure 2.14) defined in 1959 as branch of computer science that broadly aims to enable computers to "learn" without being directly programmed [172] and has evolved over several decades to focus on computational algorithms designed to mimic the iterative learning process of the human brain. Through the process of iterative learning a machine learning algorithm develops and refines a mathematical model within a defined set of rulesets and dependencies. [173, 174]. The foundational principles of these models are based on a number of mathematical formulas dependant on the task required to be performed, these include regression (logarithmic, linear), random (naïve Bayes, random forests), decision trees, and clustering (nearest neighbour, components analysis) [175]. The algorithm fine-tunes its weights and parameters using these strict mathematical principles enabling the algorithm to refine its fitting process autonomously without requiring direct human intervention. Employing this methodology opens the door to applying the model across various tasks, for instance, it can be employed in scenarios involving conditional probability. In such cases, the algorithm applies rules founded on associations, such as can be seen in classification tasks such as identifying patterns within images or data [176].

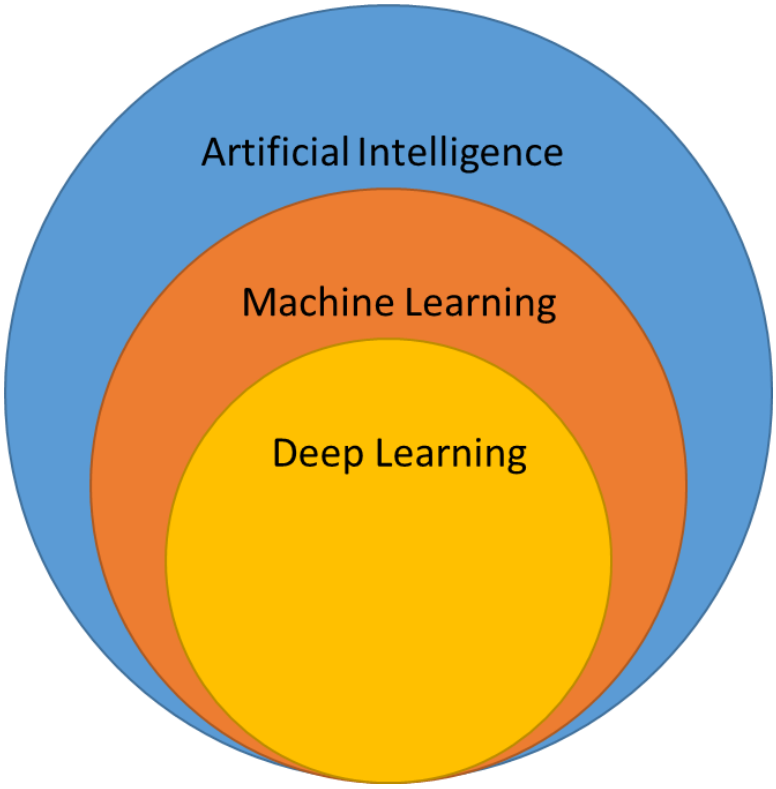


Figure 2.14 – Diagrammatic representation of machine learning subsets.

2.4.1. Early Methodologies

The early history of machine learning focused on predictive techniques such as k-nearest neighbour, Bayesian classifiers, decision trees and support vector machines (SVM) [177]. While these methods remain highly relevant today, especially for automated data mining [178], these techniques have been largely replaced by deep learning and neural networks as the tasks faced by machine learning algorithms have become more nuanced [179]. Many of these applied methods can be found in Nilsson's Foundations of trainable pattern classifying systems [180] although the statistical methods are a lot older [181]. Early examples of these rule based and pattern recognition systems proved ineffective at real world medical imaging problems [182]. It would not be until the

1980's and 90's that computing power would be good enough to further progress with Hunt et al [183] using decision trees for diagnostic medical imaging, Altman et al with K-nearest neighbour [184], Restricted Boltzmann Machines (RBM) [185] and Kononenko et al experimenting with diagnostic procedural rulesets [186]. One of the major reasoning behind the drive for deep learning algorithms is the resource requirement of earlier algorithms as seen in Figure 2.15, the more complex and nuanced the task requiring to be performed is, the more complex the algorithm must also become, conversely a deep methodology can simplify the task by spreading multiple complex analyses over multiple layers, allowing for smaller, more efficient use of computing resources.

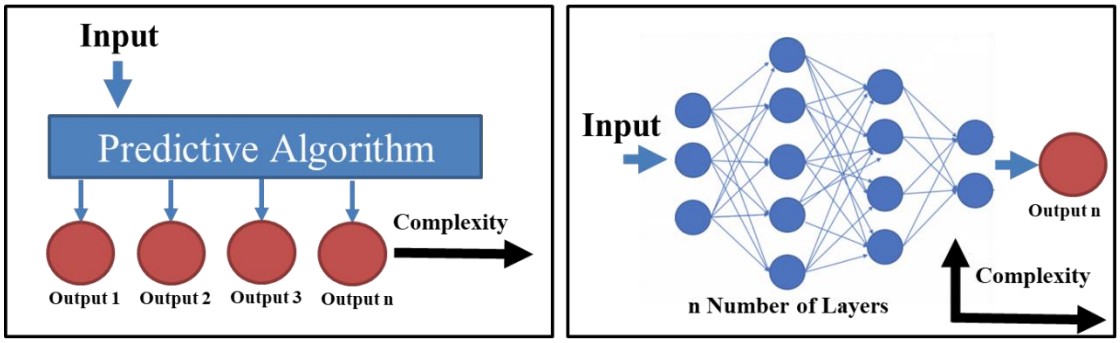


Figure 2.15 – Simple representation of flat/shallow learning vs deep learning algorithm shape.

The K-Nearest Neighbour (KNN) algorithm employs weighted parameters and distances from neighbouring data points to effectively classify data. This approach allows the algorithm to intelligently group similar data points together. This simple weighting technique can be highly successful at simple tissue based classification and segmentation

when paired with more complex analysis techniques such as histogram analysis [187] and texture-based clustering [188].

Decision trees are a widely adopted technique in which a rule-based classification system is employed to effectively categorise data [189]. The architecture of a decision tree involves a root node from which data branches into distinct nodes, each corresponding to specific attributes or categories. The accuracy level and intricacy of the data play a crucial role in shaping the size of the decision tree. Notably, as data complexity increases, the decision tree's size also expands to ensure precise outcomes. The process of node and classifier selection encompasses a range of methodologies, accommodating both single and multiple-state attributes based on the metrics under examination [190]. There are a number of methodologies for the construction of decision trees, including random forest [191] and the J48 algorithm [178] with various levels of accuracy dependant on task.

The Support Vector Machine (SVM) it is commonly used by leveraging pre-classified data or rulesets to learn patterns that can be subsequently utilised for predictions [192] but can also performs unsupervised clustering [193]. This is achieved through a process involving the mapping of data points onto vectors, followed by the application of weightings to allocate these points into distinct categories. Consequently, SVM exhibits the ability to effectively cluster data points by utilising hyperplanes to partition the dataset. One of the notable strengths of SVMs lies in its capacity to yield accurate predictions with relatively sparse data, thus mitigating the risk of overfitting, and as such has previously been used as a valuable tool for classification tasks, but it is not suitable

for large datasets or in cases where the number of features exceed the size of the dataset. SVMs have been used in research areas such as breast mass classification [194], kidney abnormalities [195], and the identification of tumours within the liver [196].

The complexity of the dataset, anatomy and image problem highlighted in Sections 2.2 and 2.3 suggests that the overlapping cross sectional data and lack of image clarity would be ultimately make these methods unsuitable for the classification task associated with this thesis.

2.4.2. Neural Networks & Deep Learning with Medical Applications

Deep learning is a subset of machine learning (as seen in Figure 2.14) that hypothesises that it is possible to intelligently solve problems by artificially mimicking the functional layout of neurons in the human brain. Initial neural network accuracy was limited by the number of neurons within the network [197]. Unlike, early machine learning algorithms complex tasks can be spread over a greater number of layers as seen in Figure 2.15, this allows for more complex analysis to be spread over multiple layers rather than be performed all at the same time. This allows for more efficient task based designs of neural networks, which has allowed for more complex non-linear datasets [198, 199]. This trend has prompted a shift towards networks featuring increased layer counts, in order to move away from the constraints posed by circuit-style functions and advance systems that mirror the multi-layered paradigm observed in the human brain. Research has indicated that incorporating a greater number of layers might facilitate enhanced

predictive capabilities in scenarios involving non-linear datasets, like those encountered in natural language processing or image recognition [200, 201].

Deep learning is an extremely powerful tool capable of being trained to recognise highly complex patterns within large datasets. Convolutional Neural Networks (CNN) (the classic layout of a CNN can be seen in Figure 2.16) have become increasingly popular for clinical research as medical datasets are primarily made up of imaging data, as convolution is already heavily used in computer vision, pattern recognition, and natural language processing. While the foundational theories surrounding CNNs can be found in Pitts and McCulloch's 1943 paper [202], it was not until 1958 that Rosenblatt outlined the principles of perceptrons [203] and neuro-computing [204]. Widrow and Hoff would publish the Widrow-Hoff learning rule in 1960 outlining what would become a common sampling method still in use today [205]. CNN research would be further progressed in 1972 with a fully-fledged theoretical framework for artificial adaptive systems published by Kloph at the Air Force Research Laboratories [206]. Two years later the initial findings regarding the back propagation optimisation technique for neural network were published by Werbos [207] in 1974 and would later be popularised by Rumelhart in 1985 [208].

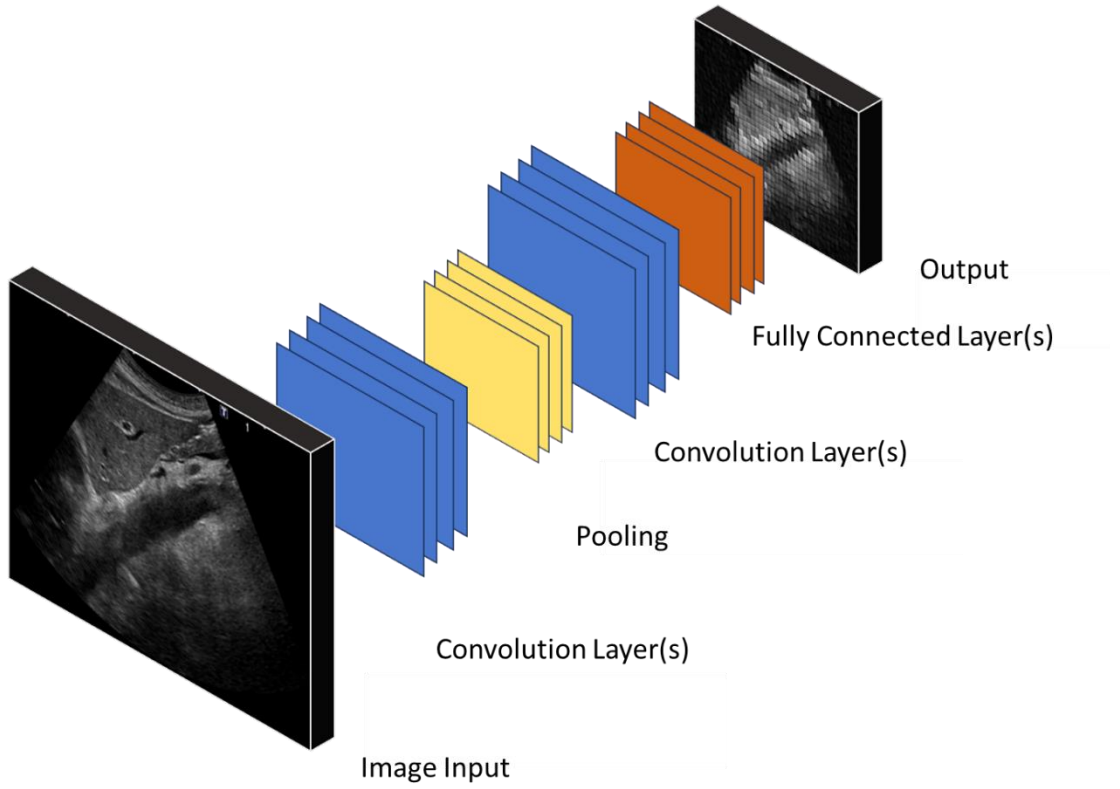


Figure 2.16 - Classic layout of a neural network. Image data passes through multiple convolution and pooling layers, and fully connected layers before the predictions are outputted.

Early examples of artificial neural networks for classification were published in the 1970s although these results were not adopted in clinical practice and were severely limited by its lack of acceptance by clinical professionals, and the limitations of the technology of the time that was not able to process complex medical data at a fast enough speed to be accepted within the medical industry [209-211]. While other areas of science saw increasing relevance to machine learning techniques, clinical medical imaging and diagnostic research lagged behind as clinicians struggled to get to grips with the complexity of the task [212]. The 1990's and early 2000's saw a resurgence of

deep learning for medical imagery in areas with a well-defined problem space such as breast mammography [213] or the diagnosis of prostate cancer [214].

Today CNNs have proven to being highly effective at processing image data, being capable of simplifying the image using a kernel to break down larger images into smaller parts that can be better processed and generalised by the neural network [215]. Ultrasound presents many challenges to the use of a neural network due to limited image detail, noise, and artifacts but is already being examined as a method of detection, segmentation, and classification in a wide range of clinical research conditions. While detecting anatomical structures can often be straightforward, ensuring precise annotation for accurate classification can present challenges. CNNs have demonstrated effectiveness in detecting and classifying masses and lesions in organs like the liver [160, 216], kidneys [217], gallbladder [154], and pancreas [218]. In addition to their classification and segmentation capabilities, CNNs offer potential value in diverse clinical diagnostic techniques. For instance, they can contribute to the evaluation of kidney function, where deep learning using ultrasound Doppler aids in predicting flow rates [219]. Similarly, CNNs are instrumental in characterising wall thickness and identifying plaque accumulation within the Circulatory system [220], as well as in examining liver fibrosis [221].

While neural network layer counts and overall size have expanded, challenges such as overfitting, where a network begins to fit too closely to the patterns in noise and other idiosyncrasies within the training data, resulting in a failure to generalise and poor performance on unseen data. Recent research highlights a striking example: a single-

pixel modification, termed the "one-pixel attack," [222] can lead a neural network to erroneously classify an image. In instances like an overfitted convolutional neural network, altering just one pixel could shift the image beyond the boundaries of its correct category. Models trained exclusively on error-free "ideal" data often lack robustness, when confronted with real-world problems and data they are unable to reconcile the imperfections within the imagery to that seen in the training, causing a reduction of the models ability to generalise [222-225]. The susceptibility of models to what can often be basic alterations or changes that can impact their categorisation raises concerns about their reliability. This vulnerability is particularly evident in domains like medical imaging data, where models have exhibited a significant susceptibility to adversarial attacks. Ongoing research aims to establish dependable strategies for mitigating these vulnerabilities [226-230].

2.4.3. Data Processing Methods

The processing and curation of data into a standardised, machine-readable format is an essential part of any machine learning method. Medical imaging modalities frequently encounter challenges related to image quality, noise and artifacts that can significantly impede the accuracy of diagnoses and may also reduce the accuracy of any applied machine learning. To address these concerns, various image processing and computer vision methods are commonly employed during pre-processing or enhancement stages to elevate the quality of imaging outcomes. The associated data processing techniques are broadly categorised into pre-processing and post-processing procedures, encompassing approaches such as contrast equalisation (for example histogram equalisation) for

contrast enhancement, noise reduction strategies, and advanced image filtering techniques.

2.4.3.1. Computer Vision Approaches for Image enhancement and standardisation

Over the past two decades, many image processing techniques derived from computer vision approaches have been adopted by researchers to enhance medical imaging for use in machine learning. Computer vision is a large diverse field of research focused on enabling computers to derive information from visual data such as images and video, consequently these techniques serve a variety of purposes, ranging from improving image quality, facilitating edge detection, feature extraction, filtering, segmentation, and classification [231]. These techniques can assist in a wide range of clinical diagnostic tasks, that are designed to examine internal body structures in a way that is minimally invasive such as X-ray, Computer Tomography (CT), Magnetic Resonance Imaging (MRI), Endoscopy, and Ultrasound as described in a previous section. By applying these techniques Computer vision has already shown to be capable of supporting reduced diagnostic uncertainty and as such increase the accuracy of diagnosis and improve workflow, it has also shown to reduce variation that characterises human-based diagnostics [232].

Due to how ultrasound is beamformed into an image, pre-processing player a large role in image enhancement. In order to produce images, ultrasound uses image forming algorithms that take the response from each individual transducer in the array and typically logarithmically compresses the amplitude and interpolates the responses

together to form an image. Due to this process the signal to noise ratio of the resulting image can be quite low, causing speckle noise artifacts [233, 234]. Modern beamforming techniques such as dynamically focused transmission and reception [235] attempts to increase depth of field without reducing lateral resolution by using a montage process, which can increase contrast, resolution and depth of field. The aperture of the transducer is weighted using a technique known as apodization examples of which can be done using FIR filtering [236] and weighted least-squares filter [237]. As computing technologies have advanced, ultrasound beam and imaging pre-processing has allowed for the forming of high quality images from ultrasound scan data [233]. Despite this improvement in image quality a number of post processing techniques are often still required to improve image quality, such as using filters. There are a number of methods that can be used to improve image quality such as: kernel based convolution [238-240], which can be used to enhance details within the image matrix; and adaptive histogram equalisation [241] to enhance images where contrast may be too low. Filters can also be used such as gaussian, median, anisotropic diffusion, but Bouma et al [242] suggests that the success of these methods are directly associated with the quality of the data and less successful on low quality data.

Ultrasound processing methods that enhance image quality are often subject to limitations that effect performance in other areas. Employing techniques that improve spatial resolution allow for better visualization of fine details but it reduces penetration depth, therefore when scanning deeper tissues, a compromise between spatial resolution and penetration [243]. Similarly, there is a trade-off between frame rate and image

quality, the speed of ultrasound through tissue is limited, meaning that at faster frame rates beam forming algorithms must form images with less data reducing the quality [244, 245]. Contrast agents allow for enhanced detection of flow within tissues ultrasound imaging, providing many important details for diagnostic assessment but may have contraindications in some patients, are present in the target tissues for only a limited time and potentially can cause artifacts [246].

These were just a selection of methods of pre- and post-image processing, which will continue to be refined using a wide variety of techniques. Image processing is an initial step within a complex process towards the continuous improvement of clinical ultrasound imagery for use in machine learning [233, 234, 247]. While these methods will improve the quality of the image being examined an additional step is to process the data itself.

2.4.3.2. Feature Extraction

Early classification and recognition algorithms were shown to be more effective at recognising features after segmentation and feature extraction, as this would allow the model to be presented with a smaller, normalised set of features to examine as such is most likely to make an accurate prediction. Machine learning algorithms are often heavily dependent on the feature engineering and extraction methods involved in the process. Feature engineering and extraction focuses on selecting, manipulating and transforming raw data into features that can be used in machine learning. These methods often enhance existing features within the dataset such as texture [248, 249], intensity

(brightness) [250], shape [251], gradient [252], wavelet transformation [253, 254], and histograms [255].

Intensity thresholding is one of the simplest forms of segmentation, where the image values are categorised across a stepped range of values, this can be done globally across the entire image or locally based upon the neighbouring values as with a kernel. Horsch et al [256] used thresholding as part of a process to segment tumours in pre-processed breast tissue scans in order to distinguish between benign and malignant masses, this work was continued in Drukker et al [257] examining how this technique could be used for further feature extraction and differentiation. Texture matching has been shown to be successful at recognising features even where ultrasound speckle may confound shape algorithms [147, 258, 259]. The Hough Transform filter has been shown to be successful at extracting long linear shapes within ultrasound imagery [260].

2.4.4. Training Methodologies

Machine learning requires training to fit the model to the characteristics of the dataset, this is typically done via supervised or unsupervised learning methodologies, with elements of both methodologies may be utilised dependant on the composition of the dataset (Figure 2.17), additional methods such as reinforcement and transfer learning are regularly used for their unique attributes to support the training process [176, 261, 262].

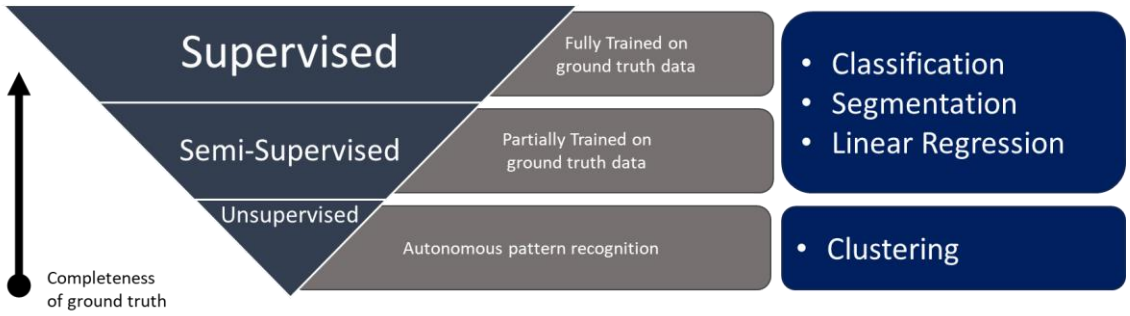


Figure 2.17 - Representation of ground truth data requirement for training method and typical tasks associated with that methodology.

In the context of a classification task, supervised learning typically uses a dataset where there is already a known classification. For example, in the case of detection and classification of cancer cells there is already an accepted methodology for classifying cancerous cells in clinical use. As such, it is common to use these well-defined classifiers when examining and training for clinical datasets [176, 261]. The model changes the weighting of its ruleset in an attempt to reduce the size of the error after each training iteration. An important attribute to note when using supervised learning is ensuring that the sample set of classifications for the training data is representative of the task it will be required to perform. Because the algorithm looks for the lowest rate of error not for the highest accuracy, it is therefore important to ensure that the dataset is balanced. Where there is an imbalance in classifiers within a training set the model will be biased towards that particular classifier [263]. This is an ongoing challenge within machine learning as many databases (such as medical imaging databases) are unbalanced. Using methodologies such as oversampling and under-sampling [264] has shown good results for improving the accuracy of unbalanced datasets [265-268]. The

aim of training is to detect generalisations for each category but there is also a risk of overfitting, where the model is weighted towards the attributes of the dataset rather than generalisations within the data itself, an overfit model will show high accuracy on the training set but low accuracy on validation [269]. Overfitting could occur due to a number of factors such as overuse of the training set, speed of training, complexity of data and under-variance [269-271]. Overfitting is a complex problem, with multiple well established, accepted methods to lower its occurrence, such as using: dropout, i.e. randomly dropping neurons during training to prevent co-dependency [272]; effective backpropagation [273, 274]; and, using regularisation techniques such as applying randomisation [275] although there are studies that suggest this has a limited effect [276].

Where labelled data for supervised training is not available an unsupervised learning methodology may be used, instead the model will examine the dataset and cluster these datapoints into categories such as through a method of pattern classification [176, 277, 278]. While these categories will not be externally validated for accuracy, there is the potential for useful categorisation or prediction to be made dependant on the mathematical formulas used within the model [279, 280].

Transfer Learning encompasses a subset of machine learning methodologies focused on the transfer of knowledge across domains [281], one method (as shown in Figure 2.18) uses a model has already been trained on a large dataset for a specific task, such as image classification, as a starting point, the pre-trained model's parameters are then fine-tuned or adjusted by training on a smaller more specific dataset for a specialist task. For

example, a generalised model could be trained to recognise a tumour, then a more specialised model could be trained using that initial trained recognition with additional categorisation parameters to recognise different types of tumours. This transfer of pre-learned generalised behaviour into more specialised models potentially allows for faster training as the complexity of the training required is reduced and will build upon the accuracy of the previous model without requiring as large a dataset as the initial training, although this is also a notable drawback with any bias, error or overfitting potentially present in the previous model also passed onto the new model [282-285]. Transfer Learning can also be used in conjunction with a semi-supervised approach (as seen in Figure 2.17), leveraging pre-trained networks to make assumptions about unlabelled data, effectively clustering these unlabelled samples which can then be labelled using these predictions [286].

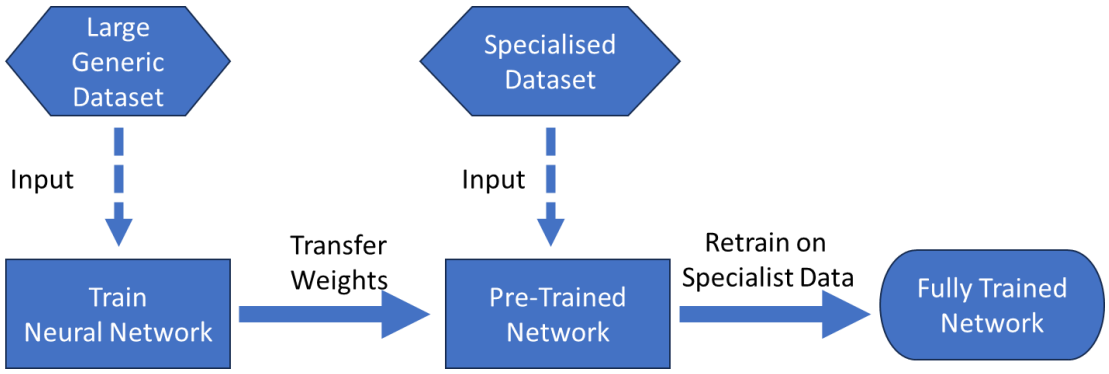


Figure 2.18 – Simplified flow chart of transfer learning from pre-trained neural networks.

2.4.5. Criticisms of Machine Learning Research

While machine learning for medical imaging has great potential to revolutionise diagnostics with ever increasing numbers of researchers focusing on this subject area, there is a great deal of criticism surrounding methodologies being used in this research. Robert et al. [287] published a systematic review of 62 studies of machine learning for COVID-19 published in the period between 1 January 2020 to 3 October 2020 and identified not a single one had the potential for clinical use. There is also a suggestion that machine learning could entrench poor practice and exacerbate bias, potentially promoting inequity through the processing of datasets [288]. Many critics of automation and machine learning in medicine cite that these methods are deskilling clinicians [289], in a study of 50 mammographers [290] there was 14% decrease in diagnostic sensitivity when readers were presented with challenging cases marked by computer-aided detection. Another study showed a 9% decrease (down to 48%) in clinician diagnostic accuracy when reading electrocardiograms (EKGs) when inaccurately labelled by machine [291].

One of the major criticisms of machine learning research for medical research is that while there are increasing publication of papers showing state-of-the-art performance on benchmark data, there is rarely any practical improvement towards solving the clinical problem [292]. Poor implementation and data leakage is a major problem with many researchers training and testing on the same dataset leading to circular analysis [293], or failing to ensure test/train splits are performed at patient level leading to algorithms to recognise individual patient markers rather than the disease itself [294]. Exclusion

criteria within the initial data processing of the dataset can play an important role in the accuracy of a machine learning output, as seen in Caruana et al [295], The initial results of this study suggested that there was a lower risk of death from pneumonia if the patient also had a history of asthma, while this neural network yielded 'state of the art accuracy' the result itself was incorrect: asthma is one of the largest risk factors associated with death from pneumonia [296].

In order to attempt to standardise publication of machine learning studies a checklist for artificial intelligence in medical imaging was published to assist researchers and reviewers in validating medical machine learning research [297]. There is also a push towards explainable AI, away from 'black box' methods so that this is especially true in medical imaging where such step by step accountability is important due to medical liability [298].

Many of these criticisms stem from core challenges faced in applying AI to real world problems:

- Medical information is often time sensitive, for example medical emergencies may necessitate real-time scanning with anomalies being detected and displayed in real time [299-301].
- The trustworthiness and interpretability of AI algorithms remains a major challenge, for AI to be used to support clinical decision making, clinicians require transparent, explainable output that they can clearly understand [302].
- Generalizability and Domain Adaptation, remains a complex problem in medical imaging, while the anatomy and physiology of the body is often consistent,

modalities may display the bones, tissues and organs in a radically different way, research is continuing to explore how different domains can be leveraged [303].

- In the case of rare medical conditions data can be scarce, limiting what can be done with AI [304].
- Regulations surrounding information governance of medical data is highly strict, limiting the ability to freely share medical imaging data. Edge computing offers a potential solution where by each clinical centre can maintain governance over its own data while collaborating over the development of a AI model via cloud technologies [305, 306].
- There are still many questions surrounding the efficacy of clinical adoption of AI, with regulatory bodies slow to provide guidance as to how to develop and deploy algorithms [307]. The legal risks of using AI in medical diagnosis are still being quantified with much work still to be done to quantify how AI should be used safely in the medical field [308].

2.5. Conclusion

This literature review has given an overview of diagnostic medical imaging based on the cross sections of the Japanese ultrasound abdominal protocol, in order to ground readers for future technical chapters. Provided an overview of the history of medical ultrasound and alternative modalities. A history and overview of machine learning was also provided before an in-depth look at deep learning history, methods, image processing requirements, training methodologies and common criticisms of medical machine learning.

There are many potential targets for applying machine learning to abdominal ultrasound due to the sheer level of complexity of each anatomical and physiological system. This thesis focuses on developing techniques to relieve the burden placed on the sonographer by providing machine learning solutions to improve the quality of data collection, reduce the operator effort required, and ensure adherence to clinical protocols and methods.

Research on machine learning-based abdominal ultrasound cross-sectional classification has been limited, therefore, it was important to establish a baseline neural network response to training on the Japanese abdominal ultrasound protocol, this was used to identify problems with classification and provide a benchmark for future results.

Chapter 3

Transfer Learning for Classification of Standard Ultrasound Abdominal Cross Sections using Neural Network Architectures

Abstract

Abdominal ultrasound screening requires the capture of multiple standardised plane views as per clinical guidelines. Currently, the extent of adherence to such guidelines is dependent entirely on the skills of the sonographer. The use of neural network classification has the potential to better standardise captured plane views and streamline plane capture reducing the time burden on operators by combatting operator variability. This chapter examines the effectiveness transfer learning through the use of 9 neural networks pre-trained on the ImageNet dataset, these networks were then trained to recognise 16 abdominal ultrasound cross sections from 64 patient sets to establish a baseline response. The highest validation accuracy was attained by both GoogLeNet and InceptionV3 is 83.9% using the pre-trained networks and the large sample set of 26,294 images. A top-2 accuracy of 95.1% was achieved using InceptionV3. Alexnet attained the highest accuracy of 79.5% (top-2 of 91.5%) for the smaller sample set of 800 images. The neural networks evaluated in this chapter were also successfully able to identify

problematic individual cross sections such as between kidneys, with right and left kidney being accurately identified 78.6% and 89.7% respectively. A further case study of mobile and small sized networks confirmed that small efficient networks could be highly effective for Ultrasound classification. This chapter builds upon existing studies, demonstrating the potential accuracy of multiple neural network architectures when classifying standard abdominal cross sections. More complex neural networks provided only limited improvement to classification accuracy with a difference of just 2.2% between the top results of the nine networks tested. Dataset size proved a more important factor with more complex neural networks providing higher accuracy as dataset size increases and simpler linear neural networks providing better results where the dataset is small.

3.1. Introduction

Diagnostic medical ultrasound offers a low risk, non-invasive method to examine anatomical features. While this often takes the form of real time diagnostic procedures, it can also facilitate longer term screening and monitoring [309]. Ultrasound has seen widespread adoption throughout healthcare due to the broad range of applications and accessibility of ultrasound equipment, especially in mid to low-income countries where access to other modalities may be limited [310, 311]. Diagnostic ultrasound within the abdominal-pelvic region has already seen extensive adoption in obstetrics with the widespread use of routine foetal monitoring in pregnancy [312] and in cardiology [313], but there is now a growing trend to develop techniques for a greater range of abdominal organs and procedures [310].

In the case of abdominal screening, the sonographer follows clinical guidelines to capture specific standard clinical cross sections of anatomical features for monitoring or reporting purposes. The obligation to ensure adherence to these guidelines lies with the individual sonographer and as such the precision with which these images are captured are subject to the attentiveness, knowledge, and experience of the individual [314]. However, there is a shortage of adequately trained sonographers meaning that these clinical protocols are often performed by users with limited training and experience [315]. The use of machine learning to aid the user in plane selection has the potential to reduce the variation caused by the operator while also reducing the time taken for the procedure due to inexperience, therefore improving workflow and patient comfort.

Image classification is a fundamental component of medical imaging research, of which deep learning is an increasingly popular subject of interest [316-318]. If we examine modern digital medical imagery at its fundamental code level, it is essentially a matrix containing information such as brightness and colour similar to that of other image files. Current machine learning frameworks such as PyTorch [319] use a more simplistic form factor such as a tensor to provide a standardised form factor that a computer algorithm can use to detect the intended features and make predictions for classification tasks. Despite being one of the most widely used medical imaging modalities in the world, ultrasound has seen comparatively little interest from deep learning research in comparison to Xray, CT and MRI [320]. This is partly due to the limited availability of ultrasound imaging data, as commonly imaging and diagnosis are performed in real time by an expert sonographer with only a limited amount of data recorded. As seen in a recent review by Avola et al [321], there are very few publicly available clinical ultrasound datasets available in comparison to other modalities where an expert clinician may examine and report on scans hours or even days later. Another contributing factor is that medical ultrasound images provide a far more restricted window of information in comparison to other medical scans. Ultrasound is produced by measuring the reflected ultrasound waves detected by a small piezoelectric array within the ultrasound probe [47], such images are typically two-dimensional, low contrast, and subject to interference such as attenuation and shadowing that can hinder classification even for experienced sonographers [51, 53].

Neural networks have already been successfully applied to many classification tasks within medical diagnostic ultrasound such as the detection of masses for cancer diagnosis [216, 322], thyroid nodules [323, 324], liver anomalies [325, 326], spine [327] and cardiac and aorta cross sections [328-330]. More generalised approaches to classification have also been explored with attempts to identify abdominal cross sections [331-333] and most commonly for foetal and obstetric cross sections [334-336]. This study uses a subset of transfer learning, that uses a neural network that has been previously trained on a larger, often more generalised dataset as its template. This has previously been shown to overcome many of the problems associated with small datasets such as poor generalisation and overtraining as is often the case with ultrasound [337, 338].

Previous studies examining classification of abdominal cross sections with machine learning are limited. Cheng & Malhi [331] proved the effectiveness of transfer learning using the ImageNet challenge dataset [339] with the successful classification of 11 standard ultrasound cross sections attaining accuracies of 77.3% using CaffeNet and 77.9% for VGGNet both of which exceeded the 71.7% accuracy achieved by a radiologist. Xu et al [340] examined classification of 11 ultrasound abdominal cross sections as part of a wider study on landmark detection, the Single-task learning (STL) ResNet-50 attained an accuracy of 81.22% in comparison to the radiologist who achieved 78.87%. Reddy et al [333], tested a number of neural networks on 6 visually distinct abdominal cross sections achieving an accuracy of 98.77% using a ResNet-50.

This chapter further expands the study of abdominal ultrasound plane classification by examining 16 abdominal cross sections. This work shall examine the effectiveness of transfer learning for a small ultrasound abdominal plane dataset, providing comparative accuracy data for a larger number of neural network architectures on standard abdominal cross sections than has been previously studied. This will serve both to aid selection of neural networks in future, but also further highlights the potential uses and difficulties of utilising deep learning for identifying and classifying abdominal cross sections.

3.2. Structure and Scope

This chapter establishes a baseline for the 16 upper abdominal cross sections as defined by the Japanese abdominal screening protocol [112]. This protocol was chosen due to its overlapping coverage of the upper abdomen, which would underline and potential difficulties applying deep learning to complex ultrasound abdominal protocols. While the Japanese abdominal screening protocol includes pelvic and bladder scans, these were excluded from this study to focus on the upper abdomen. This is the first study of machine learning classification of the full Japanese abdominal protocol and as such it is important to study network response to this dataset, to ascertain future requirements. This chapter focuses on the three most popular CNN architectures, those being the Linear, Residual, and Inception, using the method outlined in Figure 3.1. The results of this initial study suggested high accuracies could be achieved with smaller, less complex neural networks. A subsequent follow-on study analysed the performance of four popular small scale neural networks designed for mobile applications was also assessed.

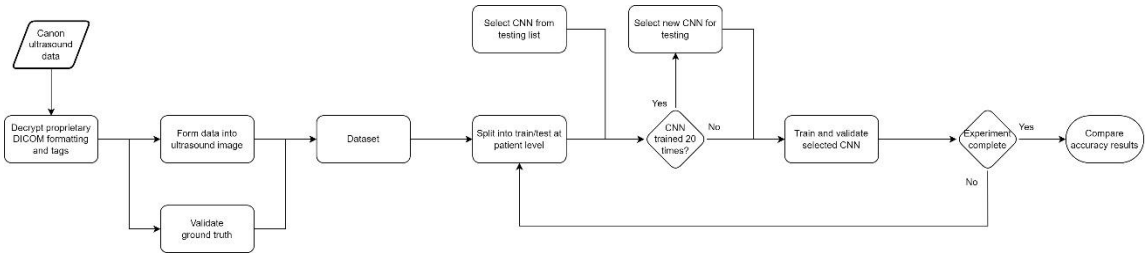


Figure 3.1 - Flow Chart showing methodology used in this experiment. Data is extracted from DICOM files and prepared for use. Train/test split performed, and CNN selected. All neural networks are then trained, and a new train/test split performed.

3.3. Method

3.3.1. Ultrasound Data Acquisition:

The ultrasound data is part of a private dataset provided via Canon Medical Europe and was captured using a Canon TUS-AI800 [341] using a curved linear array, with each of the 16 cross sections (examples of which are displayed in Figure 3.2) classified at the time of capture by a single experienced sonographer. While the data is anonymous, acceptance criteria were that participants be of adult age with no underlying pathology detected by the sonographer that may influence the results at time of recording. The sonographer strictly adhered to the standardised capture method defined by the Japanese society of sonographers [112], starting the scan in the location defined within the method and progressively sweeping through the region of interest ensuring complete coverage of the defined target anatomy. The ultrasound data was recorded as a stream of 8-bit greyscale images of varying length (between 14 and 46 seconds), these sequences were effectively raw ultrasound images and contained no text or graphical annotation

from the User Interface. These were then stored in a DICOM format [342] and anonymised before being provided for use in this work.

The dataset consists of 64 patient studies with 16 recorded anatomical cross sections each for a total of 33,093 individual images. These patient studies were split 50/14 (approximately 80/20 split) between training and test sets, both training and test sets were resampled at the patient level for each training run for cross validation purposes. Although this significantly reduces the pool of possible test images it was done to ensure no data leakage that could artificially inflate results. Because the resampling must be performed at patient level, a holdout method was chosen over folded cross validation to ensure validity of the test data. As the neural networks that form the basis for the transfer learning experiment are trained on 3 channel RGB images, the single channel greyscale images were duplicated into three channels during the process to convert the image into tensors, with only a negligible drop in performance noted. The full image was used with no cropping or adjustment beyond minor contrast normalisation using the standard method provided in PyTorch in order to ensure standardisation across the imagery.

Two training sets were produced alongside a single test set as reported in Table 3.1. The first training set was produced to provide a balanced, idealised dataset by defining a single image frame (an example of which can be seen in Figure 3.2) from each set of cross sectional sweeps for a total of 800 images, this was done to simplify the problem space, while in many cases a sonographer must move the probe to fully visualise the region of interest, reduction to a single ideal cross section provides the neural network with the most opportunity to make the correct prediction. The second training set takes

into account the entire sonographic sweep and as such essentially consists of multiple short videos centred on the correct region of interest during examination and is made up of 26,294 images, this data contains significant repetition, minor deviations such as changes in attenuation, shadowing, natural physiological changes, and the slight movements of the patient and sonographer that occur naturally during clinical examination. This provides a more realistic training set but also significantly increases the complexity of classification. The test set consists of 224 images with each of the 16 cross sections represented by 14 precise images. Those images and videos corresponding to the test set were excluded from all training datasets.

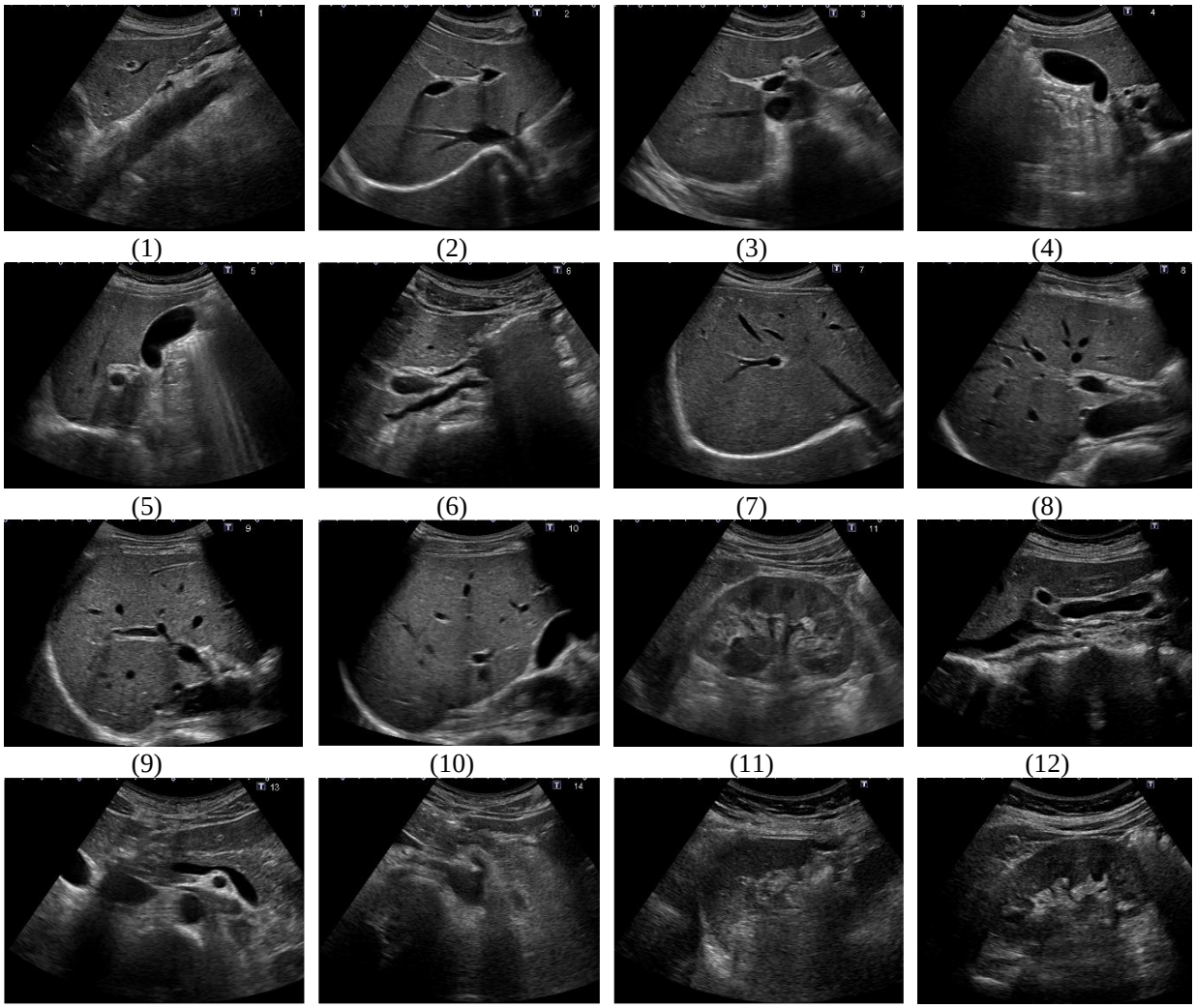


Figure 3.2 - Example of the 16 ultrasound abdominal cross sections.

Table 3.1 - Identified plane categories in training and validation sets.

Abdominal Cross Section	Training Set 1	Training Set 2	Validation Set
1. Epigastric sagittal scan: Liver/aorta	50 (6.3%)	1478 (5.6%)	14 (6.3%)
2. Epigastric horizontal scan to right subcostal scan: Hepatic vein	50 (6.3%)	1722 (6.5%)	14 (6.3%)
3. Right Epigastric oblique scan: Horizontal portal vein	50 (6.3%)	1605 (6.1%)	14 (6.3%)
4. Right Subcostal scan: Gallbladder	50 (6.3%)	1545 (5.9%)	14 (6.3%)
5. Right hypochondrium vertical scan: Gallbladder	50 (6.3%)	1539 (5.9%)	14 (6.3%)
6. Right hypochondrium vertical to oblique scan: Extrahepatic bile duct	50 (6.3%)	1575 (6%)	14 (6.3%)
7. Right subcostal scan: Liver	50 (6.3%)	1528 (5.8%)	14 (6.3%)
8. Right intercostal upper scan: Liver	50 (6.3%)	1558 (5.9%)	14 (6.3%)
9. Right intercostal mid scan: Liver	50 (6.3%)	1670 (6.4%)	14 (6.3%)
10. Right intercostal lower scan: Liver	50 (6.3%)	1609 (6.1%)	14 (6.3%)
11. Right intercostal scan: Right kidney	50 (6.3%)	1516 (5.8%)	14 (6.3%)
12. Epigastric vertical scan: Extrahepatic bile duct/pancreas	50 (6.3%)	1717 (6.5%)	14 (6.3%)
13. Epigastric horizontal scan: Pancreas	50 (6.3%)	1886 (7.2%)	14 (6.3%)
14. Epigastric oblique scan: Pancreas	50 (6.3%)	1972 (7.5%)	14 (6.3%)
15. Left intercostal scan: Spleen	50 (6.3%)	1759 (6.7%)	14 (6.3%)
16. Left intercostal scan: Left kidney	50 (6.3%)	1615 (6.1%)	14 (6.3%)
Total	**800**	**26294**	**224**

3.3.2. Neural Network Architectures

The experiment was performed on a computer with an Intel CPU with a clock speed of 2.4 GHz and a Nvidia 20 series GPU using the PyTorch framework [319] and Cuda toolkit (version 11.6). As with previous literature [331, 333, 340] publicly available neural networks pre-trained on the ImageNet challenge dataset [339] were used as the basis for transfer learning. The neural networks architectures chosen for this experiment can be classified by the principles behind their design. These being two linear convolutional neural networks (Alexnet [343, 344], VGGNet [345]), five residual networks (ResNet-18, 32, 50, 101, 152) [346], and two inception networks (GoogLeNet (Inception V1) [347] and InceptionV3 [348]). A summary of the exact number of layers and parameters used by the neural networks is provided in Table 3.2. These neural networks were chosen as typical examples of their respective architectures, with five residual networks evaluated to test how the depth of residual network effects network response to ultrasound data. Three training procedures were used: transfer learning using dataset 1, transfer learning using dataset 2, and a baseline using only training dataset 2 without pre-trained transfer learning weightings being applied at initialisation. Training used the ADAM optimiser [349] with an initial learning rate of $1x10^{-4}$ with the learning rate degrading every 5 steps, over 20 epochs. The ADAM optimiser was chosen as it scales is well suited for sparce and noisy data. It calculates a moving average of the first-order moments (the mean of gradients) and the second-order moments (the uncentered variance of gradients), allowing for fast and efficient convergence. Each network was trained 20 times with the training and test sets resampled for each training run in order to

benchmark performance while reducing performance variation from any single training run. The final layer of each neural network was adjusted from 1000 to 16 in order for the neural networks to perform the required classification task, no additional changes were made from the standard network architecture.

The earliest examples of convolutional neural networks use a linear stack of convolution, pooling and connected layers to form a hierarchical design, similar to that found in the visual cortex. This classic linear design is demonstrated in the Alexnet [343, 344] architecture which successfully competed in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 challenge achieving the lowest top-5 test error rate of 15.3% despite only consisting of 8 layers. Inspired by the Alexnet architecture, the VGGNet [345] architecture secured second place in the ILSVRC in 2014 challenge, this was achieved by producing a deeper neural network by stacking additional convolution layers such as VGG-16 which contains 13 convolution layers followed by three interconnected layers. In order to modify these linear models for use, the SoftMax function (multinomial logistic regression) was adjusted to transition the model to the required classifiers. While there are many variations of linear neural networks available the two networks exampled offer a reasonable benchmark against the other neural network architectures.

Table 3.2 - Summary of Neural Network Shape and Parameters

Model	Method	Convolution	Fully Connected	Parameters
Alexnet	Linear	5	3	57,069,392
VGG16	Linear	13	3	134,326,096
GoogLeNet	Inception	22	1	11,996,288
InceptionV3	Inception	48	1	25,145,048
ResNet18	Residual	18	1	11,184,720
ResNet34	Residual	34	1	21,292,880
ResNet50	Residual	50	1	23,540,816
ResNet101	Residual	101	1	42,532,944
ResNet152	Residual	152	1	58,176,592

Linear neural network architectures suffer from a significant drawback, known as the vanishing gradient problem, as the number of layers within the network is increased there is a significant decrease in performance as the size of the gradient is halved within rectified linear unit layers. As the network back-propagates up through the layers of parameters the size of the gradient decreases with each additional layer, effectively decreasing the effectiveness of backpropagation with each additional layer. This limits the useful depth possible with linear architectures without significant augmentation [350, 351].

In order to facilitate neural networks with deeper architectures an innovative design method was required. One such solution was residual networks or ResNet [346], which creates feature maps of specific residual identifiers from a layer. These residual feature maps are propagated higher up the neural network with each training epoch effectively

creating shortcuts within the model. The use of multiple ResNet depths tests the theory that ResNet depth should not adversely affect model accuracy, as such this experiment uses Resnet networks of 18, 34, 50, 101 and 152 layers to confirm this did not affect performance significantly.

The Inception architecture uses a modular design approach to mitigate the vanishing gradient problem, in GoogLeNet (Inception V1) [347] convolution layers are clustered together into modules instead of activated linearly. Auxiliary networks were also added to train in conjunction with the main network branch allowing the model to cross validate and enhance the gradient at these intervals. InceptionV3 [348] further refined this method by providing additional batch normalisation and increased tensor size to 299x299. Results from version 1 and version 3, as well as the other highlighted architectures, are analysed in this work.

3.3.2.1. Mobile Networks

Four networks mobile or small-scale networks from residual and modular architectures (as seen in Table 3.3) were selected for testing: MobileNetV2, MobileNetV3, EfficientNet, and SqueezeNet. All models had been pre-trained on the ImageNet challenge dataset [339] and were tested solely on Dataset 2 using the same training parameters as the previous set of neural networks.

MobileNetV2 [352] and MobileNetV3 [353] designed by Google, are based on residual architecture design but replaces a number of convolutions into pointwise 1x1

convolutions, as well as implementing residual shortcuts similar to those seen in ResNet models, allowing for enhanced performance at reduced network sizes.

EfficientNet [354] is derived from the architecture of MobileNetV2 but the structure has been inverted and uses scaling coefficients to scale the network width, depth and resolution allowing for networks to be customised to the required task without editing the structure of the network itself, allowing for the use of fewer parameters to be used dependent on scaling and required task.

SqueezeNet [355] is a modular convolutional neural network that 'squeezes' parameters using 1x1 pointwise convolutions, decreases the number of input channels throughout the network, down samples late so that convolutional layers have large activation maps, which essentially compress the model while maintaining network detail.

Table 3.3 - Mobile networks shape and Parameters selected for case study.

Model	Method	Convolution	Fully Connected	Size (Mb)	Parameters
MobileNet V2 [352]	Residual	53	1	8.81	3,400,000
MobileNet V3 [353]	Residual	28	1	16.3	5,400,000
EfficientNet [354]	Residual	237 (scalable)	1	15.6	11,000,000
SqueezeNet [355]	Modular	18	1	2.81	421,098

3.4. Results.

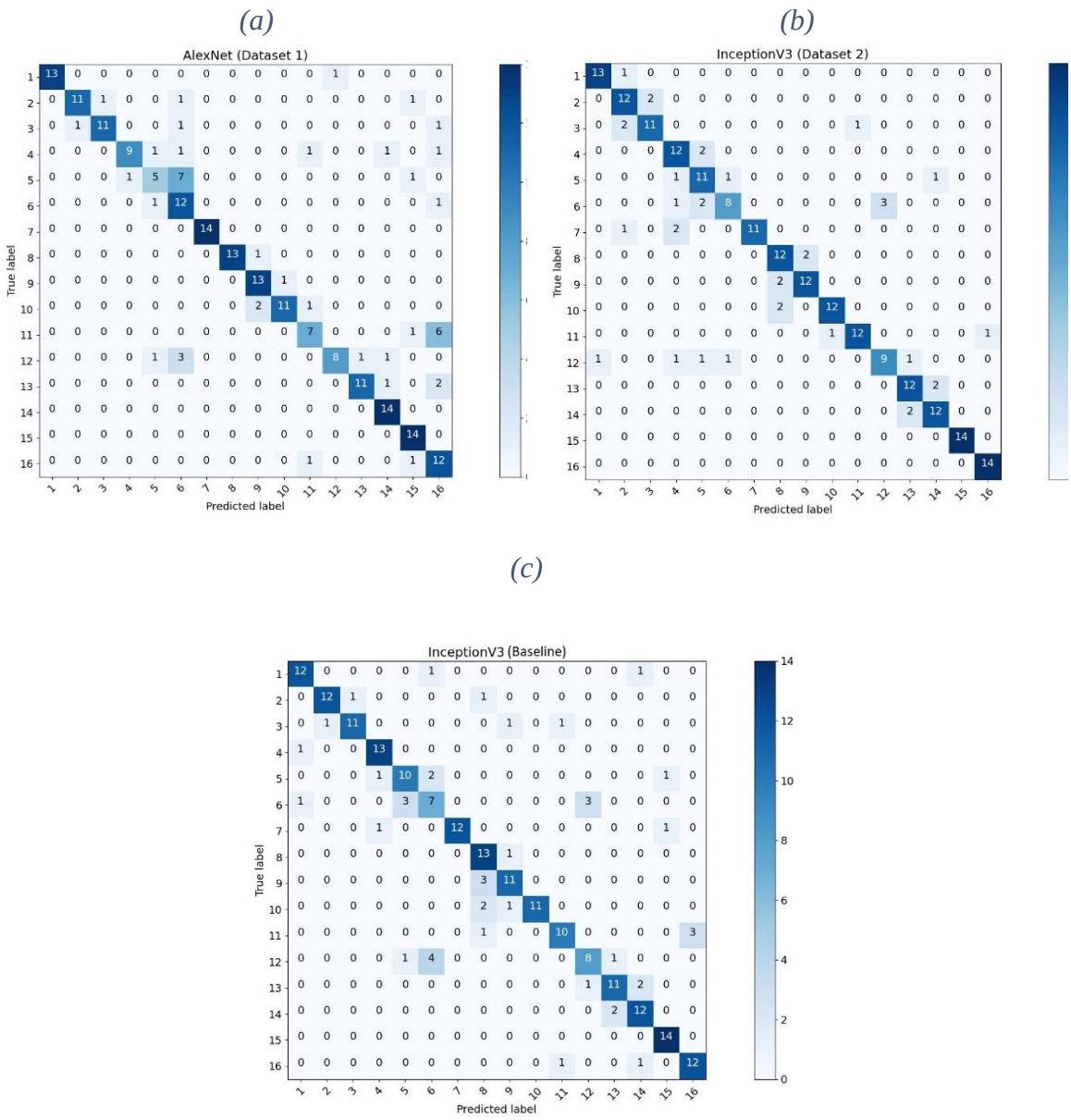
The results for highest single neural network accuracy of the nine neural networks (as shown in Table 3.4) show that the Inception architecture achieved the highest accuracies on the test set for both networks pre-trained on the ImageNet dataset and then trained with dataset 2 and the Baseline, with GoogLeNet (InceptionV1) and InceptionV3 attaining the top result of 83.93% for dataset 2, with inceptionV3 attaining 79.91% and GoogLeNet 77.68% for the Baseline. Linear neural network architectures attained the highest results for dataset 1 with Alexnet achieving 79.46% and 77.23% for VGG16.

Table 3.4 - Highest validation accuracy achieved after 20 epochs from nine neural networks over 20 training runs.

	Alexnet	**VGG16**	**GoogLeNet**	**InceptionV3**	**ResNet 18**	**ResNet 34**	**ResNet 50**	**ResNet 101**	**ResNet 152**
Baseline Accuracy	69.20%	70.09%	77.68%	79.91%	75.06%	73.66%	73.21%	71.88%	71.43%
Dataset 1 Accuracy	79.46%	77.23%	62.05%	71.88%	67.41%	73.21%	73.21%	70.98%	70.54%
Dataset 2 Accuracy	80.80%	82.59%	83.93%	83.93%	83.04%	83.48%	83.48%	82.14%	83.04%

The confusion matrix in Figure 3.3 confirms that the largest misclassification errors are: between cross sections within close proximity such as cross sections 8, 9 and 10 which focus on the liver; where anatomical structures overlap such as in cross sections 5 and 6

which focus on vertically oriented biliary system, as well as 6 and 12 which the bile duct is a significant landmark; and differentiating between the kidneys in cross sections 11 and 16.



1. Epigastric sagittal: Liver/aorta, 2. Epigastric horizontal: Hepatic vein, 3. Right Epigastric oblique: Horizontal portal vein, 4. Right Subcostal: Gallbladder, 5. Right hypochondrium vertical: Gallbladder, 6. Right hypochondrium vertical: Bile duct, 7. Right subcostal: Liver, 8. Right intercostal: Liver, 9. Right intercostal: Liver, 10. Right intercostal: Liver, 11. Right kidney, 12. Epigastric vertical: Bile duct/pancreas, 13. Epigastric horizontal: Pancreas, 14. Epigastric oblique: Pancreas, 15. Spleen, 16. Left kidney

Figure 3.3 - Confusion Matrix for top performing Neural Networks: (a) Alexnet Dataset 1, (b) InceptionV3 Dataset 2, (c) InceptionV3 Baseline Dataset

While assessment of the highest accuracy provides a baseline understanding of neural network response, examining top-2 accuracy where both the first and second highest prediction is taken into account, allows for the difficulty of the image classification task to the scaled by reducing errors due to class ambiguity. Taking this ambiguity into account provides a potentially clearer picture of overall network response. Top-2 accuracy results (shown in Table 3.5) continue the trend with InceptionV3 attaining the highest top-2 accuracy of 92.86% for Baseline with the second-best result being GoogLeNet with 90.18%. The linear architectures attained the highest top-2 accuracy in dataset 1 with Alexnet attaining 91.52% and 90.18% for VGG16. InceptionV3 also achieved the highest top-2 for dataset 2 at 95.09% but ResNet 18, 34 and 50 jointly attained the second-best result of 94.64%. The neural networks with the highest overall accuracy did not correspond to that of the highest top-2 accuracy. Those that did match were ResNet101, ResNet152 for Baseline; Alexnet, VGG16 and GoogLeNet corresponded for Dataset 1 and VGG16, ResNet34 and ResNet50 for Dataset 2. The neural networks with the highest overall accuracy did not correspond to that of the highest top-2 accuracy. Those that did match were ResNet101, ResNet152 for Baseline; Alexnet, VGG16 and GoogLeNet corresponded for Dataset 1 and VGG16, ResNet34 and ResNet50 for Dataset 2.

Table 3.5 - Highest top-2 validation accuracy attained accuracy after 20 training runs.

	Alexnet	VGG16	Google Net	InceptionV3	ResNet 18	ResNet 34	ResNet 50	ResNet 101	ResNet 152
Baseline Top-2	86.16%	83.04%	90.18%	92.86%	87.95%	87.50%	89.29%	87.05% *	87.05% *
Dataset 1 Top-2	91.52% *	90.18% *	79.46% *	88.84%	84.38%	88.84%	87.05%	86.16%	87.05%
Dataset 2 Top-2	92.86%	93.75% *	94.20%	95.09%	94.64%	94.64% *	94.64% *	94.20%	93.75%

*** = Accuracy and Top-2 attained from same neural network model.**

The testing algorithm included category specific accuracy results (shown in Table 3.6) allowing for a deeper examination of the strengths and weaknesses of ultrasound plane categorisation. When examining the plane specific categorisation results from the InceptionV3 neural network trained from Dataset 2 it was possible to correctly categorise the right kidney plane 78.57% and the left kidney plane 89.71% of the time suggesting sufficient visual information is available to achieve successful classification. When examining the overall performance of transfer learning with Dataset 2 (from Table 3.6), the cross sections with the lowest accuracy were plane 6 (Right hypochondrium vertical to oblique scan: Extrahepatic bile duct) with an average accuracy of 64.29%, and Plane 12 (Epigastric vertical scan: Extrahepatic bile duct/pancreas) with an average of 67.46%. These cross sections see the highest error in each of the three exampled confusion matrixes, this is likely due to intersecting anatomical structures within the plane classifiers.

Table 3.6 - Accuracy of Individual Cross sections: Highest single neural network accuracy trained using Dataset 2

Category	Alexnet	VGG16	GoogLeNet	InceptionV3	ResNet 18	ResNet 34	ResNet 50	ResNet 101	ResNet 152
1. Epigastric sagittal: Liver/aorta	92.86%	100%	100%	100%	100%	100%	100%	100%	100%
2. Epigastric horizontal: Hepatic vein	78.57%	92.86%	85.71%	78.57%	92.86%	85.71%	92.86%	85.71%	92.86%
3. Right Epigastric oblique: Horizontal portal vein	92.86%	78.57%	78.57%	78.57%	78.57%	78.57%	85.71%	85.71%	85.71%
4. Right Subcostal: Gallbladder	71.43%	64.29%	78.57%	78.57%	71.43%	85.71%	71.43%	64.29%	78.57%
5. Right hypochondrium vertical: Gallbladder	71.43%	71.43%	71.43%	78.57%	71.43%	85.71%	71.43%	78.57%	71.43%
6. Right hypochondrium vertical: Bile duct	71.43%	50.00%	64.29%	71.43%	64.29%	57.14%	64.29%	57.14%	78.57%
7. Right subcostal: Liver	85.71%	85.71%	78.57%	85.71%	92.86%	78.57%	85.71%	78.57%	71.43%
8. Right intercostal: Liver	78.57%	85.71%	100%	85.71%	92.86%	92.86%	85.71%	78.57%	100%
9. Right intercostal: Liver	92.86%	85.71%	85.71%	85.71%	78.57%	92.86%	85.71%	78.57%	92.86%
10. Right intercostal: Liver	78.57%	85.71%	85.71%	92.86%	78.57%	85.71%	85.71%	85.71%	71.43%
11. Right intercostal: Right kidney	64.29%	78.57%	78.57%	78.57%	85.71%	78.57%	92.86%	85.71%	64.29%
12. Epigastric vertical: Bile duct/pancreas	57.14%	71.43%	71.43%	78.57%	64.29%	64.29%	64.29%	71.43%	64.29%
13. Epigastric horizontal: Pancreas	85.71%	100%	85.71%	78.57%	92.86%	100%	100%	92.86%	100%
14. Epigastric oblique: Pancreas	85.71%	78.57%	85.71%	85.71%	71.43%	57.14%	71.43%	71.43%	71.43%
15. Left intercostal: Spleen	100%	100%	100%	100%	100%	100%	100%	100%	100%
16. Left intercostal: Left kidney	85.71%	92.86%	92.86%	85.71%	92.86%	92.86%	78.57%	100%	85.71%
Average Accuracy	**80.8%**	**82.6%**	**83.9%**	**83.9%**	**83%**	**83.5%**	**83.5%**	**82.1%**	**83%**

Examining the variation in training outcome between the 20 runs (detailed in Table 3.7), by calculating the difference between best and worst performing network, shows that in most cases using the full dataset and transfer learning (dataset 2) reduced variation in training result with the exception of ResNet-18 with a variation of 13%. Inception based neural networks achieved the lowest variance with GoogLeNet having the smallest training variation on dataset 2 of 6% and InceptionV2 achieving 7%. Alexnet achieved the highest accuracy for dataset 1 but there was notable variance in the result of 22%, GoogLeNet achieved the poorest overall accuracy but also smallest variance.

Table 3.7 - Variance in training outcome based on the standard deviation for neural networks over 20 runs.

Model	Dataset 1	Dataset 2	Baseline
Alexnet	22%	7%	11%
VGG-16	21%	8%	10%
ResNet-18	19%	13%	8%
ResNet-34	21%	9%	15%
ResNet-50	25%	10%	13%
ResNet-101	26%	10%	15%
ResNet-152	22%	10%	13%
GoogLeNet	9%	6%	14%
InceptionV3	13%	7%	10%

3.5. Discussion

This chapter examined the effectiveness of transfer learning for a small ultrasound abdominal cross-sectional dataset, providing comparative accuracy data for a larger number of neural network architectures on standard abdominal cross sections than has been previously studied. This will serve both to aid selection of neural networks in future, but also further highlights the potential uses and difficulties of utilising deep learning for identifying and classifying upper abdominal cross sections. While the size of the test set is small, this chapter provides a benchmark as to expected performance of neural networks for medical ultrasound classification tasks on 16 upper abdominal cross sections. It has been possible to compare machine learning using a relatively small medical ultrasound dataset of just 26,294 uneven non-ideal samples, with two transfer learning experiments using the ILSVRC data set [339], one leveraging a balanced idealised sample set of just 800 and the other using transfer learning the augment the entire dataset. Optimisation of techniques for convolutional neural networks has seen many improvements with machine learning using the InceptionV3 neural network able to achieve a result of 79.91%, just 4.02% lower than the highest result achieved by transfer learning in only 20 epochs. Furthermore, with transfer learning it was possible to use just 800 samples to train a network to attain an accuracy of 79.46%, just 4.47% from the best result from the larger dataset 2. The use of transfer learning and the complete dataset produced the best result of 83.93% with the result being shared by both Inception neural networks tested.

The residual network architecture did not produce the highest accuracy models (as seen in Table 3.4) but does improve in accuracy as the size dataset increases with results for dataset 2 showing accuracies typically within 1% of the highest result. As previously discussed, residual mapping should have allowed each of the ResNet models to attain similar accuracy results with some variation expected from training randomisation. ResNet 34 and 50 both achieved the highest accuracies of 73.21% for dataset 1 and 83.48% for dataset 2 but ResNet18 achieved the highest baseline accuracy of 75.06%. The difference between highest and lowest performing ResNet neural network was 3.63% for the Baseline, 5.80% for dataset 1, and 1.34% for dataset 2, suggesting that residual mapping struggled with the smaller datasets which would also partially account for the subsequent drop off in accuracy in the larger ResNet-101 and 152 models.

Despite the use of 16 upper abdominal cross sections with many overlapping anatomical structures the top performing neural networks (Table 3.6) achieved an average overall accuracy of 82.94% with greatest error occurring between cross sections containing overlapping identifiers. Where the top-2 accuracy is considered, the neural networks studied achieved an accuracy between 79.46% and 95.09% with the top 10 models being within 2.2% accuracy. The high top-2 accuracy and confusion matrix (Figure 3.3) suggests that while a positive prediction was being made the similarities between cross sections played a major role in reducing accuracy as the majority of errors correspond with cross sections containing the same anatomical structures such as right liver cross sections 8, 9 and 10, cross sections 6 and 12 which both contain the extrahepatic bile

duct as the main region of interest and differentiating the left and right kidneys in cross sections 11 and 16.

The variation in accuracy recorded suggests that larger neural networks benefitted from the larger dataset (dataset 2) and transfer learning the most, ResNet-101 and ResNet-152 displayed notably lower per-plane accuracy results for dataset 1, improved accuracy results for the baseline and then most improved with the addition of transfer learning (dataset 2). While variance itself is less relevant than accuracy as a training metric, neural networks with a smaller variance are more likely to achieve a result closer to the highest accuracy in fewer iterations. Transfer learning can significantly improve accuracy but is no substitute for data. While dataset 1 was too small to provide sufficient information for machine learning to provide a useful result it was capable of producing surprisingly accurate results rivalling the larger baseline dataset and warrants further examination of the effect of ultrasound sample size on neural network learning and generalisation in future works. This study also suggests that the number of layers was less important than dataset size when performing upper abdominal ultrasound plane classification with the difference in accuracy of neural networks for dataset 2 being just 2.2%. Transfer learning also significantly improved neural network accuracy with the larger dataset, when comparing dataset 2 with the baseline, the per-plane training variance is noticeably reduced with the addition of transfer learning along with a significant improvement in accuracy. While dataset size was a more significant factor in reducing variance and increasing accuracy, transfer learning allows for significant

improvements to ultrasound plane classification accuracy where the data is sufficient for the number of parameters in the neural network used.

While there are limitations to the amount of direct comparison that can be made as previous studies used different cross sections, it is possible to highlight a number of trends when classifying abdominal ultrasound data. As seen in Table 3.8, comparing the accuracy results of transfer learning on dataset 2, the overall the results of this chapter are in line with those of previous studies. Smaller networks such as Alexnet achieved an accuracy result just 3.13% lower than the highest accuracy network, show significant potential to classify ultrasound cross sections, CaffeNet (a variant of Alexnet) achieved just 0.6% lower than the significantly larger VGGNet used in Cheng & Malhi [331], and 3.5% lower in the case of Reddy et al [333]. Linear neural network architectures such as these traditionally suffer from the vanishing gradient problem, whereby the size of the gradient is halved in rectified linear unit layer, as the network back-propagates up through the layers of parameters the size of the gradient decreases with each additional layer, effectively decreasing the effectiveness of backpropagation with each additional layer. This limits the useful depth possible with linear architectures in complex without significant augmentation [350, 351].

Table 3.8 - Highest classification accuracy of results in comparison to those previously published abdominal ultrasound studies.

Author	Images	Sets	Cross sections	Model	Average Accuracy
***Cheng & Malhi* [331]**	5,518	185	11	CaffeNet (Alexnet)	77.30%
				VGGNet (VGG-16)	77.90%
***Xu et al* [340]**	187,219	706	11	ResNet50 (STL)	81.22%
Reddy et al [333]	1,906	983	6	Alexnet	95.27%
				VGG-16	97.37%
				VGG-19	98.03%
				GoogLeNet	96.49%
				InceptionV3	97.89%
				Resnet-18	97.37%
				Resnet-50	**98.77%**
				Resnet-101	98.24%
This studies results	26,294	64	16	Alexnet	80.80%
				VGG-16	82.59%
				Resnet-50	83.48%
				Resnet-101	82.14%
				GoogLeNet	83.93%
				InceptionV3	83.93%

As in this study, cross sections containing overlapping landmarks and regions of interest such as the kidneys are shown to be a significant cause of classification error, in Cheng & Malhi [331] and Xu et al [340] both transverse and longitudinal scans of the left and right kidneys cause significant additional classification error, Reddy et al [333] while not containing multiple kidney classifiers, experienced similar error in liver cross sections where the right kidney appeared within the ultrasound image. A small reduction in accuracy can also be noted for larger scale Resnet networks in Reddy et al [333] the

resnet-50 achieved classification accuracy results 0.53% higher than that of the Resnet-101 compared to 1.34% in this chapter. While this would be expected in linear style networks, residual networks create feature maps of specific residual identifiers. These residual feature maps are propagated higher up the neural network with each training epoch effectively creating shortcuts within the model therefore reducing the effect of vanishing gradient [37]. Despite this, results suggest that standard ultrasound data may not have enough visual information to fully utilise networks larger than Resnet-50. The inception architecture uses a modular design approach to mitigate the vanishing gradient problem in GoogLeNet (Inception V1) [347] and InceptionV3 [348] convolution layers are clustered together into modules instead of activated linearly. While more effective in this study, it did not achieve highest accuracy in Reddy et al [333] where results were 2.28% lower for GoogLeNet and 0.88% lower for InceptionV3.

The results presented in this chapter are limited by the size of the test set of 14 patients, containing just 224 samples, necessary to ensure that no data leakage occurred during training. All patient sets are within normal range with no abnormal pathology or underlying conditions noted during ultrasound screening. All images were produced by a single machine, with all classification occurring at time of sampling by a single experienced operator. Only a single manually selected ideal plane image for each of the 16 plane categories was taken, while it would have been possible to take multiple samples from each patient set, there was insufficient differences to warrant including these results with a variance of less than 1% when the sample size was quadrupled.

Despite the use of 16 abdominal cross sections with many potential overlapping anatomical structures the top performing neural networks (Table 3.6) achieved an average overall accuracy of 82.94% with greatest error occurring between cross sections containing overlapping identifiers. Where the top-2 accuracy is considered, the neural networks studied achieved an accuracy between 79.46% and 95.09% with the top 10 models being within 2.2% accuracy. The high top-2 accuracy and confusion matrix (Figure 3.3) suggests that while a positive prediction was being made the similarities between cross sections played a major role in reducing accuracy as the majority of errors correspond with cross sections containing the same anatomical structures such as right liver cross sections 8, 9 and 10, cross sections 6 and 12 which both contain the extrahepatic bile duct as the main region of interest and differentiating the left and right kidneys in cross sections 11 and 16.

When examining per-plane accuracy at the network level, the size of the dataset and use of transfer learning, were significant for reducing the variance in training results. The variation in accuracy recorded suggests that deeper neural networks benefitted from the larger dataset (dataset 2) and transfer learning the most, with ResNet-101 displaying notably lower per-plane accuracy results for dataset 1, improved accuracy results for the baseline and then most improved with the addition of transfer learning (dataset 2). While variance itself is less relevant than accuracy as a training metric, neural networks with a smaller variance are more likely to achieve a result closer to the highest accuracy in fewer iterations. Transfer learning can significantly improve accuracy but is no substitute for data. While dataset 1 was too small to provide sufficient information for

machine learning to provide a useful result it was capable of producing surprisingly accurate results rivalling the larger baseline dataset and warrants further examination of the effect of ultrasound sample size on neural network learning and generalisation in future works. This also suggests that the number of layers was less important than dataset size when performing abdominal ultrasound plane classification with the difference in accuracy of neural networks for dataset 2 being just 2.2%. Transfer learning also significantly improved neural network accuracy with the larger dataset, when comparing dataset 2 with the baseline, the per-plane training variance is noticeably reduced with the addition of transfer learning along with a significant improvement in accuracy. While dataset size was a more significant factor in reducing variance and increasing accuracy, transfer learning allows for significant improvements to ultrasound plane classification accuracy where the data is sufficient for the number of parameters in the neural network used.

3.5.1. Mobile Networks

Based on the previously discussed results of the previous nine networks, a follow-up study looked at neural networks designed to be run on mobile devices, typically these networks are smaller than 20Mb in size. This was designed to further test the potential efficiency and scalability of ultrasound image-based classification that was discussed earlier in this section. The use of small efficient networks designed for mobile deployment offers the opportunity to fully saturate the network perceptions ensuring the most effective possible accuracy result for that network.

The results of the mobile network training are compared to the top performing standard size neural network in Table 3.9. This not only shows that mobile networks such as MobileNet V2 and V3 were able to provide comparable accuracies to standard sized networks achieving 82.5% and 81.2% respectively, but EfficientNet exceeded the InceptionV3 model accuracy by 0.6%, achieving 84.5% accuracy despite being five times smaller at just 15.6Mb in size. While SqueezeNet only achieved an accuracy of 77.5%, this was done using a fraction of the system resources of the other networks tested, being both noticeably faster to run and with a network size of just 2.81Mb. This result is in line with previous expectations, suggesting that there is a limited amount of useful image information within ultrasound for use in cross section classification as such smaller more efficient networks can achieve comparable results to those of larger networks.

Table 3.9 – Accuracy of Individual Cross sections for mobile networks vs inceptionV3: Highest single neural network accuracy trained using Dataset 2

	InceptionV3	EfficientNet	mobilenetv2	MobileNetV3	SqueezeNet
Network size	83.5Mb	15.6Mb	8.81Mb	16.3Mb	2.81Mb
1. Epigastric sagittal: Liver/aorta	100%	85.94%	84.38%	85.94%	87.50%
2. Epigastric horizontal: Hepatic vein	78.57%	92.19%	87.50%	87.50%	79.69%
3. Right Epigastric oblique: Horizontal portal vein	78.57%	84.38%	78.12%	81.25%	75.00%
4. Right Subcostal: Gallbladder	78.57%	82.81%	82.81%	75.00%	78.12%
5. Right hypochondrium vertical: Gallbladder	78.57%	85.94%	92.19%	85.94%	79.69%
6. Right hypochondrium vertical: Bile duct	71.43%	46.88%	51.56%	45.31%	42.19%
7. Right subcostal: Liver	85.71%	92.19%	90.62%	93.75%	89.06%
8. Right intercostal: Liver	85.71%	84.38%	82.81%	78.12%	71.88%
9. Right intercostal: Liver	85.71%	81.25%	76.56%	82.81%	76.56%
10. Right intercostal: Liver	92.86%	92.19%	85.94%	89.06%	85.94%
11. Right intercostal: Right kidney	78.57%	71.88%	68.75%	60.94%	59.38%
12. Epigastric vertical: Bile duct/pancreas	78.57%	79.69%	81.25%	75.00%	75.00%
13. Epigastric horizontal: Pancreas	78.57%	92.19%	87.50%	84.38%	84.38%
14. Epigastric oblique: Pancreas	85.71%	90.62%	82.81%	89.06%	82.81%
15. Left intercostal: Spleen	100%	96.88%	96.88%	95.31%	89.06%
16. Left intercostal: Left kidney	85.71%	92.19%	90.62%	89.06%	84.38%
Average Accuracy	**83.9%**	**84.5%**	**82.5%**	**81.2%**	**77.5%**

3.6. Conclusion

This Chapter builds upon current knowledge by evaluating the classification accuracy of three major neural network architectures using 16 abdominal ultrasound cross sections, providing a baseline for future study. Transfer learning using linear, residual and inception neural network architectures were all shown to be effective in classifying abdominal cross sections with the number of layers in the neural network being a less significant factor than the size of the datasets. Transfer learning was capable of significantly augmenting dataset size compared to training using the data alone. The inception and residual architectures were more effective with larger datasets, while classic linear neural architectures remain useful for smaller dataset where the limited number of parameters was more effective than the deeper neural networks tested. As neural network architectures further develop for image classification techniques it is important to continue to test their effectiveness on medical imaging such as ultrasound which provides more constrained visualisation data than that of traditional imagery.

Neural network selection proved to be a less crucial factor when compared to providing enough data for training with results suggesting that increasing dataset size would likely further reduce the variance between neural network accuracy results. While the inception architecture produced the highest accuracy when provided with a sufficiently large dataset, the difference in accuracy between neural networks was fairly small. In the case of a small dataset transfer learning and a small linear neural network of just 8 layers was able to attain the most accurate result. Transfer learning significantly improves accuracy and reduces training variance, even where dataset size is small as is often the

case with medical datasets in comparison to traditional supervised learning. While the limitations of the validation set should be noted, the results are encouraging in that a sufficiently accurate classification result can be achieved for multiple abdominal cross sections even where overlapping anatomical structures are present.

Further study of mobile networks confirms that high accuracy results can be achieved for ultrasound classification using lightweight networks, with mobile and small network classifying comparably to that of substantially larger networks. This is likely due to the limited visual information available within ultrasound data, not fully utilising the parameters of larger neural networks although this case study suffers from the same limitations and the original study, with more data required for further training and testing to ensure that the networks are fully utilised.

The study of neural networks for abdominal plane classification has so far been limited, this chapter provides much needed context for further research in the area suggesting the potential of gaining high accuracy results while utilising neural networks with fewer parameters by using transfer learning as a baseline for further training. The significant accuracy gained from transfer learning on the small sized dataset suggests further research into what size of dataset is required to achieve a meaningful high accuracy classification result for ultrasound data.

Chapter 4

A Cost Focused Framework for Optimising Collection and Annotation of Ultrasound Datasets

Abstract

The process of collecting primary medical ultrasound images, presents a notable hurdle in the form of the high costs associated with clinical data generation and annotation. The challenge of balancing costs against dataset size is a concept well-recognised within the realm of clinical trials. Consequently, the strategies employed in this domain can be adapted to streamline data collection and annotation procedures, thereby mitigating expenses and timelines in the context of machine learning-driven feasibility studies.

This chapter introduces a biphasic framework designed to evaluate the cost of data collection via iterative predictions of accuracy in relation to sample size. The framework also incorporates active learning techniques to guide and optimise comprehensive human annotation specifically for machine learning applications within the domain of medical ultrasound imaging. The chapter showcases the potential reduction in costs against publicly available breast, foetal, and lung ultrasound datasets, as well as presenting a practical case study centred around the breast ultrasound dataset.

The findings underline the ability to predict the correlation between dataset size and subsequent accuracy, echoing a pattern akin to that seen in clinical trials. Substantial enhancements in accuracy are observed with the utilisation of just 40-50% of the data, contingent on the applied tolerance metric. The employment of active learning further reduces the necessity for manual annotation, resulting in a marked cost reduction of approximately 66%, while maintaining a permissible accuracy deviation of around 4% of theoretical maxima.

The significance of this work lies in its ability to predict how much additional data and annotation will be required to appropriately train a neural network to the accuracy level required. Using methodologies such as power theory to scale trials is already well understood by clinical funders and so provide a valuable and effective framework for feasibility and pilot studies. This framework can therefore be applied to machine learning studies to maximise predictive gains while adhering to a fixed budget [356, 357].

4.1. Introduction

While this EngD was initially planned to be an image and data analysis-based machine learning project, the industrial partner was unable to provide any additional data, nor had the bandwidth to provide assistance in annotation. It was therefore necessary to plan and implement a collection protocol that could take into account the limited resources available. The lack of data is a common problem in medical data science so finding an efficient, cost-effective solution for collection and annotation has merit in itself, especially in proof of concept and pilot studies such as those described in later chapters.

4.1.1. Motivation for Cost Analysis and Optimisation

Ultrasound is one of the most commonly used diagnostic modalities in the world today due to its low cost and minimally invasive approach [358]. Despite this, there are very few large-scale public ultrasound datasets available [321] and where clinical data does exist there is often no useful annotation to produce an effective ground truth. This is not a problem unique to ultrasound. The inherent cost of producing high quality data and subsequent complex clinical annotation required to inform the neural network means that generating appropriate datasets for diagnostic quality deep learning is a major investment [359, 360]. Therefore, when designing or commissioning a research project applying machine learning to ultrasound, it is important to factor in the financial and clinical cost of producing and annotating the data as well as the machine learning itself. There are many methods aimed at optimising neural network response to training, such as transfer learning [361], as well as methods for reducing the burden of annotation by

reducing human-model supervision [362] and self-supervision [363, 364] such as masked autoencoders [365] and few-shot learning [366, 367], as well as methods for reducing the burden of annotation by reducing human-model supervision [362] and self-supervision [363, 364] such as masked autoencoders [365]. These methods while effective are not designed to consider the real-world barriers to machine learning research, factors such as the cost and time of data collection that could completely prevent a study from being performed. Fortunately, there is already a tried and tested methodologies within medical research for performing this type of analysis that is well known to clinical funding bodies and commissioners: those used for designing clinical and random control trials, where is often not clinically or financially viable to sample a large population, therefore a smaller feasibility study is first performed, and the results analysed to calculate the size of subsequent trials [368].

The process of developing a dataset for machine learning has many similarities to designing random control trials, the addition of more data has diminishing returns on how much it improves the accuracy of the results, some trade-offs can be made to maintain the validity of the study while making it cost effective [369]. Where funding and clinical resources are finite, it is important to weigh the value of additional data and annotation against the time and cost of producing it. This chapter applies a framework similar to that of clinical trials [370], such as using a statistical power curve function during sampling to quantify diminishing returns in training result. Cost will then be further optimised by applying active learning to reduce the cost of data annotation after collection by targeting manual annotation to those parts of the data that most require

additional attention by an expert clinician. Time and cost are critical metrics to decision makers attempting to balance the risks and priorities of medical device and imaging research involving machine learning but has seen limited focus in the current literature.

4.1.2. Cost / Time Optimisation Methods and Applications

This section introduces the use case for power curve theory in determining dataset size before comparing the proposed method to common methods from the literature. The challenge of labelling and annotation is then discussed, examining the potential for active learning algorithm using an uncertainty sampling methodology to guide manual labelling.

Statistical power analysis is the concept of estimating the effect of a result within a given sample and to what extent this can be generalised to a larger sample size based upon its statistical significance using a fitted curve. This power curve represents every arrangement of power and difference for each sample size when the significance level and the standard deviation are held constant. Factors that may affect statistical power are as follows: the statistical significance criterion used in the test, the magnitude of the effect of interest in the population and the sample size used to detect the effect [371]. Statistical power analysis has been previously used to predict classification performance [372], to predict dataset sample size with retinal optical coherence tomography (OCT) [373] and in magnetic resonance imaging (MRI) [374] but has yet to be explored for ultrasound.

When proposing a machine learning study, cost can be a limiting factor, especially in medical imaging where sample size can play a major role in study cost. One common method of sample selection is to use a derivation of the Widrow-Hoff learning rule [375] that suggests the use of a number (such as 10, 100 or 1000) of sets of data for every imaging feature that will be used in the model. This method is somewhat arbitrary and may come up with sample sizes that are too small or large for actual training purposes depending on the feature-set being examined with limited possibility for cost to performance comparison. Model-based sampling based on the algorithmic characteristics such as generalisation [376, 377], or convergence [378] can provide good baseline for sample size selection based on threshold criteria but can be more difficult to directly relate to costs. The proposed method uses empirical curve fitting for sample size determination [379], allowing for accurate prediction of time and cost of producing the data similar to that used for control trials [380, 381].

Where ultrasound data is available without annotation, there is an opportunity to apply a targeted approach to sample labelling. There are many ways to reduce the cost of annotation in the early stages of data analysis such as using unsupervised clustering methods [382], in this case active learning was used to target manual clinical annotation time more effectively. While more expensive than self-supervised and automated methods, full human annotation is already recognised as appropriate by regulatory bodies for medical device research and as such was used as the benchmark in this study.

Active learning is a subfield of machine learning aimed at minimising cost of obtaining labels for data by allowing the algorithm to directly query the labelling source in this

case the clinician performing the annotation [383, 384]. This feedback method allows for greater accuracy while using fewer labels. [385]. There are many common methods of active learning within machine learning, such as using an unlabelled pool where a network chooses the best examples of a classifier known as diversity sampling [386]. In this chapter selective uncertainty sampling [387] is used to identify where the neural network has the lowest confidence in its prediction and target those images for annotation. This sampling method has shown to be highly effective with classification problems [388] and has been used previously as a method of dataset selection criteria for ultrasound data [389]. This forms an active learning loop, allowing for the consistent querying of the learning network to better inform the annotation process (Figure 4.1). Active learning has already been successfully applied to breast ultrasound using a weakly supervised approach, as well as in the detection of breast masses [390, 391], in the multi-model detection of liver fibrosis for ultrasound elastography [392], and in semi-supervised covid lung disease classification [393].

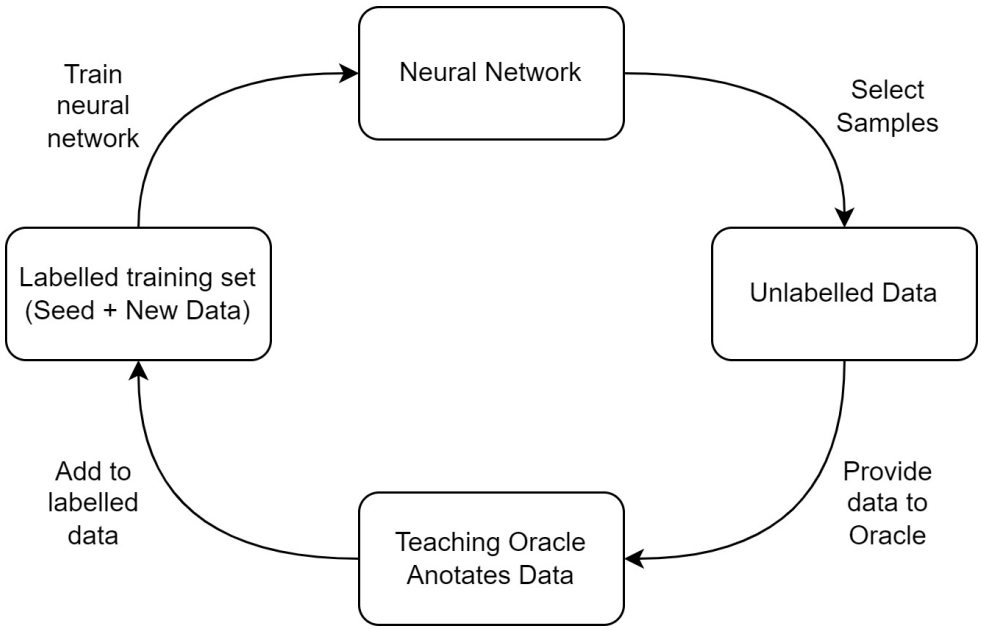


Figure 4.1 - Active Learning Cycle based on Settles [383]. This shows the cyclical iterative nature of active learning within machine learning.

4.1.3. Structure and Scope

This chapter proposes a biphasic prescriptive framework for optimising data capture and data annotation. The datasets, machine learning algorithms, and data control measures of each phase is shown in section 2. The efficacy of each phase is shown independently in sections 3.1 and 3.2. A case study is presented in section 3.3 using publicly available data, demonstrating the framework for reducing the cost of data capture and annotation compared to the common approach of using fully annotated arbitrary data sets. This chapter examines how:

- Ultrasound dataset size effects neural network accuracy performance for three publicly available datasets to determine sampling size effectiveness.

- Uses power curves to predict data performance from a small sample to inform further data collection for machine learning based on a cost benefit analysis.
- Compares that sample size prediction compared to the real result from the dataset.
- Tests the effectiveness of uncertainty sampled active learning for ultrasound data for reducing the cost of annotation.
- Combines these methods to determine a sample size and annotation level for maximising accuracy whilst minimising cost.

The use of curve fitting for determination of sample size is not as efficient as formulaic or model-based sample size selection methods but uses empirical testing to provide a simple robust basis for predictive modelling of dataset cost. While semi-supervised, fully automated, or clustering methods may provide less expensive labelling options than manual annotation, they are subject to a number of separate difficulties which may be expensive to overcome. Where manual annotation must be maintained as the primary form of annotation, such as cases where regulatory approval is a consideration, uncertainty sampled active learning allows for manual annotation to be targeted at those classifications with the weakest predictions while stronger classifiers could potentially be labelled using a semi-automated labelling process. When combined, these methods form a novel 2 phase process to reduce the cost of producing a dataset for machine learning for ultrasound by optimising collection and labelling of data.

4.2. Method

4.2.1. Proposed Method for Optimising Sampling/Annotation

Phase 1 uses power curves, a method common in determining size of clinical trials based on factors such as population size and available resources. Applying this technique allows these same factors to be considered during the collection of data for machine learning and also to determine a rough performance estimate from the size of the dataset. Ensuring a representative sampling within the training set assists in the subsequent extrapolation of the statistical power curve. This experiment simulates the data collection process, producing a teacher oracle on a small subset of the data, and then using the result to extrapolate the power curve. Each subsequent iteration adds data to the oracle training subset, representing an additional round of data collection. Phase 2 uses semi-supervised active learning to automatically annotate a proportion of the dataset, where a reduction in performance can be accepted as part of the experimental parameters, an error tolerance threshold can be applied, leading to significant cost and time savings. In this work, a convolutional neural network (CNN) is used as described in section 2.3, Alexnet is used a well-known and understood benchmark. Hyperparameters are also exemplar and not intended as a recommendation of optimal settings, but merely to demonstrate the framework in action.

4.2.1.1. Phase 1 – Optimised Data Set Capture:

Phase 1: Estimate the power curve and predict required dataset size (Figure 4.2). Below are explanatory notes for the flow chart:

1. A neural network is trained on a small sample of annotated data (dependent on experimental constraints e.g., 10-100 samples). The dataset should be split into training and test sets. Validation metrics (such as accuracy) are saved.
2. An additional subset of annotated samples (ideally in equal chunks) is added to the dataset (re-randomising training and validation sets is advised).
3. Neural network is retrained and tested. The chosen validation metrics are saved.
4. The validation metrics are plotted against dataset size and a power curve is fitted to the data.
5. Repeat steps 2-4 until curve fit is 'stable' at desired statistical power and accuracy. Stability is when subsequent sample groups predict end accuracies within your desired tolerance (such as within 2%).
6. Plots of the power curve (e.g., accuracy vs sample size) can then be used to determine the required dataset size for desired/acceptable validation metric.

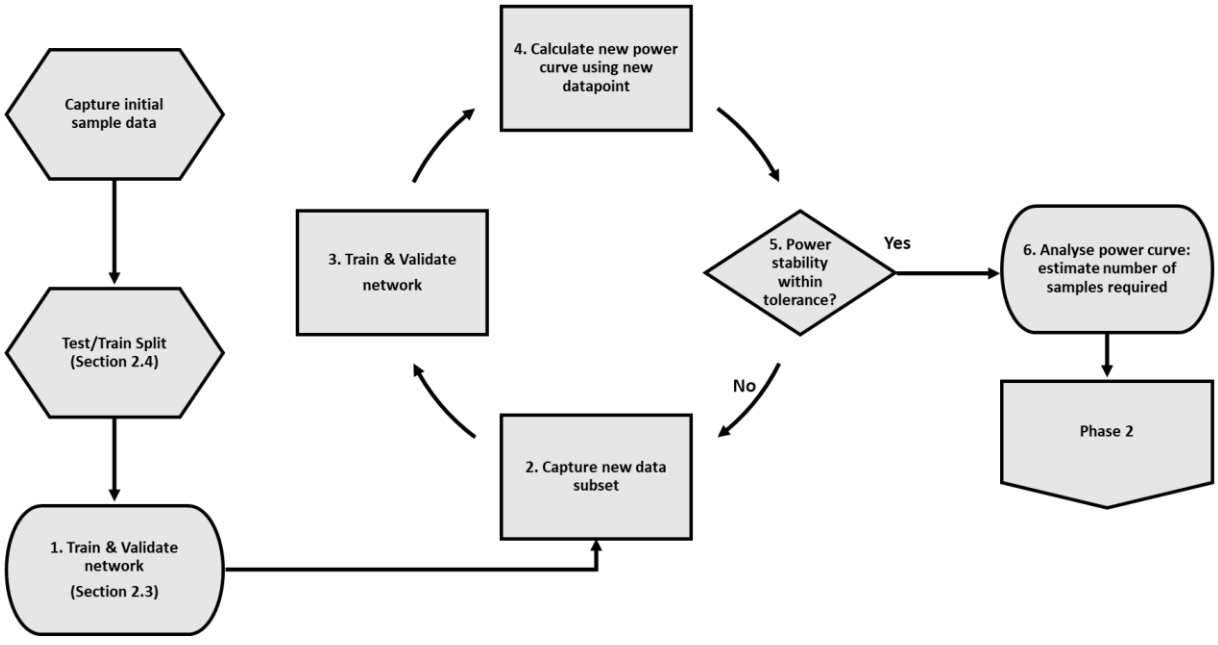


Figure 4.2 – Flow diagram of phase 1: Collection cycle and subsequent power curve analysis leading to the determination of dataset size based on curve fit.

4.2.1.2. Phase 2 – Optimising annotation:

In cases where excess samples have been captured, particularly in large unannotated datasets or where data is being repurposed, active learning, detailed in section 2.5, can be used to selectively target samples that the CNN has the most difficulty identifying for manual annotation, by selecting samples where the neural network has provided the lowest predictive accuracy as seen in Figure 4.3. This process uses selective uncertainty sampling to minimise manual annotation of remaining data. Below are explanatory notes for the flow chart:

1. Train and validate a neural network on available annotated data (such as the sample set produced in phase one).
2. Identify least certain samples on unannotated data, where the CNN has least certainty detecting particular classifiers (e.g., bottom 50 samples).
3. Manually annotate next batch of data with additional focus on identified weak classifiers.
4. Combine new and old batches and reshuffle the dataset.
5. Train new neural network and evaluate result using the validation set.
6. Use the validation metrics (such as accuracy level) to decide if additional samples are required. Consideration should be given to the effect of dataset balance on classification as well as comparisons made to predicted values from Phase 1 to ensure training validity.

7. Cease annotation when (cost and accuracy) metrics are within acceptable parameters.

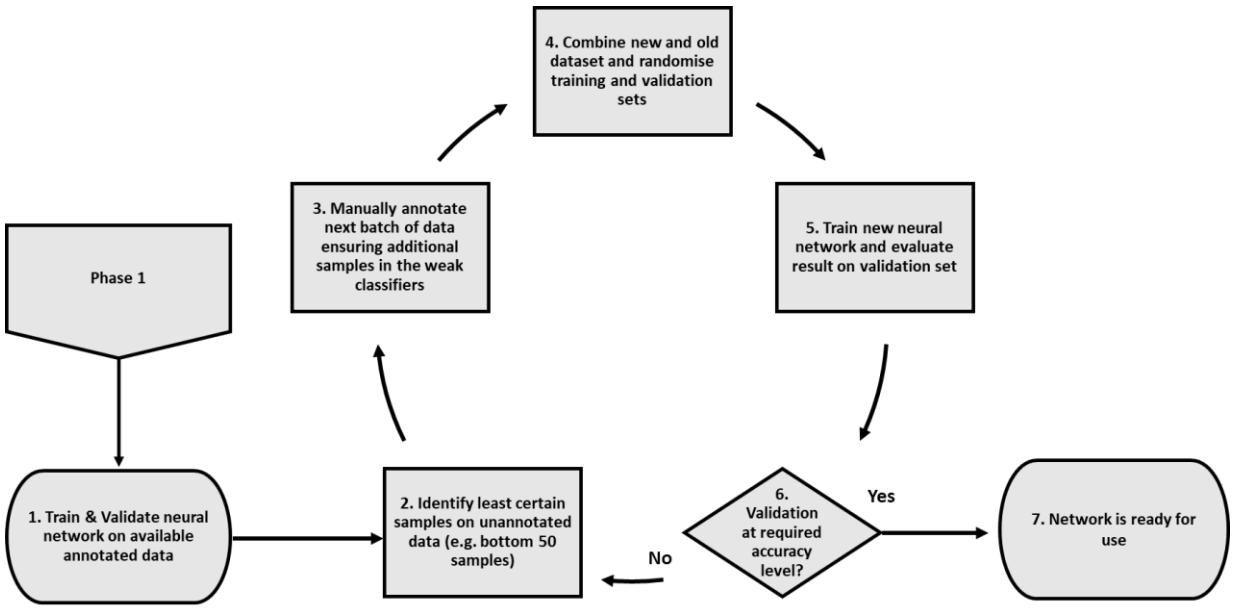


Figure 4.3 – Flow diagram of phase 2: Active learning cycle for annotation.

This iterative process allows this technique to be applied naturally during the data collection and annotation process, such as during a pilot study. A new round of training can be performed upon receipt of a new batch of data, adding an additional datapoint for the power curve. If data is in a single large batch it, like those introduced in subsection 2.2, can be divided into percentages such as in this study, to produce the required increments. This is shown in the case study, Section 3.3.

4.2.2. Datasets

4.2.2.1. Breast Lesion

The BUSI breast lesion ultrasound dataset [394] (Figure 4.4) consists of breast ultrasound images of 600 women between the ages of 25 and 75. The ground truth images were presented with original images. The images were categorised into three classifiers: normal, benign, and malignant as confirmed by biopsy in the original study. The original ground truth for this dataset contained image masks for segmentation, as this work focuses on classification, these segmentation masks were not used.

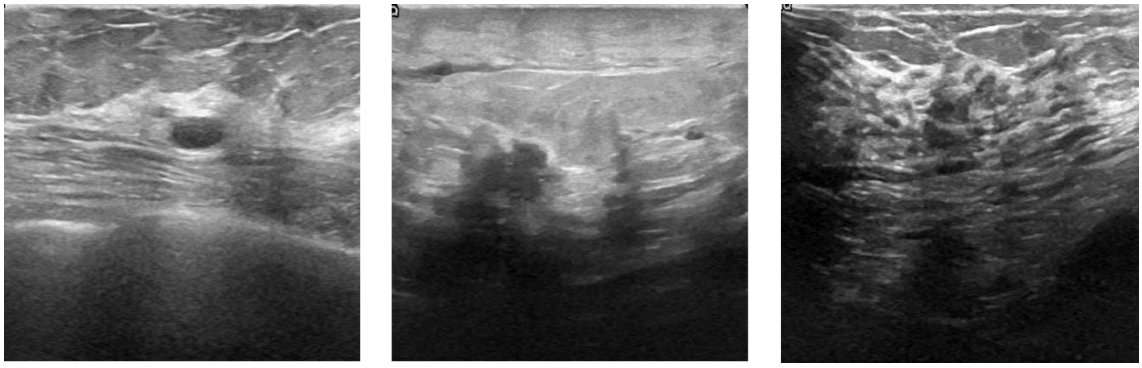


Figure 4.4 - Examples of breast lesion ultrasound classifiers from the BUSI dataset [24], left benign, centre: malignant, right: normal.

4.2.2.2. Covid Lung

The lung ultrasound dataset [395, 396] (Figure 4.5), consists of 179 videos (64 COVID, 49 bacterial pneumonia, 66 healthy), 53 images (18x COVID, 20x bacterial pneumonia, 15x healthy) from convex probes and 17 videos (6 COVID, 2 bacterial pneumonia, 9 healthy) and 6 images (4 COVID, 2 bacterial pneumonia) from linear probes. Cases of viral pneumonia in the dataset were excluded as it consisted of only 6 cases and there is evidence to suggest ultrasound can differentiate between viral and bacterial pneumonia [397, 398] meaning including it in a single pneumonia classifier would be counter intuitive.

Figure 4.5 - Examples of Covid Lung Ultrasound Dataset [25]. left: Covid, centre: bacterial pneumonia, right: normal.

4.2.2.3. Foetal Planes

The foetal ultrasound dataset [334] consists of around 12,000 images from 1792 patients and is split into 6 classifiers: foetal abdomen, brain, femur, thorax, maternal cervix, and a generic 'other' classifier as exampled in Figure 4.6.

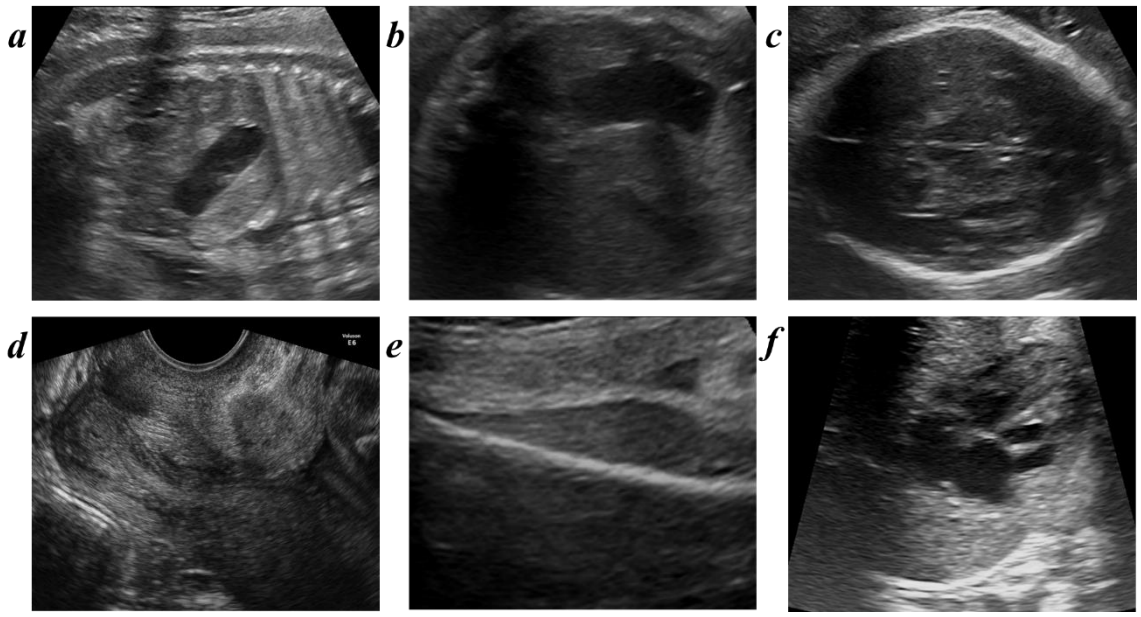


Figure 4.6 - Examples of foetal Plane Ultrasound Dataset [29]. a: other, b: abdomen, c: brain, d: maternal cervix, e: femur, f: thorax.

4.2.3. Deep Learning

The experimentation was performed using the PyTorch framework [319], on a computer with an Intel CPU with a clock speed of 2.4 Ghz and a Nvidia 3060 GPU. A standard Alexnet neural network that had been pretrained using the ImageNet Challenge dataset [339] was used with the final layer output reduced to fit the classification requirements of the dataset. Alexnet [215] was selected to provide a baseline to study dataset selection

size, using the training parameters in Table 4.1. Network selection and hyperparameter settings are simple and provides an example of the framework in action but not to be optimal for high precision machine learning tasks. Hyperparameter values were based on those found to produce stable results for the ultrasound experiments reported in chapter 3. The complexity of the data and requisite predictors and feature sets should be considered when designing machine learning studies using techniques as seen in these recent studies [399-402]. The images were contrast normalised and compressed into tensors sized 299×299. Test/train split was performed on a per image basis as there is no known repeated data within the set. No additional data augmentation was performed as this would potentially confound results.

Table 4.1- Sample Hyperparameters used in training of the example network.

Hyperparameters	Value
Activation Functions	SoftMax
Learning rate	0.001
Training iterations	80
Epochs	20
Optimiser	ADAM
Momentum	0.9
Dropout	0.5

4.2.4. Training set size

The dataset was split twice for each experimental run. The dataset was initially split using stratified random sampling [403] at a ratio of 80/20 at the subject level to create a holdout test set, this was performed to ensure a representative sampling of each classifier. The training set was then resampled with a percentage retaining their original labels for training the oracle networks. The percentage of the breast and foetal datasets were between 1% and 5% then in increments of 10% thereafter. The Lung dataset has a smaller sample size and so the smallest split tested was 5%, then in increments of 10%. This was done to simulate the collection of data over time. The percentage of data is increased with each iteration in order to determine the relationship between sample size and accuracy. The labels for the remaining data was removed and new labels produced by the teacher oracle network.

Each experimental training run was performed 80 times over 20 epoch each. Each epoch represents a single completed pass of training data through the algorithm. In order to account for variability in the training result, that occurs naturally when performing machine learning [404], 80 resampling runs for each dataset increment were performed in 4 sets of 20. This produced 1,120 networks trained on the breast and foetal datasets and 800 networks trained on the lung dataset. The use of an unseen test set means that any overfitting that has occurred will already be reflected in the test result that is used in the calculation of the power curve, with the accuracy of an overfitted network substantially lower when validated on the unseen test data. Each subsequent round of collection increases the size of both training and test sets and testing adds a new

datapoint to the power curve further refining the curve fit. The exampled approach is overly simplified, performing a stratified random split with each new experimental run and looking at only very simple classification tasks, more complex datasets should consider folded-based cross validation and the potential improvements that could be made through careful feature selection and ensemble learning models as methods to reduce overfitting as exampled in recent studies [405-407].

4.2.5. Active Learning

At its core, active learning is a technique whereby the learner plays a role in specifying the content they learn [385]. An uncertainty sampling [386, 389] method is used whereby the images with the lowest confidence was selected for annotation. This was performed for each percentage of the dataset from 10-90% with the active learning performed on the remaining percentage of the dataset also using a threshold percentage to specify an additional proportion of the dataset for annotation (as can be seen in Figure 4.7). Each active learning threshold was tested 20 times over 20 epochs.

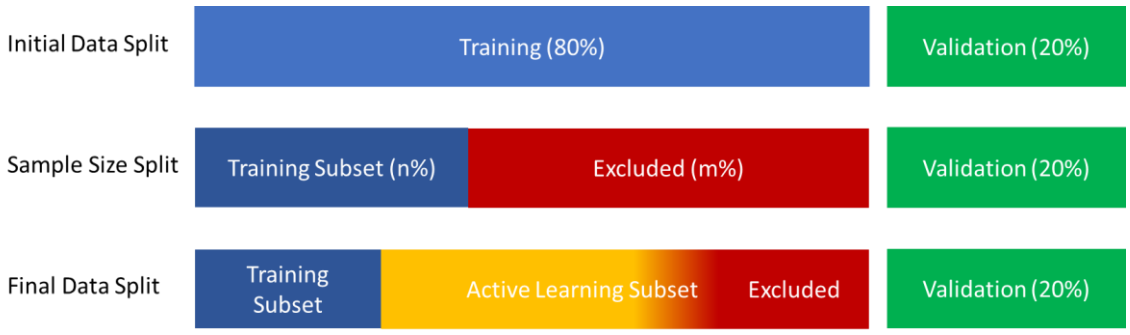


Figure 4.7- Diagram of active learning dataset split method showing proportion of data used for training and threshold for additional annotation.

4.3. Results.

4.3.1. Size to accuracy of dataset

Initial experimentation examined the effect of dataset size on accuracy, examining how power curves could be used to determine data requirements without applying active learning to the annotation. Figure 4.8, Figure 4.9 and Figure 4.10 provide both mean and highest accuracy results to facilitate comparison. The mean curve fit result from 80 networks controls for variation in training outcome while the curve fit from highest accuracy networks suggests what can be achieved with an optimal training path. Examining the breast dataset mean accuracy results for 80 neural networks per threshold percentage (Figure 4.8), the highest mean accuracy of 85.42% was achieved using 90% of the data, contrast to 79.6% using 40% of the data and 75.29% at 20%, a difference of 5.82% and 10.13% respectively. Increasing dataset size reduces the variation as seen in the standard deviation between neural networks with an average of 6.17 at 1% of the data, down to 2.42 at 90%. Selecting the neural network with the highest accuracy for each percentile shows that the highest accuracy network with 91.72% was produced with only 60% of the dataset, in comparison to 82.8% using 20% of the dataset a difference of just 8.92%, there were significant diminishing returns on data investment after 30-40% of the data is used.

Using the mean accuracy data, it is possible to extrapolate a close approximation of data to fit classification accuracy, a fitted curve from just 10% of the data can be used to approximate the amount of additional data required to reach a certain level of accuracy

similar to that used in clinical studies, with each additional data point improving the fit further.

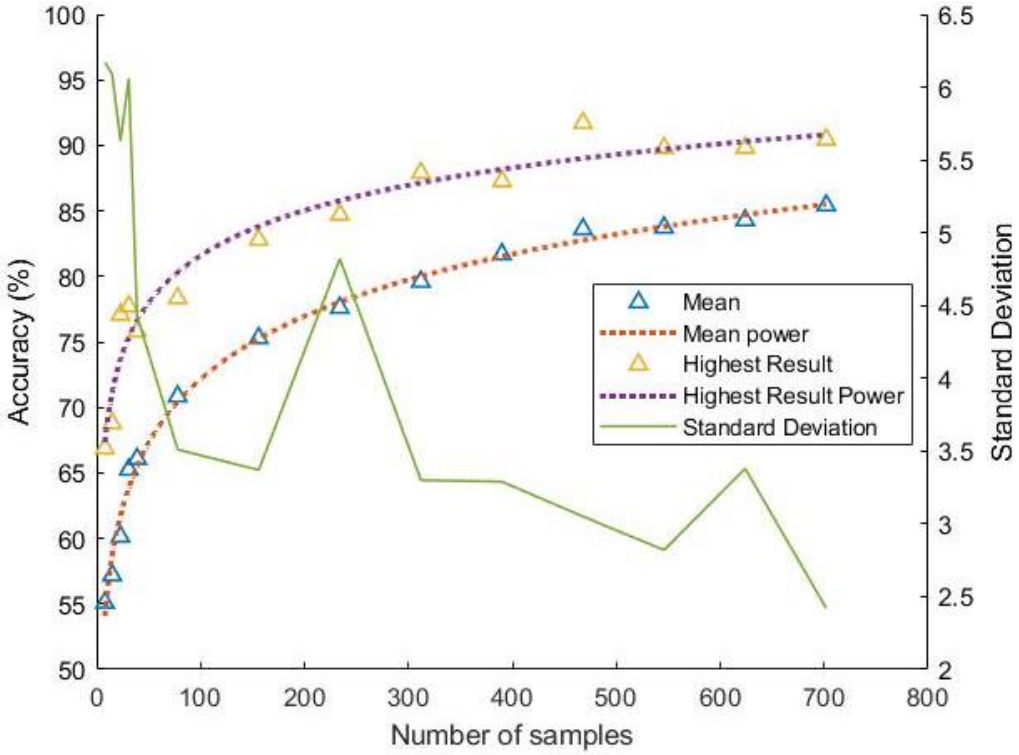


Figure 4.8 - Accuracy of mean and highest result with associated power curves for neural network response for breast dataset.

In the lung dataset (Figure 4.9) the difference between the highest mean accuracy of 83.66% and 80.87% using 40% of the dataset was just 2.79%. The trend of reducing standard deviation as dataset size increases is less obvious, while the initial deviation is 12.34 at only 5% of the dataset it is reduced to 7.62 by around 10% of the dataset but remains unstable but does achieve the lowest standard deviation at 5.25 at around 90% of the dataset. When the highest accuracy neural networks are considered, an accuracy of 89.93% is achieved at just 30% of the data, with diminishing returns until 70-80% where a significant improvement is achieved with results of 94.36% at 70% and 95.96%

at 80% of the data. As previously seen in Figure 4.8 the same data trend is possible and is visible in the mean accuracy data for the lung dataset. A statistical curve is used to predict CNN accuracy for sample sizes.

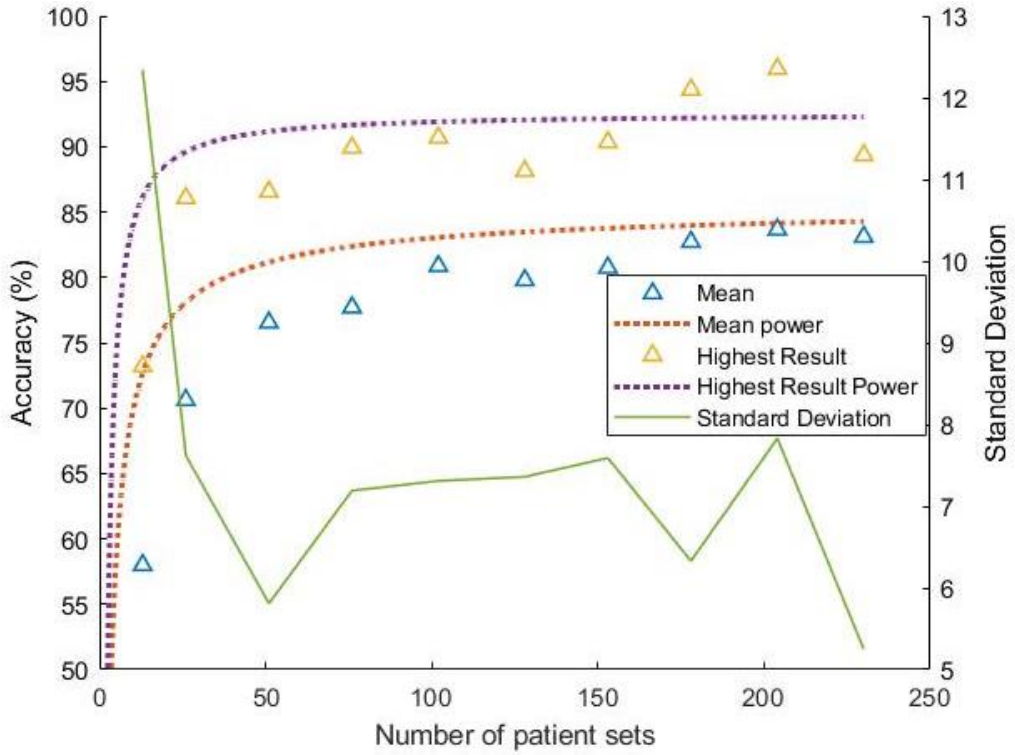


Figure 4.9 - Accuracy of mean and highest result with associated power curves for neural network response for lung dataset.

The foetal plane dataset contains over 12,000 samples from over 1700 patients making it the largest dataset assessed, using only 20% (around 2400 samples) the mean accuracy reached over 90%, with additional data providing diminishing returns for the additional data added. The standard deviation trends downwards from 3.40 to 0.54 using 80% of the dataset. The foetal data also exhibits the same trend from the power curve (Figure 4.10) despite containing substantially more data and classifiers than the previous two datasets, the difference between mean and highest result after 50% of the dataset (6000

samples) is 0.66% of that achieved with 90% of the dataset, it is also within 0.56% of the highest achieved result of 94.97%.

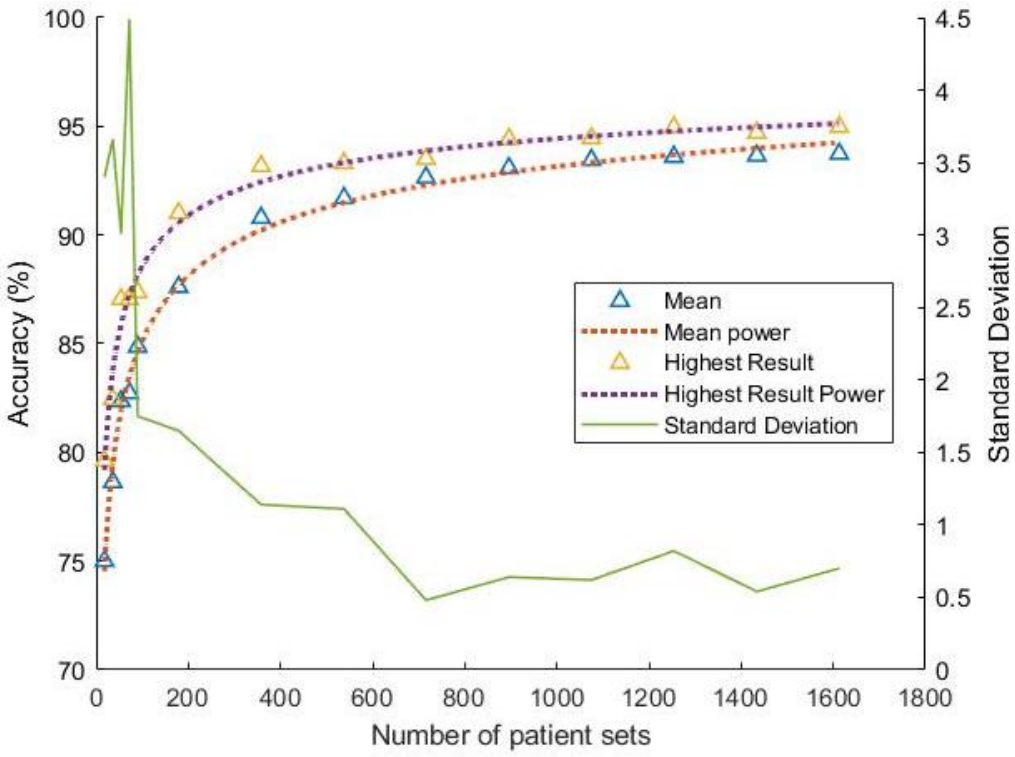


Figure 4.10 - Accuracy of mean and highest result with associated power curves for neural network response for foetal plane.

The use of curve fitting as a method of sample selection is empirical and has clearly shown to be effective at determining sample size in all three datasets, a clear followable trend that can be seen in the network response and can be used to extrapolate data requirements based off this trend. The high initial standard deviation in accuracy result in training seen in all three datasets is due to factors such as overfitting, sample randomisation and training performance. As sample size increases the so does the stability of the training process, due the test set being unseen the results from the networks form a clear accuracy trend regardless of these factors. Where sample size will

be consistently small harmonic mean (F-1 Score in Table 4.2) should be factored into result metrics to ensure network response is truly representative of learnt classification. The F1 score is the harmonic mean of the accuracy and recall, in cases where there are imbalanced classes, using these two values as part of the evaluation metric provides a more accurate predictor of performance than accuracy alone.

4.3.2. Active Learning

Comparing the results of using active learning to target the lowest predicted accuracy using a threshold to that of annotating the same percentage with no targeting shows a small consistent improvement. The highest accuracy of 92.99% is achieved at 60% of the dataset, as can be seen in Table 4.2, the neural network is already performing consistently with a weighted average precision of 90%, recall of 92% leading to an F-1 Score, the combination of precision and recall of 0.91.

Table 4.2 - Precision, Recall and F-1 score for top performing network (BUSI (breast) dataset based off a network trained on 10% of the dataset with an additional 40% annotated using active learning.

Classifier	Precision	Recall	F1-Score
0 – Benign	90	88	0.89
1 – Malignant	94	92	0.93
2 – Normal	87	96	0.91
Average	**90**	**92**	**0.91**

Comparing the default annotation values to those using active learning shown in Table 4.3 and Figure 4.9, shows that the majority of learning can be achieved using between

40-50% of the data (in the region of 300-400 sample sets). For example, when 10% of the data is used for the teacher oracle network and an additional 30% is annotated using active learning then a mean result of 82.99 was achieved that is only 4.3% less than when trained with the complete dataset where a mean result of 87.29% was achieved. Where 20% of the dataset is used to train the teacher oracle network then this difference drops to just 3.28%. The variation of accuracy after 60% of the dataset is likely due to the probabilistic nature of neural network training rather than the dataset itself. The statistical maximum result of 92.99% was achieved at all subsequent dataset proportions above 50% when trained exhaustively, but this may not be feasible to achieve in practice. This supports the hypothesis that additional data provides limited, to no, return on investment after this point.

Table 4.3 – Mean comparative accuracy of neural networks trained using Active Learning to label a percentage of the BUSI (breast)dataset.

		Percentage of dataset used for training									
Percentage of training data actively learned	**-**	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
	0%	**70.84**	**75.29**	**77.60**	**79.60**	**81.69**	**83.61**	**83.73**	**84.29**	**85.42**	**85.42**
	10%		**77.39**	**80.51**	**82.99**	**85.25**	**86.56**	**86.62**	**85.67**	**85.86**	**86.85**
	20%			**80.10**	**84.01**	**84.78**	**85.73**	**85.57**	**85.35**	**86.69**	**87.20**
	30%				**83.03**	**84.94**	**86.59**	**85.70**	**86.08**	**87.17**	**86.78**
	40%					**84.32**	**85.49**	**85.46**	**85.56**	**85.73**	**85.99**
	50%						**84.28**	**85.16**	**86.56**	**86.37**	**86.07**
	60%							**85.29**	**85.64**	**86.88**	**87.29**
	70%								**85.42**	**85.42**	**85.42**
	80%									**85.92**	**86.14**
	90%										**86.02**

When comparing the data for the default annotation technique with active learning for the breast dataset (Figure 4.11), the mean active learning consistently is above that of the default annotation technique, although this improvement suffers from diminishing returns after 50% of the dataset is used. While there is significant improvement in the highest accuracy results achieved even with only 10% active learning there is significant training variance at lower dataset sizes that would need to be accounted for in the training methodology.

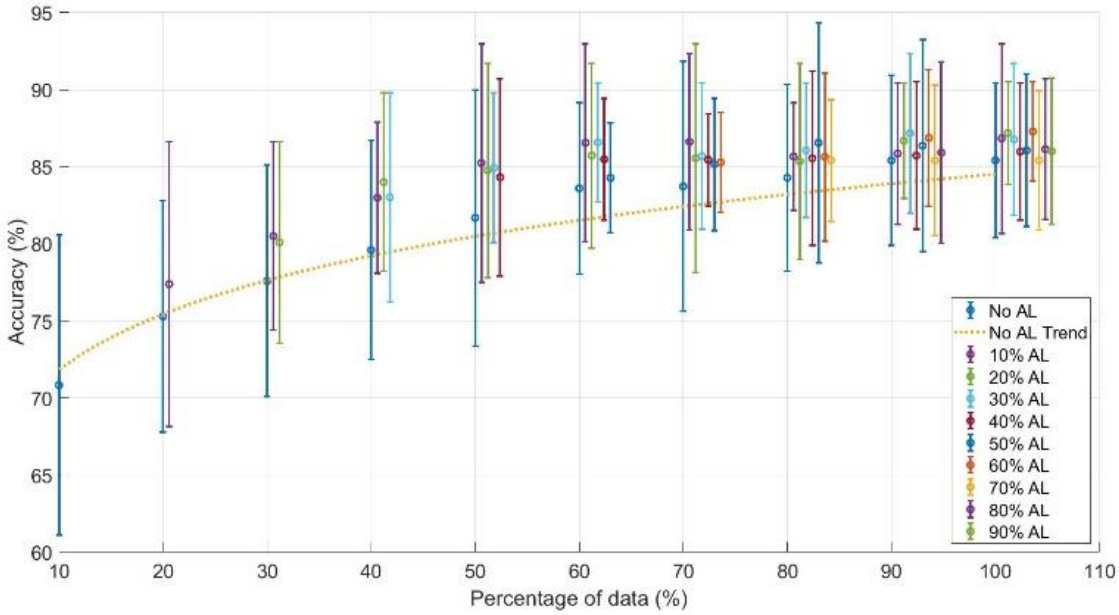


Figure 4.11 - Comparison of mean Active Learning (AL) to default annotation for breast dataset. Error bars denote variation in accuracy result between the 20 networks trained at that data percentage and level of active learning.

Lung ultrasound also performed well with active learning (Table 4.4), with a network trained on 20% and an additional 10% active learning was able to achieve a mean

accuracy of 82.35%, just 4.3% less than the highest achieved mean accuracy of 86.65% from a network trained on 30% of the data and an additional 50% targeted through active learning.

Table 4.4 - Mean comparative accuracy of neural networks trained using Active Learning to label a percentage of the lung dataset.

		Percentage of dataset used for training									
Percentage of training data actively learned	**-**	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
	0%	**70.62**	**76.55**	**77.71**	**80.87**	**79.79**	**80.75**	**82.72**	**83.66**	**83.11**	**84.43**
	10%		**78.41**	**78.66**	**81.09**	**80.86**	**84.33**	**84.24**	**83.37**	**84.59**	**84.59**
	20%			**82.35**	**83.77**	**80.09**	**83.22**	**85.69**	**83.77**	**85.09**	**84.43**
	30%				**81.46**	**81.91**	**83.51**	**82.97**	**86.65**	**85.23**	**84.81**
	40%					**80.86**	**83.33**	**83.90**	**84.36**	**84.51**	**84.57**
	50%						**84.33**	**84.81**	**85.23**	**83.11**	**84.59**
	60%							**84.24**	**84.36**	**84.81**	**85.23**
	70%								**83.50**	**85.23**	**84.43**
	80%									**84.84**	**84.81**
	90%										**84.59**

When comparing active learning to default annotation methods (Figure 4.12), the active learning does improve mean accuracy results but does not significantly improve training of high accuracy models after 60% of the data is in use due to the wide variation in training accuracy achieved.

The variation in training accuracy is highest for this dataset, out of the three, which is attributed to the low base dataset size, meaning that the CNN training is more susceptible to statistical anomalies in the data, a well-known phenomenon. Despite this,

the trend improvement is still evident, although with smaller returns initially than larger datasets.

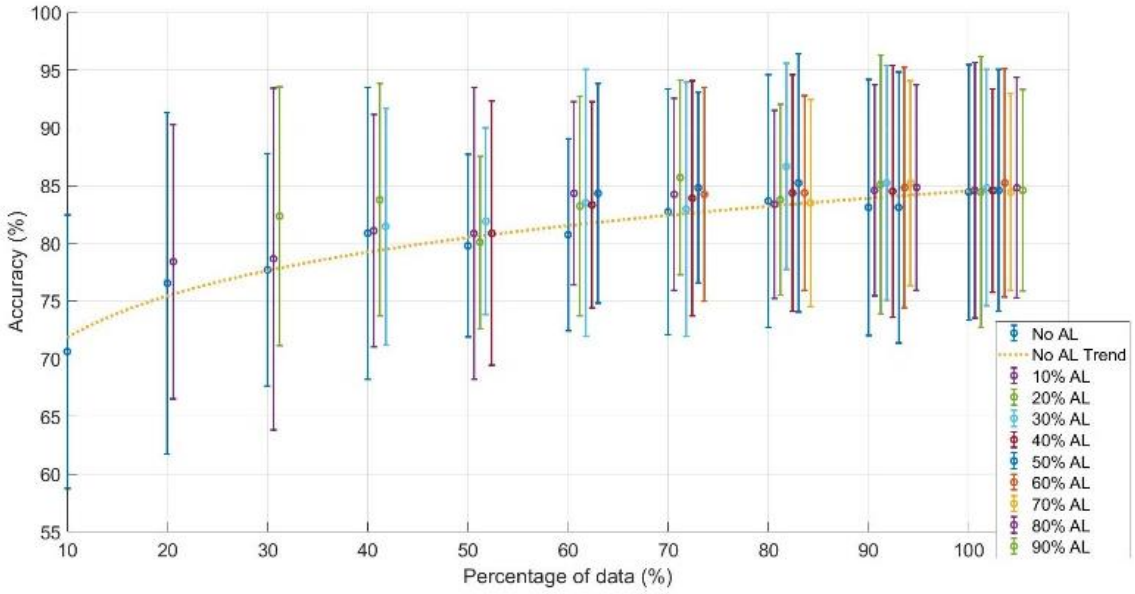


Figure 4.12 - Comparison of mean active learning to default annotation for lung dataset. Error bars denote variation in accuracy result between the 20 networks trained at that data percentage and level of active learning.

The foetal ultrasound data has a significantly larger sample size and double the number of classifiers than the previous two datasets. As seen in Table 4.5, there was a 2.22% improvement when active learning was used to annotate 10% of the dataset but is subject to diminishing returns as the highest accuracy result achieved was 94.40% using 80% of the dataset where the active learning had been trained using a dataset with 60% of the data an improvement of only 1.39%.

Table 4.5 - Mean comparative accuracy of neural networks trained using Active Learning to label a percentage of the foetal ultrasound dataset.

	Percentage of dataset used for training										
Percentage of training data actively learned	**-**	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
	0%	**87.61**	**90.79**	**91.70**	**92.65**	**93.08**	**93.45**	**93.60**	**93.65**	**93.74**	**94.13**
	10%		**93.01**	**93.83**	**93.95**	**93.89**	**93.89**	**94.19**	**93.78**	**94.15**	**93.92**
	20%			**93.85**	**94.10**	**93.92**	**93.78**	**94.02**	**93.97**	**93.98**	**94.09**
	30%				**94.15**	**93.91**	**93.83**	**94.03**	**94.00**	**93.84**	**94.06**
	40%					**94.01**	**93.80**	**93.77**	**94.13**	**94.18**	**94.08**
	50%						**93.89**	**93.92**	**93.94**	**94.00**	**93.98**
	60%							**93.31**	**94.40**	**93.68**	**94.34**
	70%								**93.92**	**93.91**	**93.95**
	80%									**93.75**	**94.23**
	90%										**94.34**

When comparing active learning to default annotation methods in Figure 4.13, an initial accuracy boost, after using above 60% data, training variance becomes a significant factor with active learning achieving only limited improvements.

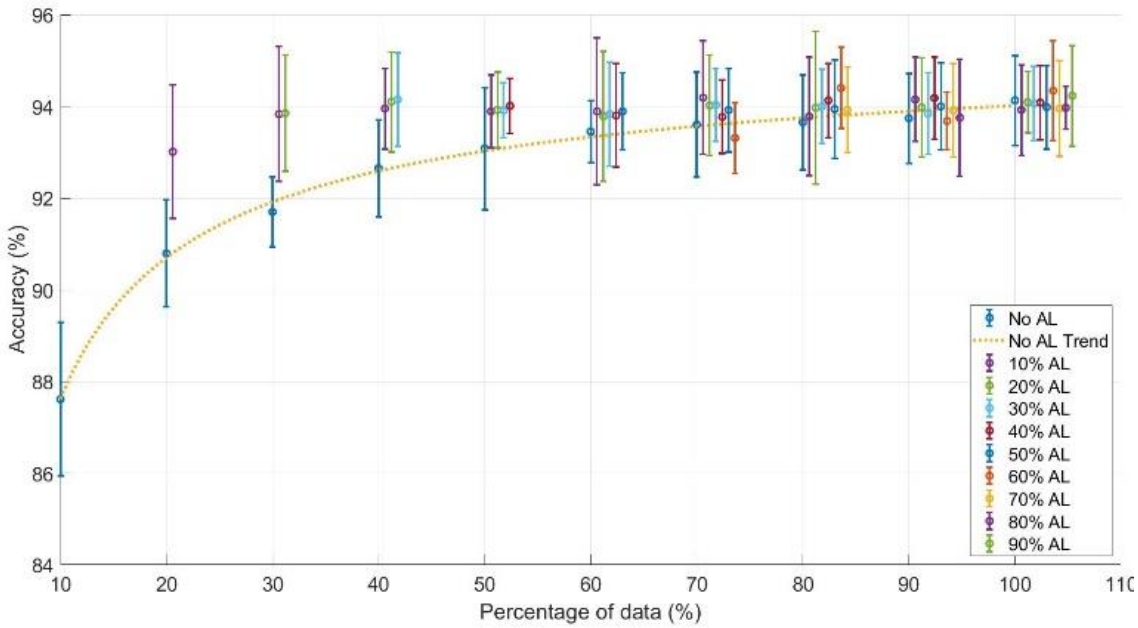


Figure 4.13 - Comparison of mean active learning to default annotation for lung dataset foetal dataset. Error bars denote variation in accuracy result between the 20 networks trained at that data percentage and level of active learning.

The use of uncertainty sampled active learning is shown to boost classification accuracy performance of all three datasets with most improvement seen prior to 50% human annotation of the dataset with substantial diminishing returns after this point. Where more than 50% of the data has been manually annotated, there is little to no performance drop, as seen in Table 4.3, Table 4.4 and Table 4.5, suggesting this data could be annotated using active learning with almost no loss of reliability despite the potential of labelling error.

4.3.3. Case Study

Having shown in section 3.1 that accuracy vs. sample size follows a power law and therefore has significantly diminishing returns after a certain point. As shown in section 3.2, active learning produces improvements in accuracy with low amounts of initially annotated data, again with diminishing returns as annotation proportion increases, we can now consider what this means in terms of costs for collection and annotation.

In order to demonstrate the potential saving incrementally, phase 1 and phase 2 of the prescribed method were applied independently to the BUSI data set, and then as a combined method considering mean response and max response of the CNNs respectively.

Phase one was applied to the BUSI breast dataset, with an initial sample size of 15. The process was iterated until power curve stability was achieved at 150 samples as shown in Figure 4.14. This allowed a prediction that 400 samples were required to be within 4% of the theoretical maximum accuracy achievable with the full dataset. These remaining samples (250) were then ‘collected’ by randomly sampling the BUSI dataset. All remaining BUSI data was used for validation.

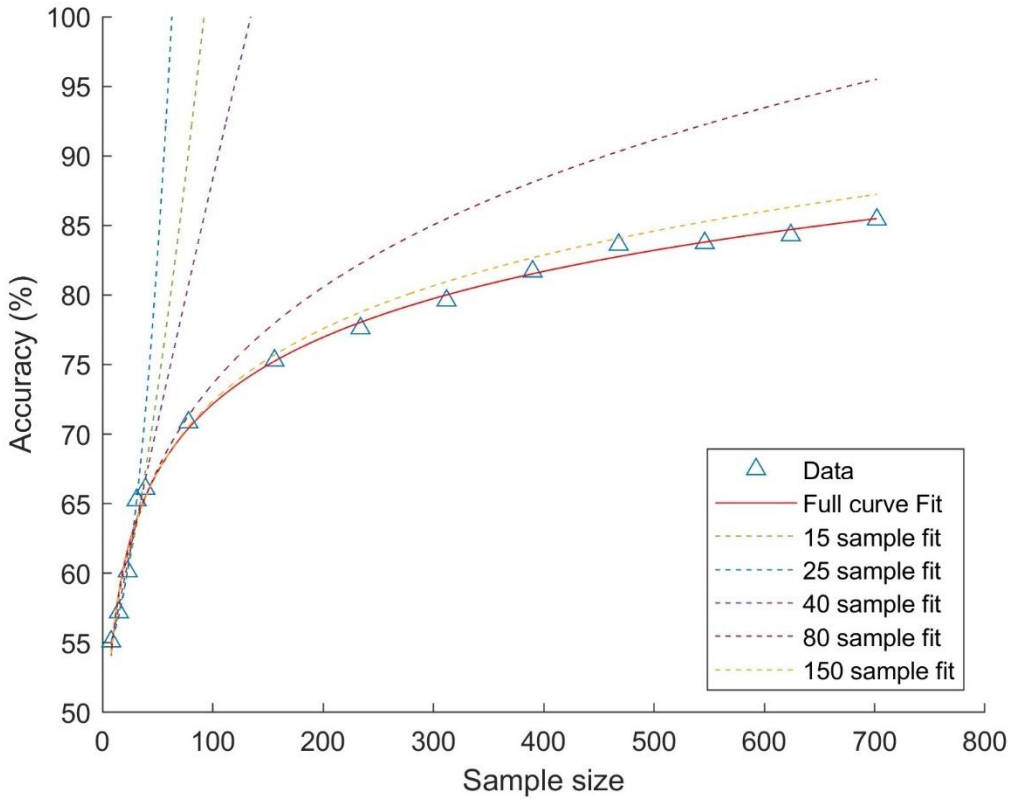


Figure 4.14 – Comparison of power curve fit for networks trained with 15-150 samples and those trained on the full breast dataset. Each additional set of samples added to the dataset improves the fit.

Phase 2 was then applied with an initial CNN trained on a sub sample of 50, and then predicted the annotation for the remaining 350 unannotated datasets with 50 chosen for annotation using uncertainty sampling and added to the training set. A new CNN was then trained, validated, and then used to select an additional 50 samples from the remaining unannotated patient sets. This was repeated until all 400 samples were selected as shown in Figure 4.15 for illustrative purposes, but the process would stop once the acceptable tolerance is reached. All networks were trained for 100 times for a cycle of 20 epoch, using Alexnet and ADAM optimisation method. Depending on the

experimental robustness requirements, the best result of the training epochs or the mean result can be considered with differing conclusions.

For the BUSI dataset with 400 samples from Phase 1, with an acceptable tolerance of 2%, 350 samples of the 400 need annotation for the mean response to be within tolerance but only 200 samples of the 400 need be annotated for the maximum result to be within acceptable tolerance.

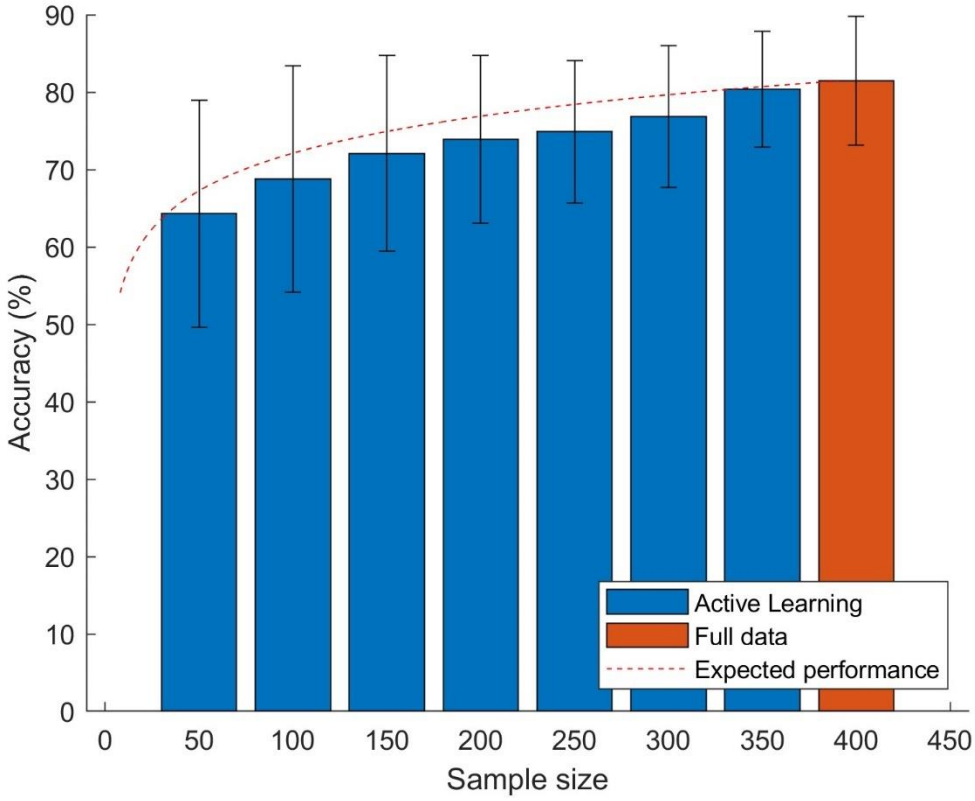


Figure 4.15 - Comparison of the mean accuracy of neural networks trained on a dataset produced by active learning compared to training on a fully human annotated dataset. Each iteration adds 50 samples to the oracles training set.

The combined method of phase 1 and phase 2, considering the maximum response from the CNNs, an accuracy of 85% was achievable using only 400 samples compared with

the theoretical maximum of 88% at the full BUSI dataset size (from Figure 4.8). Additionally with only 200 of the 400 samples manually annotated, accuracy only drops to 84.7% for a 50% reduction in annotation burden, directly translatable into costs.

For completeness, the cases of simply performing phase 1 alone (with 400 captured and annotated samples) and performing phase 2 alone on the full BUSI dataset, yielding 50% annotation, were also considered to illustrated cost differences. Using an initial representative costing model of 1:2 for data collection and annotation the relative costs of each method and phase can be seen in Figure 4.14 and Figure 4.15, calculated using Equation (1), where P is the price of collection or annotation, and N is the numbers predicted by phase 1 and 2 respectively.

$$Cost = (P_{Collect} \times N_{Collect}) + (P_{Annotate} \times N_{Annotate}) \quad (1)$$

Dependent on accuracy and robustness requirements, significant cost savings can be made by optimising collection using a statistical power curve, and by targeting annotation by applying active learning as described in our method. Combining the methods shows the potential to reduce costs even further, when a 1:2 unit cost for collection and annotation is applied, where the best performing network is taken into account, cost savings of up to 66% are possible as shown in Figure 4.16. A similar analysis has been performed for differing overall acceptable tolerances from the maximum prediction from Figure 4.8. This allows for further optimisation of costs when accuracy can be acceptably traded. The shape of this graph shows that regardless of the

initial costing model used, the prescribed method will always yield a cost reduction in comparison to capturing arbitrary amounts of data and annotating it all, which is an important result allowing decision makers to optimise their clinical applications of machine learning. The scale of the cost saving is related to the complexity of the data, the CNN type used, and the costing model, but this method is always expected to return a cost reduction for minimal accuracy loss.

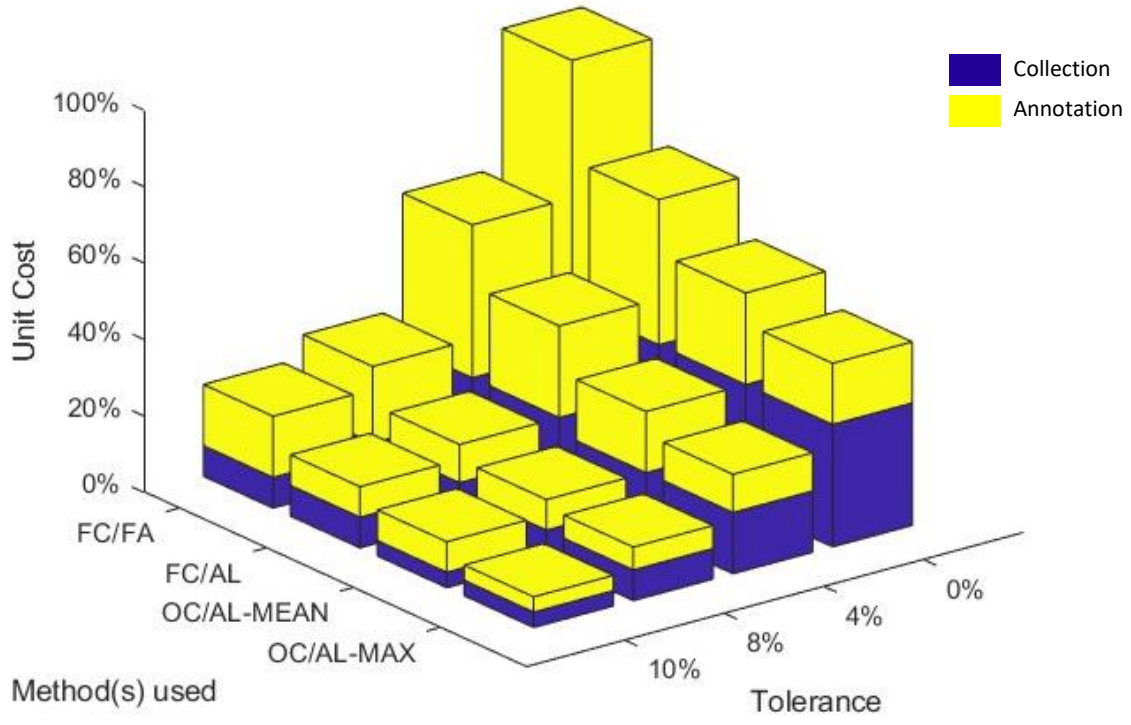


Figure 4.16- Cost saving of capture and annotation for methods: Full capture/Full annotate (FC/FA), Full capture/Active learning (FC/AL), optimised capture/Active learning from mean accuracy (OC/Mean-AL), optimised capture/Active learning from max accuracy (OC/Max)

This case study has shown statistical power curves and active learning allow for significant optimisation in both sample and annotation set size. This reduction in sample

size represents a direct cost reduction in producing a viable dataset. Through the example case study on the BUSI dataset, this gave a 50% cost reduction for an accuracy loss of 4% when considering mean response or a 66% cost reduction for an accuracy loss of 3.75% from theoretical maximums at full dataset size using Alexnet as a performance benchmark. Similarly, if theoretical maximum accuracy is required, the method allows for a 40-50% cost reduction with negligible loss in accuracy depending on the robustness criterion used, demonstrating the power of active learning in boosting accuracy at low sample numbers. Even when using just phase 2, a cost reduction of ~25% is feasible for no accuracy drop using active learning to take some of the annotation burden. If the case study were a 'quick pass' feasibility study, then a massive 90% cost saving can be made for an accuracy trade-off of 10%. Although cost is important, this would be most significant in terms of time as it allows proof of concepts to be demonstrated quickly and efficiently. This method is a powerful tool for planners to maximise gains and productivity under a fixed budget or time frame.

4.4. Discussion

Estimating clinical trial sample size is a standardised practice allowing clinical researchers to fit the size of studies, so they are feasible clinically and financially within the timeframe available. This same approach has been used to predict the effectiveness of increasing the dataset for machine learning and inform researchers as to the usefulness of further annotation. Examining the results from the three datasets in this simple classifier per image study, the neural network response to sample size trend is clear and can be exploited to cap data collection and annotation costs. The breast and

lung datasets both showed diminished returns after 40-50% of the dataset with the much larger foetal dataset reaching diminishing returns between 10-20%. This means that data collection and annotation can be reduced without significant accuracy loss, although the exact cost savings is dependent on the dataset.

Of course, in this study we have access to the full datasets, which we use to demonstrate relative cost reductions and accuracy deviation, but this process is designed to work where the dataset is unknown or incomplete, to predict how big it needs to be and how much should be annotated. When considering the cost saving threshold, it is important to consider the cost of misclassification as part of the empirical fit, understanding the importance and precision requirement of each classification will greatly affect where cost savings can be made. This is an iterative process that can be done throughout the data collection process to plan subsequent data capture and annotation with each successive cycle providing a more accurate indicator of how much additional data is required to achieve the desired results. This is due to the power curve convergence observed (Figure 4.14), where the power curve converges onto a stable value predicting accuracy for arbitrary sample sizes like that shown in Figure 4.7.

This process can also be used retrospectively to determine possible accuracy increases from additional sampling. Considering Figure 4.17, for the BUSI dataset, extrapolation of the BUSI power curve suggests that additional accuracy can be achieved but a doubling of sample size will only yield a 4% improvement in accuracy. Similarly, no improvement is expected for the large foetal dataset with increasing sample size, allowing decision makers to plan appropriately.

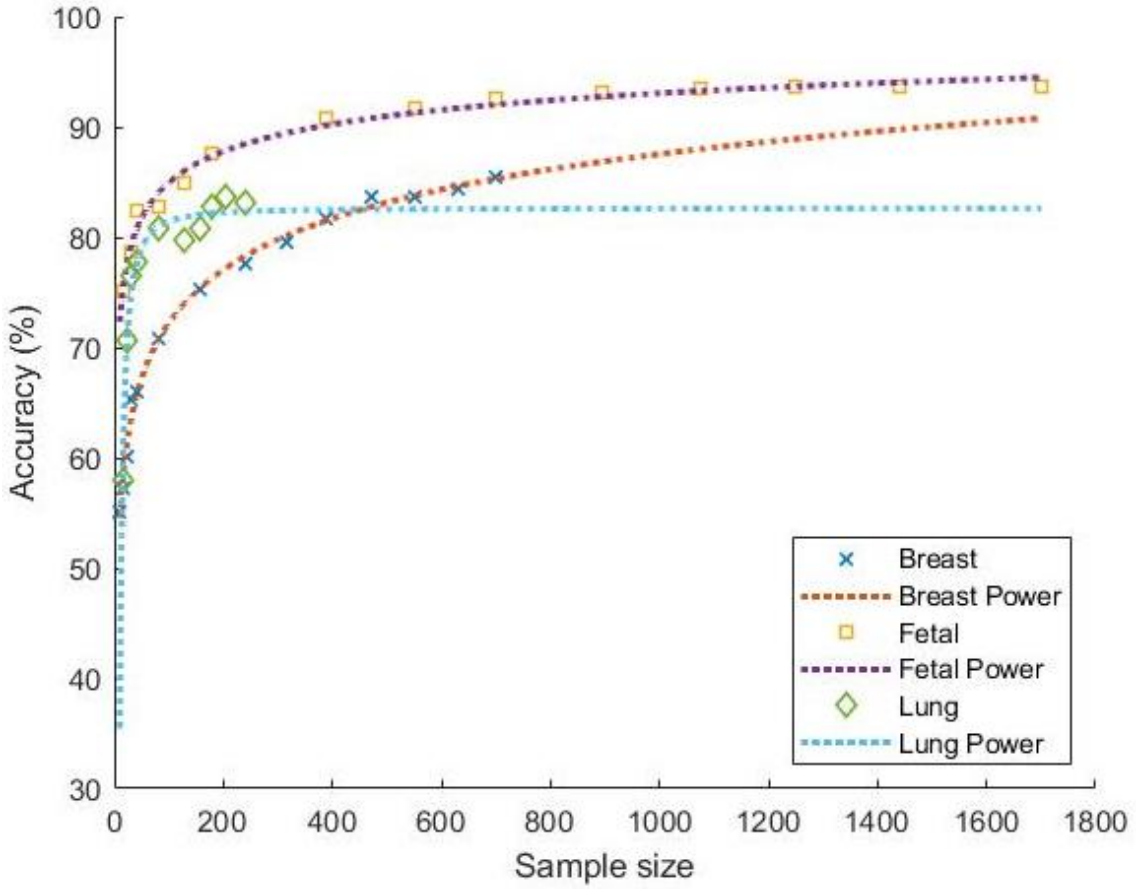


Figure 4.17 - Power curve extrapolated trends from mean neural network accuracy results of all three datasets normalised to a patient set sample size. This chart highlights the extrapolation trend which can be used to estimate additional sample size requirements for each dataset.

When evaluating the effectiveness of active learning across each of the datasets using Figure 4.11, Figure 4.12, and Figure 4.13, there is a measurable decrease in dataset annotation requirement but only up to ~60% dataset usage, above this point, it provided minimal improvement over untargeted annotation methods. Using a targeted approach such as active learning it is possible to further reduce the sample size by identifying

where the neural network is least sure of the result, but this effect also diminishes in value as the training set increases in size.

4.5. Conclusions

This chapter demonstrated a biphasic method that can be used to perform a cost analysis for the collection and annotation of data. Three publicly available ultrasound datasets were investigated using the prescribed method:

- Using an empirical curve-fit model of sample size determination was shown to provide an indicative method for determining cost and providing a method for scaling research studies at the cost of stability.
- Uncertainty sampling with active learning provides a cost-effective method of augmenting manual annotation by targeting samples with the lowest confidence for human annotation while those with high confidence can be annotated using active learning.
- Defining an error tolerance on required performance metrics can be used with this framework to substantially lower cost. With a 50% cost reduction possible in the case study with an error tolerance of 4%.
- The point of diminishing returns can be clearly defined using this method, potentially reducing over collection for that specific use case.
- Using this framework aligns machine learning research with other clinical trials that already use these and similar sampling techniques.

This framework provides ultrasound researchers with an empirical method to answer the question ‘how much data do you need?’ using power theory to identify the most effective sample size and therefore cost of future collection but also a method to maximise the effectiveness of manual annotation, balancing the requirements of many regulatory bodies with the need to control annotation costs in smaller studies. In order to progress machine learning research further in ultrasound, significant investment in data collection and annotation will be required, but this burden can be significantly reduced by scaling feasibility studies and using targeted sampling methods. Extrapolating results from pilot studies using power theory to design future clinical trials that are feasible both clinically and financially is common practice within medical research but has yet to be widely applied in the context of machine learning for medical imagery despite facing the same sampling problem.

The use of the proposed methodology will allow researchers to not only predict the cost of future studies, but also provides a framework for scaling studies that will allow more ultrasound machine learning studies to be funded in future by providing a clear empiric indicator of expected performance that is easily converted to cost metrics. Targeted annotation using uncertainty sampling provides a robust method of augmenting manual labelling maintaining the focus on human annotation as primary focus of labelling, allowing the clinician to focus on targets with low predictive certainty, with semi-supervised active learning labelling those with high confidence. The implications of lower cost studies with clear empirical indicators of results that can be expected in future studies, is that machine learning research into ultrasound will become a less risky

endeavour allowing for more prospective studies to be conducted and more ultrasound data suitable for machine learning to become available to the academic community.

Chapter 5

Using Positional Tracking to Improve Abdominal Ultrasound Machine Learning Classification

Abstract

Diagnostic abdominal ultrasound protocol is based around gathering a set of image cross sections that ensure the coverage of relevant anatomical structures during the collection procedure. This allows clinicians to make diagnostic decisions with the best picture available from that modality. For large protocols like those commonly performed on the abdomen, traditional image only machine learning classification can provide only limited functionality, for example it can be difficult to differentiate between multiple liver cross sections or those of the left and right kidney from image alone. In this proof of concept, positional tracking information was added alongside image data as an additional input to a neural network. This was done to provide the additional context required to recognise these otherwise difficult to identify cross sections. In this chapter optical and sensor based infrared tracking (IR) was used to track the position of an ultrasound probe during the collection of clinical cross sections on an abdominal phantom. Convolutional neural networks were then trained using both image-only and image with positional data, the classification accuracy results were then compared. The

addition of positional information significantly improved average classification results from ~90% for image-only to 95% for optical IR position tracking and 93% for Sensor-based IR in six common abdominal cross sections. The addition of low-cost positional tracking to machine learning ultrasound classification will allow for significantly increased accuracy for identifying important diagnostic cross sections, with the potential to not only provide validation of adherence to protocol but also could provide navigation prompts to assist the user in capturing cross sections in future.

5.1. Introduction

Diagnostic ultrasound relies on the capture of cross-sectional images of anatomical structures within the body to provide a clinician with the requisite information to make a clinical decision. Capturing these anatomical cross sections is time consuming and requires a high level of user skill in anatomy and ultrasound operation [28, 408]. Machine learning has the potential to reduce the skill floor by assisting and automating ultrasound capture procedures [409], but to do so it must overcome the two fundamental difficulties described in Chapter 3: The differentiation of anatomical cross sections that are in close proximity and those that are visually similar. This is exampled in previous studies [331, 332] showing that both experienced clinicians and neural networks have substantial difficulty classifying abdominal cross sections where the anatomical structures were visually similar from image alone.

Machine learning has previously been used in the classification of 11 abdominal cross sections [331, 332] achieving respective accuracies of 77.9% and 82.2% using transfer learning. The use of segmentation and landmarking [332, 410] was also shown to improve accuracy with models achieving 85.2% and 83.4% respectively, with increased accuracy possible if errors from similar cross sections were excluded. These studies show reduced accuracy where cross sections overlap or have visual similarities. Where a distinct dataset is used, that avoids these overlaps and visual similarities, accuracies of between 95.7% and 98.6% can be achieved [333]. This further highlights the limitations of using an image-only approach for abdominal cross sections, due to the lack of distinctive landmarks where there are overlapping classes within the imagery. Therefore,

additional identifiers should be sought. Positional data has been previously used in medical ultrasound applications [411] such as 3D image reconstruction [412] and biopsy [413], but has not been utilised to assist machine learning in improving classification of diagnostic abdominal cross sections.

In order to test the efficacy of positional based tracking of an ultrasound probe for machine learning, two separate systems were tested: Optical infrared tracking (IR) using a Vicon system, and an IR system based upon low-cost application specific integrated circuits (ASIC) IR sensors. Vicon has been shown to be highly accurate to within 2mm [414], and is effective as a positional and registration reference measurement in other medical imaging applications [77, 415]. It also has been shown to achieve high accuracies in motion capture, as part of complex automated classification processes such as respiratory tracking [416] and pose estimation [417]. The use of optical tracking would be difficult to implement within a clinical environment, due to the need for a large camera gantry. This necessitated the design of a smaller more mobile IR tracking system, better suited for the small spaces found in clinical areas. IR tracking has shown to be highly accurate at tracking while maintaining a low latency [418] with previous studies of similar positional systems being capable of tracking an ultrasound probe mounted to a robotic arm [419], spinal column tracking [420, 421], and tracking operator movements when applying machine learning to scanning the median nerve and radial artery [422].

5.2. Structure and Scope

This chapter seeks to present a proof-of-concept method to improve machine learning classification accuracy for abdominal scanning using positional information to augment image-based classification. This chapter first compares image only machine learning classification to optical IR tracking within a Vicon system. Sensor-based IR tracking was then tested using a modified HTC virtual reality tracking system. The use of the sensor-based IR tracking, while less accurate than Vicon, is to demonstrate the addition of positional tracking using a mobile sensor which would be more indicative of what could be used in a clinical environment. This chapter does not seek to compare positional tracking precision, but the resultant output of the neural network classification using these tracking systems. This is to show how effective positional information is at improving classification of difficult to identify ultrasound cross sections and edge cases.

5.3. Method

In order to make an effective comparison between image-only neural networks and those augmented with positional information, image and positional data was collected for six standard clinical abdominal cross sections and three normalisation points using an ultrasound abdominal phantom. This was performed within a laboratory environment using a medical ultrasound device and the optical or IR sensor positional tracking systems respectively. This data was then pre-processed into an image tensor and file containing classifier and raw coordinate output from the positional device to produce the dataset. The dataset was then split 80/20 at the session level to prevent data leakage and

used to train a three-channel image only model and then a four-channel image and positional model. This model was validated using the unseen validation set data and results outputted, the dataset was then re-split and the experiment repeated. This method is shown in the flow chart in Figure 5.1.

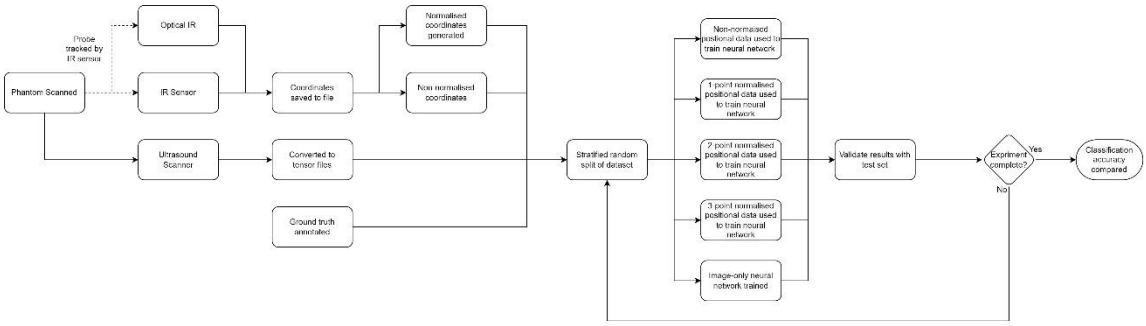


Figure 5.1 - Flow chart of positional tracking pipeline. Ultrasound and positional data are collected. Ground truth and additional normalised coordinates generated. Dataset is split and networks trained. Cycle continues until experiment complete.

5.3.1. Dataset

A Kyoto Kagaku ‘Echozy’ ultrasound phantom (Kyoto Kagaku Co., Ltd., Japan) as seen in Figure 5.3, was scanned using a SonixTouch Q+ medical ultrasound system (SonixTouch, BK Ultrasound, USA) using a curved array, 5-2/60 ultrasound probe. These images were captured via HDMI cable using OpenCV [423] and were stored as .jpeg and .pt 3-dimensional tensor files. Six cross sections were chosen as regions of interest (Figure 5.2):

- Right hypochondrium transverse approach for common bile duct.
- Right intercostal approach sweeping through the liver to visualise the right portal vein.
- Right hypochondrium longitudinal approach for the Gall Bladder.
- Epigastric longitudinal approach sweeping through the aorta.
- Transverse approach of the left kidney
- Transverse approach of the right kidney

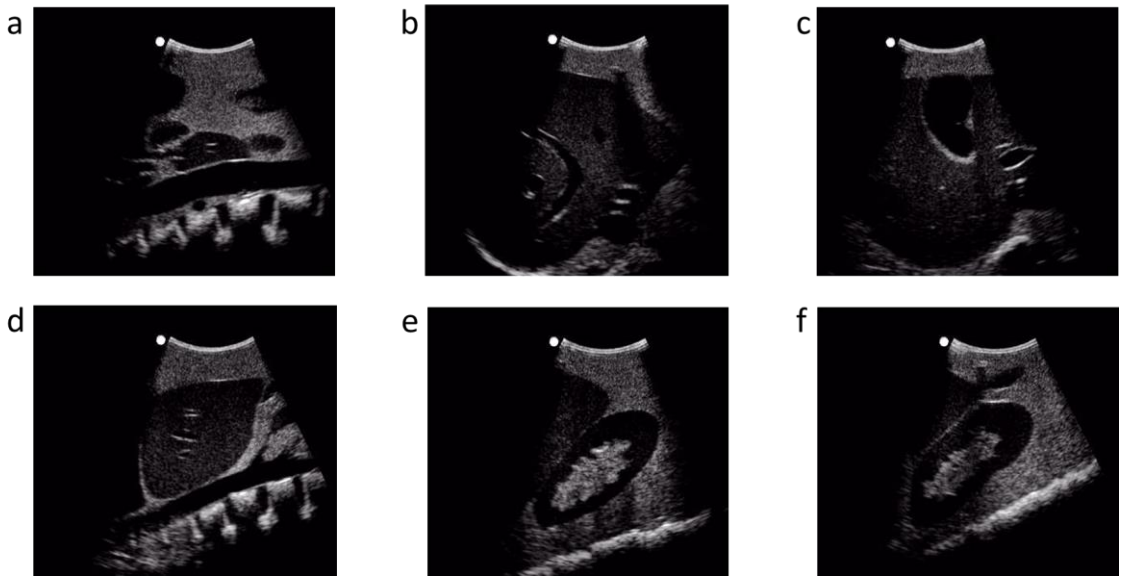


Figure 5.2 – Examples of cross-sectional ultrasound scans of the phantom. a) common bile duct. b) portal vein. c) gall bladder. d) aorta. e) left kidney. f) right kidney.

These cross sections were chosen specifically based on classification error seen in chapter 3 [424] and in previous studies [332, 333, 425] and due to visual similarity, such as with the left and right kidneys and over lapping region of interest such as with Gall bladder and Common Bile Duct. Complex sweep scans of aorta and portal veins

that contain both visual similarities and overlapping anatomical structures were also chosen to provide added complexity to classification.

Figure 5.3 - Image of Kyoto Kagaku phantom.

The optical IR dataset is made up of 137 sets of scans totalling 18,614 images, the IR sensor dataset is made up of 22 sets of scans totalling 3410 images (Table 5.1). Images were captured at a rate of 5 frames per second. Each set was performed as if scanning an individual patient with the sonographer using minor pressure and angle variation during the capture process while ensuring that the target region of interest (ROI) was visible and would adhere to standard clinical collection protocols. This was done to provide additional natural variation in the images.

Table 5.1 – Comparison of size and composition for the optical IR and IR sensor phantom datasets.

Cross Section	Optical	Infrared
Left Kidney	3611	630
Right Kidney	2629	563
Aorta	3335	725
Bile Duct	3488	438
Gall Bladder	2854	426
Portal Vein	2697	628
Total Images	**18614**	**3410**

5.3.2. Tracking system

Two methods of probe tracking for generation of coordinate data were tested: Vicon optical IR tracking and IR sensor-based tracking. While the ultrasound and positional tracking systems were all capable of a high rate of capture, a capture rate of 5 frames per second was used to prevent any de-synchronisation due to potential changes in system latency throughout the scanning process. As both Vicon optical IR and IR sensor tracker require line of sight and operate within the same frequency band, separate sessions were performed for each positional system to minimise any potential interference. Electromagnetic sensors would provide a system that does not require line of sight, but these systems are expensive and not readily available.

5.3.2.1. Vicon optical-based IR tracking

The Phantom was placed on a non-reflective surface within a fully calibrated Vicon optical measurement volume utilising a Vicon MX Giganet system [426] with 12 Vicon

T160 cameras (16 MP, 18 mm focal length lens) mounted to a professional camera rig (Figure 5.4). These cameras detect light reflected off tracking dots at a wavelength of ~850nm. Volume calibration was performed by placing the origin point on the floor 1 meter from the lab desk ensuring that coordinates were as similar as possible between sessions. The ultrasound machine with the screen at its lowest position and laptop were placed at least 2 metres from the phantom and masked in the calibration setup to prevent interference with tracking. Vicon tracking markers were affixed to the probe, phantom and a Y frame that had been secured to the probe. The addition of the Y frame allowed for additional distance between tracking dots therefore increasing the sensitivity of the optical camera imagery and also ensuring line of sight can be maintained while the operator was positioning the probe. The Vicon API was used to stream the coordinates into python which was captured at a rate of 5 frames per second via a Wi-Fi connection from the laptop capturing the ultrasound images to a computer running the Vicon optical tracking development kit.

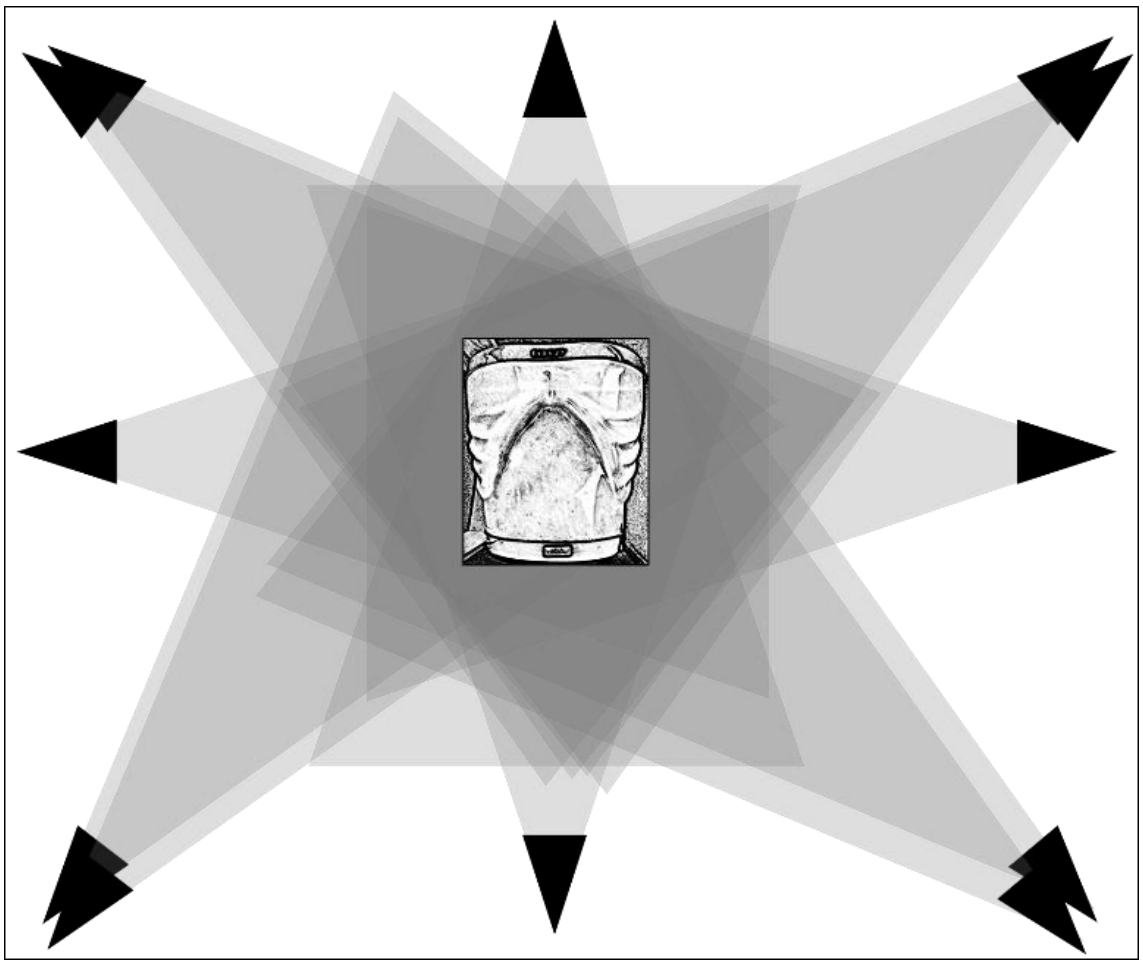


Figure 5.4 – Diagram showing the optical IR camera tracking rig setup with the phantom inside the visual field.

5.3.2.2. IR Tracking

This positional system is a modified setup based on those used for full body tracking for VR [427]. The system itself had been modified to track a single HTC VIVE (3.0) tracker (pictured Figure 5.5(a)) [418, 428] which was affixed to the ultrasound probe using a strap and hot glue. A Steam VR base station (2.0) (pictured Figure 5.5(b)) [429] was attached via a mounting strap to the ultrasound cart which was positioned anteroinferior to the Phantom ensuring clear line of sight (Figure 5.6). The base station produces pulses

of infrared light at a wavelength of ~850nm which is then detected by simple ASIC IR sensors on the VIVE tracker. VIVE tracker has previously been shown to be accurate to within 0.68 ± 0.32 cm translationally and 1.64 ± 0.18° rotationally [420] in comparison to the Vicon tracking system.

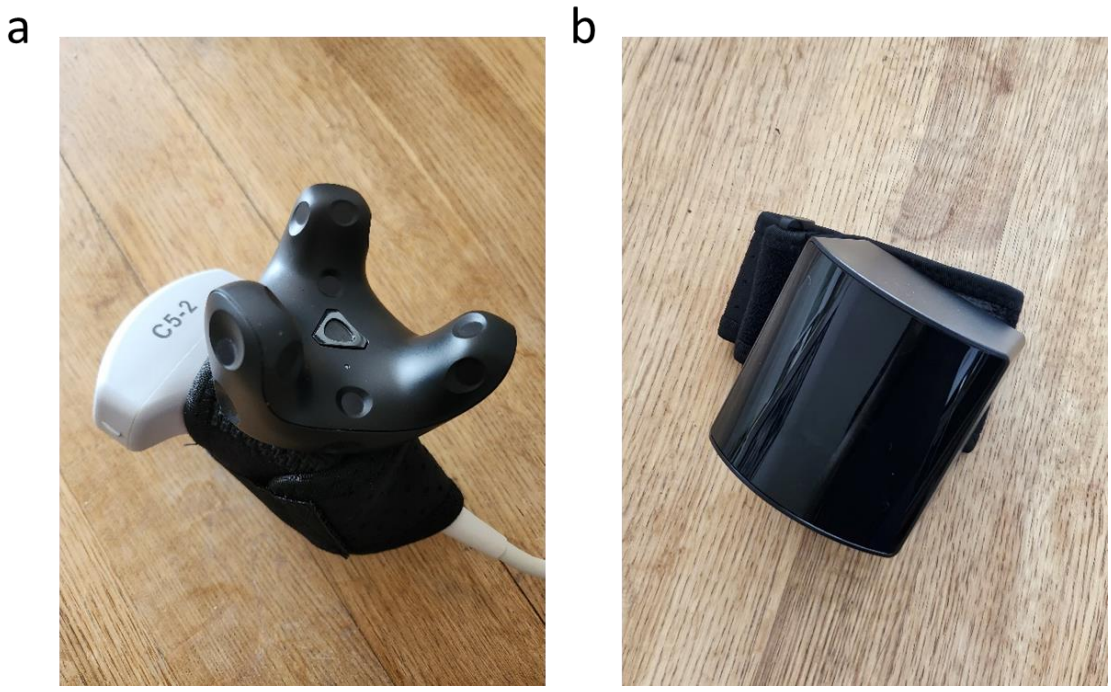


Figure 5.5 -Images of the IR positional sensor system a) probe with VIVE tracker. b) Steam lighthouse sensor with strap.

The Base station was moved after each collection set to mimic moving to a new patient or clinical space. Note that anterosuperior scans were performed but excluded as they provided conflicting reversed positional data, this data could have been used if the angle of the phantom was tracked during the IR experiment, or a second base station used to provide additional point of reference. Software requirements for a headset and VR stage were bypassed by using a modified system profile and using developer options within

the Steam VR software. OpenXR [430] was used to extract the coordinates from the VR runtime with a modified API used to stream the coordinates into python which was captured at a rate of 5 frames per second using a USB cable to the Steam VR base station.

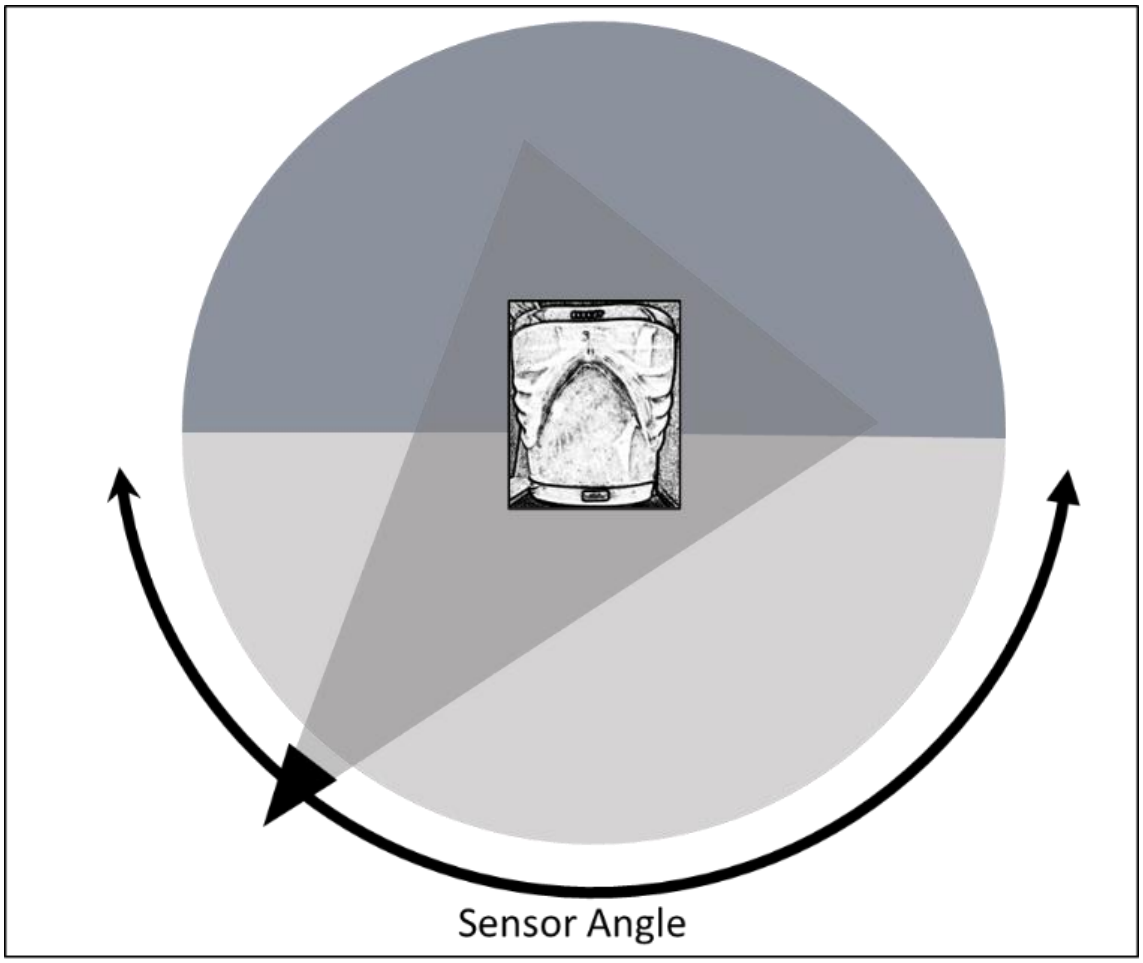


Figure 5.6 – Diagram showing the position of IR tracking rig setup in relation to the phantom during experimental data capture.

5.3.2.3. Phantom Coordinate Normalisation

In order for the positional data to be used to effectively track the ultrasound probes movement it is necessary to normalise coordinates provided to the neural network so that they are of similar scale. In order to test normalisation methods, scans of three fixed points on the phantom were taken before each set of scans was performed as shown in Figure 5.7:

- On the right midclavicular line, between the right 9^{th} and 10^{th} ribs.
- The probe is positioned on the xiphoid notch along the midsternal line with the probe positioned anteriorly.
- On the left midclavicular line, between the left 9^{th} and 10^{th} ribs.

These anatomical points on the ribcage, are less subject to variation due to patient positioning or disease process, are not subject to patient dignity concerns, and can be precisely and consistently pinpointed by a clinician. Use of soft tissue landmarks such as the umbilicus would be impractical in cases with abdominal distension where these features would be subject to greater variation. These defined points on the abdomen were used to normalise the coordinates for each axis, where multiple points are used, a simple mean is used to provide a single normalisation point. This normalisation point was then applied during the conversion to positional tensor.

Post-capture normalisation for the optical data was not required, as the optical IR positional data was automatically calibrated to a point within the measurement volume during each collection session, meaning that differences in coordinates between scan

sessions was very small. However, the IR sensor base station was moved after each cycle of data collection to represent moving between patents and potential changes in clinical area.

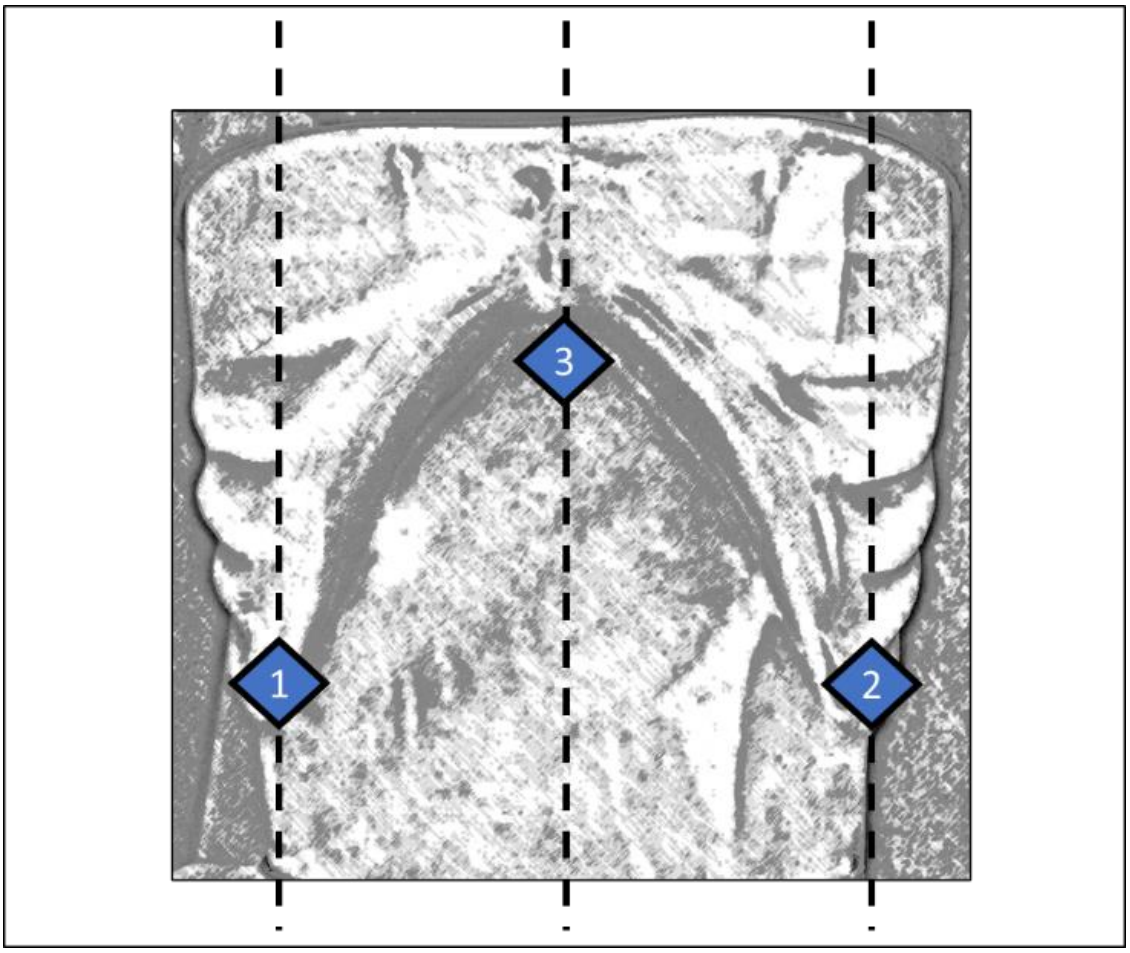


Figure 5.7 Representation of the relative position of the three points of normalization on the human abdomen. The three points are: 1) between ribs 9-10 on right midclavicular line, 2) between ribs 9-10 on left midclavicular line, 3) horizontally positioned on xiphoid notch.

In order to evaluate the amount of normalisation required prior to input into the neural networks four sets of normalisation data were produced by setting a new zero point:

- No normalisation – using the original captured values.
- 1-point normalisation from point one on the anatomical right side of the phantom. This locates a single point on the abdomen within the tracked volume.
- 2-point normalisation using a simple mean of points 1 and 2 on the phantom. These measurements would allow for the sizing of the abdomen along a single dimension.
- 3-point normalisation using a combined simple mean of all three normalisation points. This would allow for the two-dimensional sizing of the abdomen.

These points are used to generate an origin with which all subsequent coordinate data is normalised to fit. Where multiple normalisation points are used, a mean point is generated and used as the origin. Training of the neural network was performed for image only and for each of the normalisation points using the same dataset split so that results could be compared.

5.3.3. Machine Learning Implementation

All training and testing was performed on a 64bit version of Windows 10, using a intel core i9 and Nvidia 40 series GPU using python [319] (version 11.4) and the CUDA toolkit (version 11.7). The SciPy metrics library was used to analyse model output. A pre-trained ResNet-50 [346] convolutional neural network from the torchvision library was used as the basis for study, with weights based on ImageNet challenge dataset [339].

The final layer of this network is adjusted to output 6 classes. Image-only method uses the default 3 channel neural network. For the positional study, the neural network was adjusted to accept a 4th channel for the inclusion of the positional data. While Inception-based networks were highlighted as providing the highest accuracy in chapter 3, Resnet-50 was selected as it produced a more consistent training result across the baseline, as well as datasets 1 and 2.

As the dataset consists of scans of a single phantom, overfitting is a concern, as such the experiment was repeated 50 times to provide an average training response, over a maximum of 5 epoch using early stopping [431] and a small batch size of 64 to promote better generalisation [432]. Training used a learning rate of 1.00e-04 using the ADAM optimiser [349]. Training and validation methodology was identical for both 3 and 4 channel versions of the network.

The dataset images were converted into tensors with 3 channels of size 330x370 pixels. The optical dataset was split 80/20 into training and validation sets for each experimental run, with both the 3 and 4 channels networks trained and validated using this split, so that a direct comparison between image-only and positional tracking could be performed. For the positional experiments the IR dataset was split 50/50 between training and validation sets, this was done to increase the size of the validation set in the IR sensor experiment, to reduce the effects of overfitting due to the small sample size. The positional data was converted into a tensor, the normalisation sum performed and input into the network alongside the 3 image channels. Training was repeated for each

normalisation state on the same data split to ensure comparison could be performed. The datasets were split at session level to prevent bias due to data leakage.

5.4. Results.

5.4.1. Image only vs optical positional tracking classification

Neural networks trained using the optical dataset produced average accuracies of 91.47% for image-only based training and 95.75% with the addition of positional data, an average improvement of 4.3% (Table 5.2). The largest accuracy improvements can be seen in classification of the bile duct (6.4%) and portal vein (7.8%). The highest performing image-only network achieved an accuracy of 96.34% with the largest error in the classification of Gall Bladder and Bile Duct. The highest performing optically tracked network achieved an overall accuracy of 98.84%, with errors in aorta, bile duct and gall bladder classification. Variance in training outcome decreases significantly between the networks trained with positional data, achieving an average reduction in variance of 23%, this can be seen clearly in Figure 5.8 with significant reductions in variance in all classifications.

When statistically comparing image-only and optical IR tracked results by performing a twin tailed T-Test with the assumption of heteroscedastic variance (Table 5.2), the optical IR tracking results proved to be statically significant with an averaged P value of 0.0482. When the results are analysed on a class by class basis, results are shown to be highly significant achieving P-values <0.003, however there is insufficient statistical significance when comparing aorta classification results (p-value 0.2436).

Table 5.2 – Comparison of the average accuracy results for networks trained with image-only and augmented optical IR tracking. The training variance of the 50 neural networks trained with each method is shown.

	Image Only	**Optical**	**Accuracy Improvement**	**Training Variance**	**Comparative P Value**
Left Kidney	97.7%	99.6%	1.9%	19.6%	0.0104
Right Kidney	96.2%	99.2%	3.0%	29.7%	0.0033
Aorta	95.1%	96.9%	1.8%	14.7%	0.2436
Bile Duct	85.2%	91.6%	6.4%	15.4%	0.0019
Gall Bladder	85.2%	90.0%	4.8%	35.1%	0.0298
Portal Vein	89.3%	97.1%	7.8%	23.6%	0.0003
Average	**91.47%**	**95.75%**	**4.3%**	**23.0%**	**0.0482**

Using Figure 5.8 to compare the accuracy all 50 trained networks, the deviation in trained per class accuracy was substantially higher in image only trained network in comparison to those using optical tracking data. Misclassification of gall bladder, bile duct, aorta and portal vein were the largest cause of deviation for both image-only and optically tracked networks. Misclassification of left and right kidney was reduced to less than 3% in optically tracked networks with the average network achieving above 99% accuracy for the kidney cross section images.

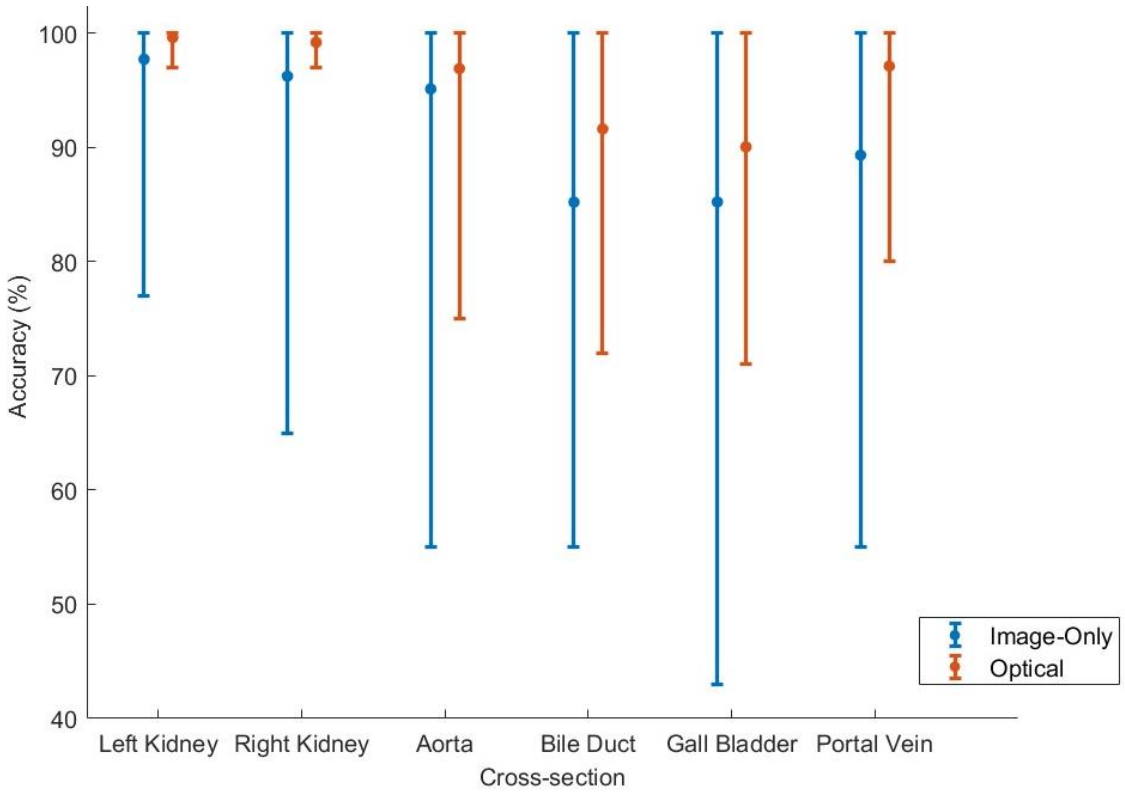


Figure 5.8 – Comparison of mean classification accuracy of abdominal cross sections for image only and optical tracking methods. Error bars represent the deviation in classification accuracy for each cross section over 100 neural networks.

When examining the networks with the highest accuracy using a confusion matrix (Figure 5.9), the largest source of error for both image-only and optically tracked network is between bile duct and gall bladder, this error is present throughout both network types, and was consistent across all 100 networks tested. If we compare the image-only networks to its optically tracked network trained on the same dataset split, the optically tracked network improves upon the image only accuracy result by an average of 4% with optically tracked always improving on its image only counterpart.

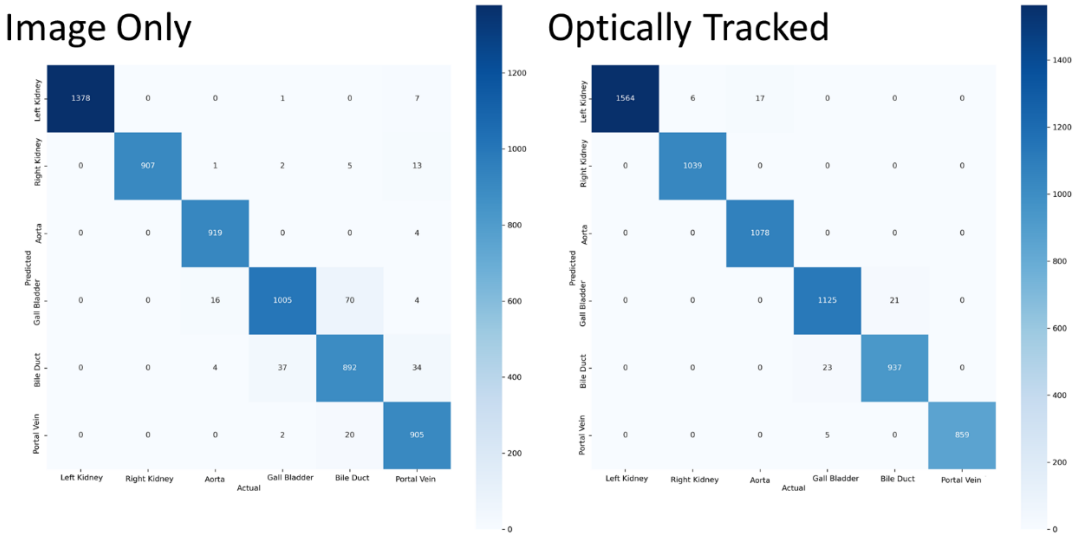


Figure 5.9 – Side by side comparison of the confusion matrix of the highest accuracy neural networks trained on image-only and optical tracking datasets.

The inclusion of the IR positional coordinate data into the training and validation sets (Table 5.3) saw an average accuracy of 89.70% for image only classification. Positional accuracy achieved average accuracy of 92.71% without any form of normalisation, 93.61% for single point normalisation, 93.73% for two-point normalisation, and 93.28% for three-point normalisation respectively. When compared against image-only accuracy results, networks trained on non-calibrated positional data achieved an average improvement of ~3% improvement, with calibrated positional data achieving ~4% improvement in cross section classification. Common bile duct and gall bladder classification were the largest sources of error in both image-only and positional tracked networks. In comparison, when looking at the maximum achieved network accuracy was 97.5% for image only, 98.5% for no normalisation, 98.7% one point of normalisation, 98.3% for two points of normalisation and 97.5% for 3 points of normalisation.

While there is a overall improvement in classification accuracy when using IR sensor tracking with normalisation, a single factor ANOVA test showed that there is insufficient statistical significance (F-Value 0.3521, P-value 0.7038) in the results to distinguish between 1, 2 and 3 points of normalisation. This is likely due to a limitation of this study as there is insufficient difference in the size of the abdominal cavity to confirm efficacy of normalisation.

Table 5.3 – Comparison of average accuracy for networks trained on image-only and IR positional sensor datasets. Average accuracy results for networks trained with coordinates normalized with 0, 1, 2 & 3 points are shown.

	Image only	**No normalisation**	**1-Point normalisation**	**2-Point normalisation**	**3-Point normalisation**
Left Kidney	96.5%	98.2%	99.5%	99.3%	99.5%
Right Kidney	96.7%	99.1%	99.0%	98.9%	99.0%
Aorta	95.4%	95.7%	95.6%	95.3%	92.6%
Bile Duct	81.1%	86.6%	87.1%	85.2%	85.8%
Gall Bladder	78.7%	84.9%	82.7%	88.2%	85.8%
Portal Vein	89.8%	91.7%	97.7%	95.4%	97.0%
Average	**89.7%**	**92.71%**	**93.61%**	**93.73%**	**93.28%**

A comparison of the accuracy of the 50 trained networks (Figure 5.10) shows that while overall accuracy was improved, there was increased training variance in comparison to optical IR. Despite an increase in overall accuracy, networks trained with positional data with no normalisation saw an increase in training variance by 5.3%, compared to

improvements of 21.7% for one-point normalisation, 18.6% for two-point, and 15.4% improvement for three-point normalisation. Accuracy values were also lower than image-only results for 12 out of the 50 no normalisation networks, as this also occurred in a number of calibrated networks this is most likely due to training variance. This notably did not occur in the more accurately calibrated optically tracked networks.

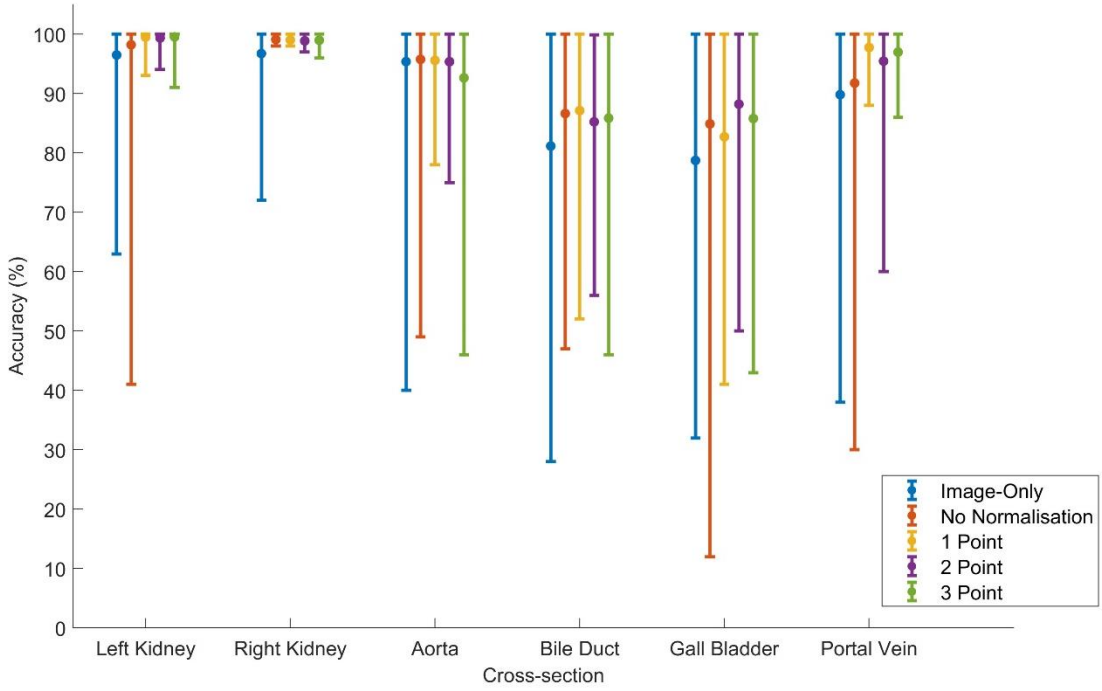


Figure 5.10 – Comparison of mean classification accuracy of networks trained to classify abdominal cross sections using image only and positional tracking datasets. Positional augmented networks show no normalisation and where 1, 2 & 3 points of normalisation have been applied. Error bars represent the deviation in classification accuracy for each cross section over 250 neural networks.

When comparing the confusion matrix for the IR tracking networks (Figure 5.11), the gall bladder and bile duct are the most confused classifications. This result is directly comparable to that seen in Figure 5.9.

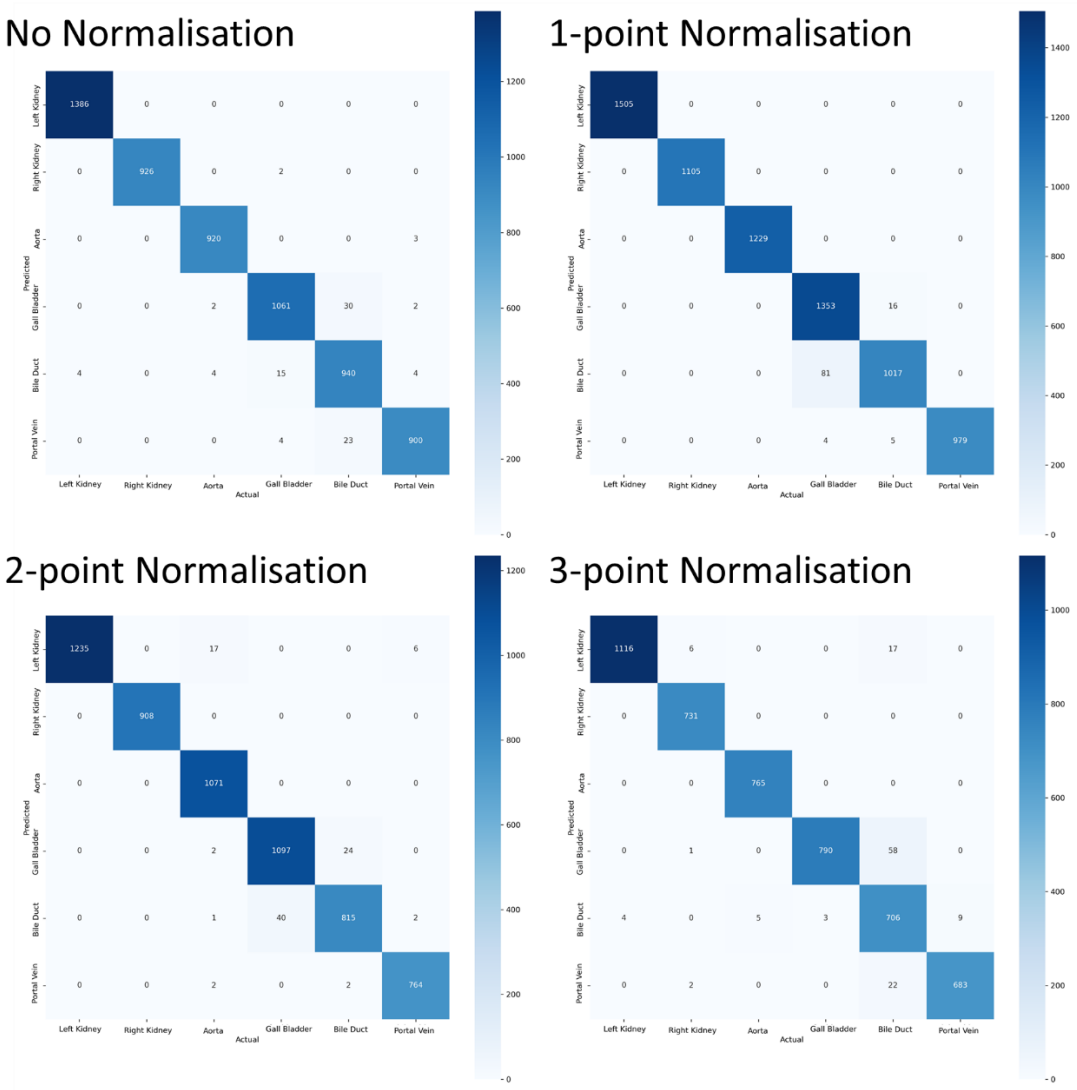


Figure 5.11 – Side by side comparison of confusion matrix examining the effect of normalisation of coordinates on neural networks trained data augmented with IR Positional tracking.

5.5. Discussion

This chapter demonstrates the use of positional data to improve classification of abdominal ultrasound cross sections on an ultrasound phantom using both optical and IR tracking systems. On average neural networks trained on optical tracking data provided the highest accuracy network, followed by IR tracking, and standard image-only classification.

5.5.1. Study Limitations

While the phantom was designed to provide an accurate representation of ultrasound cross sections for clinical training purposes, it is an idealised representation of a human abdomen and cannot fully represent the difficulties usually encountered during image acquisition in ultrasound scanning such as:

- Shadowing is limited as phantom materials do not have the same density range that would reduce amplitude and obscure ROI.
- Attenuation changes from different tissue thicknesses or densities present in the anterior abdominal wall are not represented in the phantom.
- Image artifacts common in ultrasound such as those from the digestive tract are not present in the phantom, as these artifacts, such as gas pockets are not represented in the phantom.The phantom is of fixed size and as such cannot represent different sized abdominal cavities, however this does not reduce clinical applicability as while the scale would change, in a high proportion of

cases the position of cross sectional landmarks would remain the same in relation to one another, and as such can still be used for positional identification.

The use of a single subject (the phantom) also increased image-only classification accuracy as there is a high chance that the neural network will recognise the identical structural features. It is also important to note that as the data was collected by a single sonographer, operator bias in the collection of cross-sectional imagery cannot be ruled out. This does not however prevent comparison of image-only and positional performance as the same subject and image set is used therefore any bias is present in both experimental tests.

The use of normalisation did increase overall accuracy by ~4%, but there is very little difference between 1, 2 and 3 points of normalisation. This is likely a limitation of the experimental setup, the phantom is a fixed size, once a fixed point on the abdomen is located, no additional variation in abdomen shape or volume is required to be taken into account. In a human trial the abdomen could potentially have much greater levels of variation and therefore the requirement of additional normalisation points should not be discounted in future experimental trials. The networks using positional data with no normalisation had the most variance in accuracy result, achieving the worst performing network at 73.3%, but still outperforming the image-only networks on average. This is likely partially due to the error in rotatory angle [a, b, c] being much smaller than that of positional [x, y, z] data. This would be particularly useful in the recognition between left and right kidney, which maintained accuracy comparable to calibrated trained networks

despite providing positional [x, y, z] values that were likely incompatible with those already seen by the network during training.

5.5.2. Accuracy

While accuracy has been used as the main metric throughout this chapter, examining the harmonic mean for the highest accuracy neural networks (Table 5.4) confirms high precision and recall for all methodologies used. This is due to the limited subject matter available with using only one phantom. Despite using a single phantom, overfitting has been sufficiently reduced using variation in the image capture technique, early stopping, small batch size and experimental repetition to provide indicative results, there is a distinct correlation between the use of positional information when training a neural network and the improvement of classification result.

Table 5.4 – Harmonic mean F-1 score for highest accuracy neural networks trained with optical and IR tracking augmented dataset in comparison to image-only.

	Image Only	Optical Tracking	IR No Normalisation	IR 1-Point	IR 2-Point	IR 3-Point
Left Kidney	1	0.99	1	1	0.99	0.99
Right Kidney	0.99	1	1	1	1	0.99
Aorta	0.99	0.99	1	1	0.99	1
Bile Duct	0.94	0.98	0.97	0.96	0.97	0.96
Gall Bladder	0.91	0.98	0.96	0.95	0.96	0.92
Portal Vein	0.96	1	0.98	1	0.99	0.98

When comparing the image-only accuracy results between Table 5.2 and Table 5.3, there is a significant drop in the average classification accuracy of the gall bladder and bile duct between networks trained on the optical set alone vs those trained with additional images from the IR sensor dataset. This is not however reflected by the level of accuracy achieved by the highest performing image-only networks with the optical image-only achieving 96.34% compared to 97.5% with the addition of images from the IR dataset. There was also significantly more variance in training result in the IR image-only networks, with the lowest network result being 79.5%, which is 1.8% lower than achieved by the worst performing optical image-only network. It is important to note that image-only results would be likely be lower with a larger sample size, imagery would also lack the same level of clarity in human trials, where body shape and differences in thickness and density would cause changes to attenuation properties that that would have to be considered.

5.5.3. Dataset Variance and Overlap

As seen in the results for both Optical and IR sensor experiments, the largest source of error for all models was between gall bladder and common bile duct. As this was consistent across both network types it is important to rule out an error within the dataset itself. Analysing the images where this error had occurred revealed that in adjusting the probe position to add variation to the dataset, a number of the images capture both gall bladder and bile duct (Figure 5.12). These images still contain the target features but also cover the other anatomical structure as well. This overlapping visual information is the exact type of edge case that was targeted during cross section selection and exists

within a number of real clinical protocols. There was a significant improvement in accuracy suggesting that probe angle information is making a significant difference in the classification of cross sections where the target ROI overlaps.

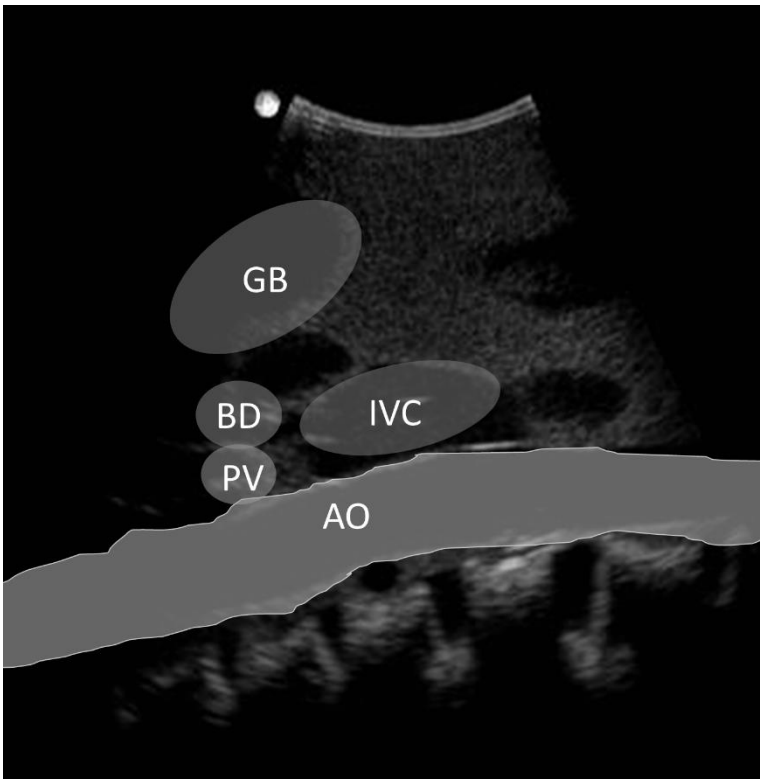


Figure 5.12 – Example of a transverse ultrasound scan on the phantom showing the right hypochondrium: Anatomical regions of interest are labelled: gall bladder (GB), Bile Duct (BD), Inferior vina cava (IVC), Portal Vein (PV), Aorta (AO)

When looking at the variance results in Figure 5.10, there is significantly more training variance in the left kidney when compared to the right kidney, this cannot be quantified by overfitting or dataset imbalance as in Table 5.1, there is significantly more left kidney images within the dataset than the right suggesting that the left kidney should instead be favoured as seen in Figure 5.8. The likely cause is that the left kidney slightly more difficult to visualise clearly in transverse ultrasound scans as it is located slightly

superior compared to the right kidney [433]. A better visualisation could have been achieved by placing the phantom in a lateral decubitus or prone position. The identification of sweeps of the aorta and portal vein were consistently good despite multiple anatomical structures such as the right kidney often being visualised during intercostal scanning.

5.5.4. Clinical practicality

Optical IR tracking was the most precise position tracking used, providing a 99% accurate neural network with the average accuracy result being ~95%, but this was achieved using an expensive Vicon measurement volume with an external calibration software, and required additional hardware attached to the ultrasound probe to maintain line of sight with the camera rig. While excellent at validating positional data as a viable method of improving neural network recognition of abdominal cross section views, it is not clinically practical outside of specialist facilities. Sensor-based IR tracking while being less accurate overall, still achieved ~98% accuracy using simple normalisation techniques with an average of ~93%, higher than that achieved by image-only. The IR tracking system still required line of sight but was compact with the base station able to be easily mounted to the screen arm. The HTC VIVE tracker was small enough not to interfere with scanning, initially attachment strap would slip from the probes ergonomic design, but this problem was easily solved with hot glue which ensured a tracker position on the probe was maintained throughout data collection. During collection of IR scans, the position of the base station relative to the phantom was initially a full 360-degree loop with the base station moved after each set of scans, but due to the use of a

single base station, and the fact no tracker was attached to the phantom, it was not possible to fully localise the positional data. Therefore, scans taken with the base station in an anterosuperior position were mirrored in comparison to the optical tracking data. As such all anterosuperior scans were excluded from the IR dataset instead of manually adjusting these values and potentially adding additional human error to the training set.

5.6. Conclusion

This chapter highlights the potential of positional sensor information as an additional data source when training neural networks on diagnostic cross sections that may be hard to differentiate using image alone. Optical IR positional tracking was highly accurate and substantially increased classification accuracy. Mobile sensor-based IR tracking provided a less accurate, but more practical example of applying positional information to machine learning for clinical use cases but also highlighted a number of difficulties that would need to be overcome before such technologies could be used. With the addition of positional data to contextualise cross section imagery that are in close proximity, or where there is a high level of visual similarity is no longer a challenge as the position of the probe is known both relation to other scans and the patient.

The collection and use of use of positional information as part of an ultrasound scan will allow a neural network to know the position of the probe relative to the patient, opening up many exciting opportunities for future research. Immediate future work should focus on increasing the sample size using a cadaver study to further test data and normalisation requirements. Electromagnetic sensors will be tested as a method of probe tracking as

this technology does not require line of sight. An Electromagnetic sensor system is already in use for the formation of 3D ultrasound images and could be repurposed for use in positional sensing. Neural networks that can localise the probe to the position on the abdomen can provide feedback to the sonographer to assist in the positioning and fine tuning of the probe for the collection of potentially higher quality ultrasound cross sections that fully capture the required anatomical structures as mandated in the clinical protocol. It would also allow a more experienced user to sweep the probe over the region of interest with the neural network selecting and potentially annotating the required cross sections automatically, speeding up scan times and reducing workload.

Chapter 6

Suitability of Theil Cadaver for Classification of Abdominal Ultrasound Cross Sections

Abstract

The use of cadavers as a teaching and research tool is well known to medical scientists, thanks to their vast contribution to the field they are often known as the ‘silent teachers’, allowing clinicians and researchers the opportunity to study areas otherwise difficult to do so. As researchers increasingly include machine learning into their studies, it is important to validate the role of the cadaver as part of a dataset for machine learning. This chapter examines six common abdominal ultrasound scans from eleven cadaver for their suitability for furthering the study of image and positional sensor-based machine learning for abdominal cross section classification. This dataset was then used to train a neural network to examine the efficacy of positional tracking in comparison to image-only classification of cadaver and if normalisation coordinates in the abdominal cavity improves classification. Findings suggest that Thiel cadavers provide a good testbed for ultrasound imaging research, as the tissue maintains very similar ultrasonic properties. The variation in anatomical visibility coupled with the variability of abdominal cavity size of a cadaver allowed for additional validation of the positional tracking and

calibration system for machine learning classification in comparison to the use of phantom, improving overall generalisability of machine learning models. However, the visual parameters of physiological structures within the cadaver can be drastically different, potentially confounding any image-based training if the dataset is not sufficiently large. While an image-only approach failed due to the large variation within the cadaver image sample set. The addition of positional inferred sensor data allowed for the networks to achieve average classification accuracies of 88.3% for one point, 91.5% for two-points and 92.8% for three-point patient normalisation.

This result suggests that positional tracking could therefore substantially improve recognition of edge case and difficult to identify diagnostic ultrasound cross sections. Cadavers must be carefully selected to ensure that the target organs are clearly visible using collection techniques that ensure that the anatomy is not altered too drastically by the method of collection. The use of machine learning to assist in the collection of ultrasound diagnostic cross sections could not only improve clinical workflows by automatically collecting the best image and supporting decision making, but it also provides a route towards automating the collection process.

6.1. Introduction

The use of machine learning for diagnostic ultrasound of the abdomen is a valuable tool in medical imaging research. However, the collection of ultrasound scans can be difficult and expensive, especially if the disease being studied is rare or difficult to detect. Cadavers have long been used by clinicians and researchers for the collection of medical data, but the effectiveness and suitability of cadavers for image-based abdominal ultrasound machine learning has seen limited study. The previous phantom study suggested that positional data could improve the overall classification of ultrasound abdominal cross sections, but the study was limited due to the small sample size. A cadaver study has the potential to increase sample size, providing greater visual variation and abdominal cavity size, increasing the validity of the proof-of-concept results while not being as costly and time consuming as a clinical trial. Cadavers embalmed using the Theil method are soft and malleable maintaining much of their original tissue properties [434-437] allowing for much more pliability and range of motion, because of this Theil cadavers are increasingly seen as useful tools for medical teaching and research [438-440] and share many of the attributes of fresh cadavers, such as the density and hardness of the tissue, maintaining similar visual attributes on ultrasound [441] compared to formalin-based cadavers [442-444]. Theil cadavers have been used heavily used in ultrasound research of regional analgesia [441, 445, 446], shear wave elastography [447, 448], Colour Doppler [449] and in therapeutic focused ultrasound [450].

There has been limited assessment of the acoustic properties of Theil cadaver, Joy at al [448] reported that quality of ultrasound images were representative with good tissue differentiation but noted progressively tissue stiffening over time, these results remained within range of those recorded in-vivo confirming that ultrasound images in Thiel cadaver retained life-like properties despite preservation.

The image quality of diagnostic medical ultrasound on cadavers has previously been assessed with mixed results. A study of imaging quality firm embalmed cadaveric subjects suggested that ultrasound imagery would be very poor due to the changes in tissue firmness in formalin cadavers that reduces the visual distinctiveness of organs and tissues [451]. This would completely negate any worthwhile ultrasound machine learning on firm embalmed cadaver as the difference in tissue density completely alter the visual and anatomical composition.

The visual properties of abdominal structures for Thiel cadavers was examined by Balta et al [452], who noted not only that tissue densities were very similar to that of fresh tissue, that there was a significant improvement in visualisation of the anatomical structure of the kidneys, and an overall improved ability to identify the left and right lobes of the liver. They also claimed that the aorta, common bile duct and portal vein were not visible under ultrasound in cadavers. It was therefore necessary to analyse the collected ultrasound imagery from the sampled Thiel cadavers, to ensure the target region of interest (ROI) contains the visual information required for a clinician to make a positive identification.

The Theil embalming method was formulated to reduce the amount of formaldehyde in the preservation of cadavers, maintaining a similar tissue density to that of a fresh cadaver while allowing for long term preservation and use. The process involves perfusing the cadavers via the carotid or femoral arteries with an injection solution (Table 6.1) which fixes and sterilises the cadaver before being immersed into a solution bath allowing for Theil cadavers to be preserved, and is occasionally washed with a maintenance solution allowing for the cadaver to be used for years while maintaining much of the tissue density and pliability of fresh cadavers.

Table 6.1 – Composition of Theil embalming solution used at the University of Dundee based off the work of W. Thiel [434, 435]

	Perfusion Solution	**Immersion Solution**	**Maintenance Solution**
Hot water	6.8L	1250L	20L
Boric acid	250g	45kg	600g
Ammonium nitrate	1680g	150kg	-
Potassium nitrate	420g	75kg	-
Sodium sulphate	700g	105kg	1kg
Propylene Glycol	2.5L	105kg	1L
Stock II (chlorocresol & glycol)	500mL	30L	200ml
Formaldehyde solution (8.9%)	2.1L	125L	-
Morpholine	150mL	-	-
Alcohol	1L	-	-
Total Volume	**13L**	**1720L**	**22L**

The phantom study in chapter 5 showed that infrared positional sensors can substantially improve classification of ultrasound cross sections from 89.7% to 93.3% an average improvement of ~3.6% using machine learning, but phantoms provide an idealised view of the abdomen. Rarely so they contain variations such as different thicknesses of tissues and fat, that can cause unexpected attenuation and shadowing. It was also not possible to fully test the patient normalisation requirements as the phantom abdominal cavity is of uniform size. In order to continue study of this method it was therefore important to examine the patient normalisation requirement of positional tracking coordinates both within the 3D space of the sensor area of detection and of the size of the abdomen.

A cadaver model was selected as the method to further study how positional data could be used to improve machine learning classification of abdominal cross sections. Using subjects from the Theil Cadaver facility at the University of Dundee [453], cadavers were scanned using an ultrasound probe with infrared sensor attachment which tracked the movement of the probe during the procedure. The use of cadavers for machine learning of ultrasound usually has a much more limited focus such as the estimation of nerve volume [454], regional anaesthesia [455, 456], vascular access [457], and guided biopsy [458].

6.2. Structure and Scope

This chapter examines six common ultrasound abdominal scans from eleven cadavers for the potential use in an image-based machine learning classification study of abdominal cross sections. While Theil cadavers provide a good stand in for human

subjects, as previously shown in the literature, some physiological changes do occur that effect the ease of collection. This chapter examines the suitability of Thiel cadaver as a machine learning dataset and tests the efficacy of the positional tracking system from chapter 5 on the cadaver data.

This chapter first seeks to highlight anomalies found in the collected Theil ultrasound image data and discusses potential workarounds to reduce variances that might cause confusion to a neural network attempting to classify the required cross sections. This also highlights a number of procedural issues and physiological conditions encountered during collection that may have affected image or classification quality and compares the results to those reported in the literature. Not all cross sections were visible in all eleven subjects, limiting the scope of this study. In cases where anatomy was not visible the probe positioning and movement was verified against the abdominal protocol and the cross section was collected regardless of visual markers, this has been documented in Table 6.3 to identify when this occurred.

Once the data has been analysed, this chapter builds upon the previous phantom study in chapter 5 by examining the classification and normalisation of infrared positional tracking using human cadavers. This chapter targets one of the limitations of the phantom study by adding variability to the cross-sectional images, expanding the sample size therefore reducing the likelihood of overfitting. This also adds variation to the size of the abdominal cavity therefore allowing for the testing of the patient normalisation algorithm within 3D space. This remains a limited pilot study but provides additional indication that normalisation to the size of the abdominal cavity improves classification

accuracy performance, whereas the previous study did not have the variation within the dataset to reliably do so.

6.3. Method

6.3.1. Image and Positional Data Collection

Collection was performed within the Theil Cadaver facility mortuary at the University of Dundee [453], using a SonixTouch Q+ medical ultrasound system (SonixTouch, BK Ultrasound, USA) with a curved array 5-2/60 ultrasound probe. Scans were collected by a trained sonographer with previous expertise in diagnostics and medical imaging interpretation to ensure the correct location and anatomical features were collected during the scan. Each scan was performed in sequence within a time frame of ~40 minutes for each subject, with scan sequences captured at a rate of five frames per second.

The ultrasound scans were performed on 11 soft body cadavers, preserved using the Thiel embalming system. The subjects consisted of 5 men and 6 women of Caucasian decent within an age range of 60 and 90. As part of the requirements for ethical approval, this was a blind study with neither the mortuary staff nor the sonographer having prior access to the medical records but criteria for acceptance was that the subject would have no visual signs of prior surgical intervention on the abdomen on examination. As part of the blinding process no identifying features are reported.

Six abdominal ultrasound cross sections were selected as regions of interest (Figure 6.1):

a) Right hypochondrium transverse approach for common bile duct.
b) Right intercostal approach sweeping through the liver to visualise the right portal vein.
c) Right hypochondrium longitudinal approach for the Gall Bladder.
d) Epigastric longitudinal approach sweeping through the aorta.
e) Transverse approach of the left kidney
f) Transverse approach of the right kidney

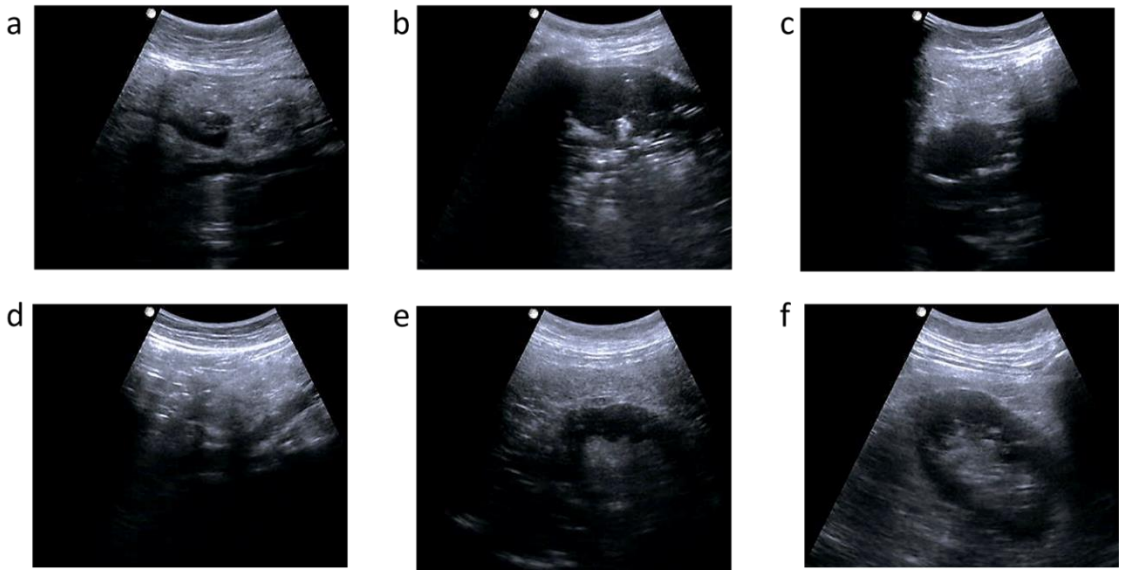


Figure 6.1 – Example ultrasound scans of the six cross sections of the abdomen from Theil cadaver.

These cross sections are all present on the Japanese abdominal ultrasound protocol and are identical to the one used in the phantom experiment in the previous chapter. They were specifically chosen due to the presence of overlapping regions of interest such as

with gallbladder and bile duct or due to the visual similarity such as with the left and right kidneys. While a supine or lateral decubitus approach could have improved visualisation of the kidneys, it was not possible to safely rotate, position and secure the cadaver in these positions without additional personnel, therefore the less optimal transverse approach was selected. A transverse approach remains valid as it is used in clinical practice, such as in trauma situations where it might not be possible to reposition the patient prior to scanning. Complex sweep scans of aorta and portal veins containing both visual similarities and overlapping anatomical structures were again chosen to provide added complexity to classification, the additional variation in abdomen size was also a factor in how much overlapping anatomy is seen in these cross sections adding additional complexity.

The cadavers could not be transported to the Vicon facility used in chapter 5 and as such this experiment uses the ASIC infrared positional sensor setup described in section 5.3.3.2. The HTC VIVE (3.0) tracker (as seen in Figure 5.5(a)) [418, 428] was wrapped in protective plastic with the sensors left exposed to prevent accidental contamination by viscera and affixed to the ultrasound probe using a strap and hot glue. A Steam VR base station (2.0) (as seen in Figure 5.5(b)) [429] was attached via a mounting strap to the ultrasound scanner which was positioned anteroinferior to the cadaver ensuring clear line of sight (as seen in Figure 6.2). The collection of positional and normalisation point data adhered to the method described in section 5.3.2.3, while it was not fully possible to test normalisation on the phantom, there was significant variation in the size of the

abdominal cavities of the cadaver, allowing for further testing of coordinate normalisation.

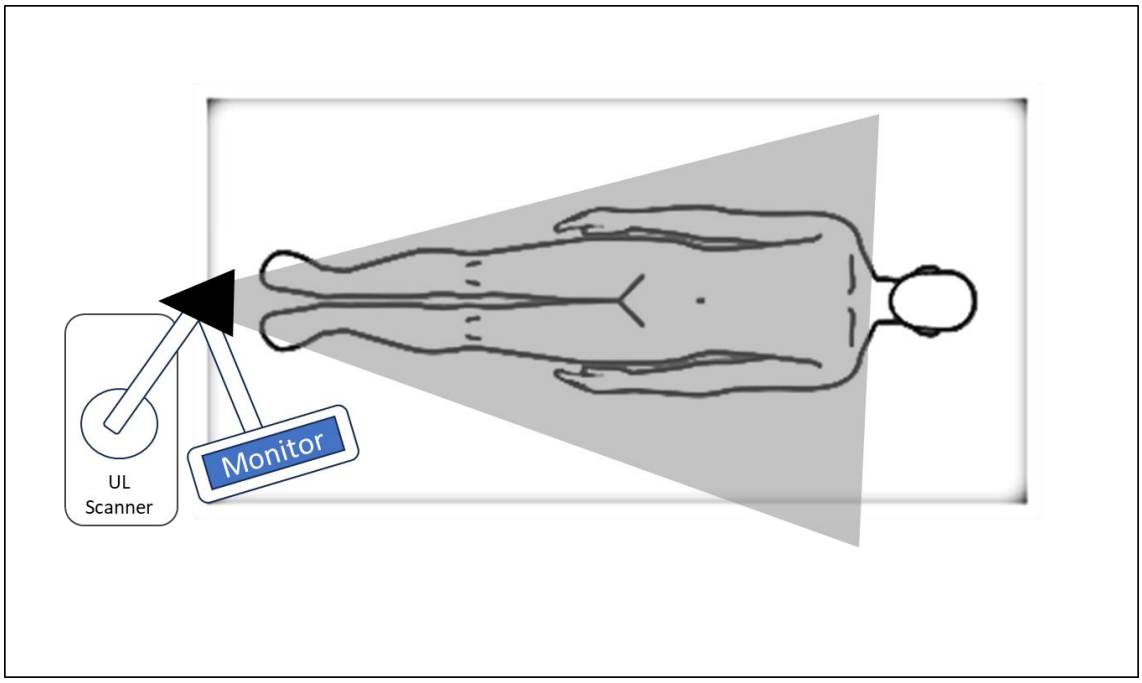


Figure 6.2 – Representation of equipment setup for collection of cadaver data during experiment at cadaver facility. The base station is shown mounted to the monitor arm of ultrasound scanner, which is positioned approximately on the midline, inferior to the subject.

6.3.2. Cadaver Analysis

The scoring and analysis of the cadaver scans was performed by a single subject expert several days after the collection session had taken place. Each set of scans was analysed within the context of the subject, and looked specifically for visual anomalies that might cause difficulties with classification of those cross sections by machine learning. The sonographer then rated these deviations as to how disruptive they would potentially be to machine learning classification.

6.3.3. Machine Learning

The machine learning methodology of this study is identical to that described in the described in section 5.3.3. Training was performed using the hyper parameters in Table 6.2, training was performed over a maximum of 10 epoch using early stopping [31] and a small batch size of 64 to promote better generalisation [32]. The CNN was trained 50 times for each of the five ground truth variations. The training and validation methodology was identical for both 3 and 4 channel versions of the network to ensure performance comparisons could be made. A holdout method was chosen over folded cross validation because the data must be split along patient sets and not every patient had examples of every classifier, this ensured that the training set was representative of the available data.

Table 6.2 - Neural Network Hyperparameters

Hyperparameters	Value
Activation Functions	SoftMax
Learning Rate	0.001
Training Iterations	50
Epochs	10
Optimiser	ADAM
Momentum	0.9
Dropout	0.5

6.4. Cadaver Study

6.4.1. Cadaver Results

The ultrasound images of the cadavers contain a number of anomalous findings that are significant enough to change the appearance of the anatomy being classified and therefore, affect the result of training and validation when attempting to apply machine learning. The ability to visualise the required anatomical structures as well as any notable deviations can be seen in Table 6.3. In cases where images of the organs and structures have not been successfully generated, this data was still used as the probe was correctly positioned on the abdomen.

Table 6.3 – Visibility results of cadaver ultrasound scans with additional information where key details are obscured.

Subject	Aorta	Gall Bladder	Bile Duct	Portal Vein	Left Kidney	Right Kidney
1	Yes (compressed)	Yes	No	Yes (shadows)	Yes (shadow)	Yes (shadow)
2	Yes	Yes	No	Yes (attenuated)	Yes (shadow)	Yes
3	Yes	Yes	No	Yes (shadow)	Yes	Yes
4	Yes (anomaly)	Yes	No	No (anomaly)	Yes	No
5	Yes (enlarged)	Not Visible	Yes	No (shadows)	Yes	Yes
6	Yes (compressed)	No	No	No	Yes	Yes
7	No	No	Yes	No (shadows)	No	No
8	Yes (compressed)	Yes (sludge)	No	Yes (attenuated)	Yes	Yes
9	Yes	Yes (collection)	Yes	Yes (attenuated)	Yes	Yes
10	No	No	Yes	No	Yes	No
11	No	No	No	No	Yes	Yes

Cadavers due their nature as diseased individuals and the processing required for preservation, often contain deviations from normal anatomical structures, such as degradation owing to the subjects advanced age, or injury from cause of death or disease process. The use of just 11 subjects means that variation drastically affected neural network quality.

6.4.1.1. Findings when scanning the Aorta.

The aorta is the main artery that carries blood from the heart and branches off to supply the blood of every major organ and structure throughout the body. There was a number of difficulties identified during collection including aortic compression owing to the lack of pulsatile volume, and physiological anomalies due to disease processes.

Aortic Compression

The Aorta is subject to compression by the sonographer during scanning, while in a living subject this would not pose a significant issue, the pressure in the aorta is within a cadaver is significantly lower as it is not being maintained by pulsatile flow. Therefore, it is possible to fully compress the aorta, making it difficult to visualise. The sonographer should avoid undue compressive force when performing an aortic scan on a cadaver especially where the chest is already sunken as was seen in subject one. This can be particularly difficult is cases where the subject has excess fluid in the abdominal cavity or in cases of obesity where pressure may be required to compress fluid or tissues away from the region of interest. Compression was not permanent, with the tissue returning back to its original state prior to compression, allowing for the sonographer to make several attempts at correctly capturing the cross section.

Figure 6.3 – Epigastric longitudinal ultrasound scan of the aorta, the edges are difficult to visualise due to excessive compressive force.

Aortic Anomaly

There is substantial deviation in the appearance of the aorta in subject four, potentially suggesting an abdominal aortic aneurysm. This should be considered when scanning cadavers for images of the aorta, as it will reduce the neural networks potential to recognise the typical aortic structure in subjects without anomalies. While the overall surrounding landmarks of the liver and overall shape remains the same, there is substantial deviation in shape of the aortic walls, that not only appear enlarged but have become ill defined on ultrasound, with the potential for an aortic dissection being present.

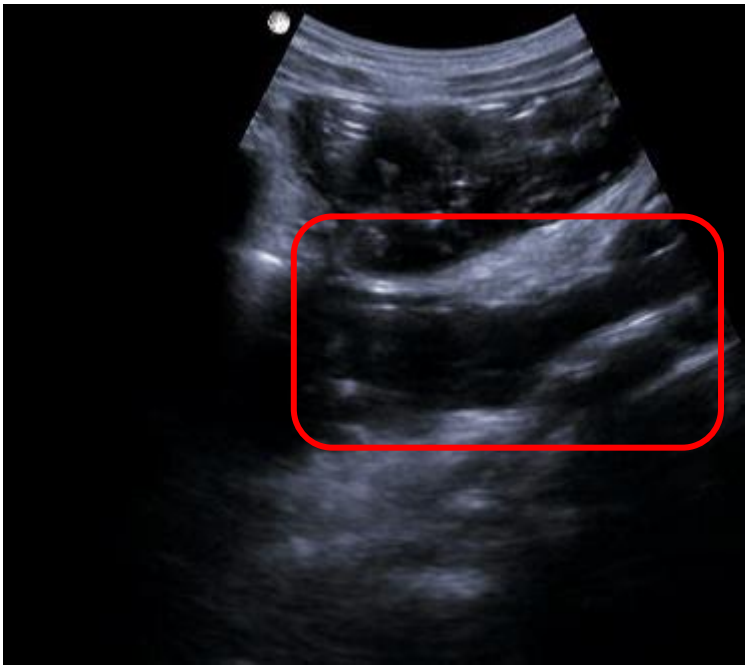


Figure 6.4 – Epigastric longitudinal ultrasound scan of the aorta with signs of anomalous dilatation of the aorta, the edges of the aorta appear of have sheared suggestive of aortic dissection.

Aortic Enlargement

Aortic enlargement can occur without anomalous deviation of the aortic wall appearing enlarged but well defined as in Figure 6.5. In this case the aorta has become substantially enlarged as we scan towards the heart. While the aorta is enlarged the edges remain well defined. While less likely to cause major difficulties in classification experiments, if should be noted where segmentation is being performed or in automated anomaly detection tasks.

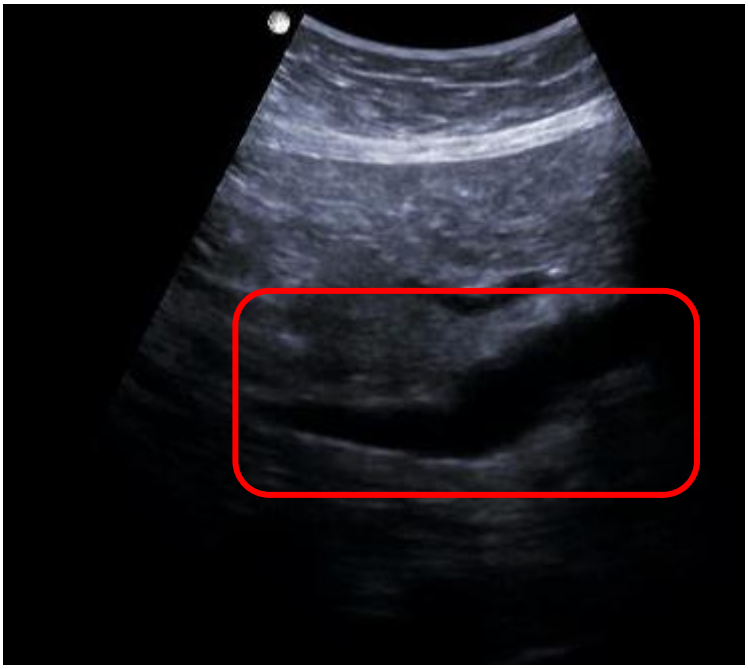


Figure 6.5 - Epigastric longitudinal ultrasound scan of the aorta. The aorta appears enlarged but the edges are clearly defined.

6.4.1.2. Gallbladder

The Gallbladder is a small pouch like organ that stores bile for subsequent use during digestion. It was not always possible to positively visualise the gallbladder, nor was it known if any subject had received a surgical intervention to remove it.

Cholelithiasis

The gall bladder can contain gall stones which may change the gallbladders appearance. In the case of subject nine (Figure 6.6), there was a substantial collection of gall stones within the gall bladder creating dense hyperechoic structures that cause acoustic shadowing.

Figure 6.6 – Longitudinal ultrasound scan of the right hypochondrium showing the gallbladder deformed with multiple gallstones.

Sludge in the Gallbladder

After a long period without secretion, bile within the biliary tract can thicken to a sludge-like consistency which can be visualised on ultrasound as a layered mass within the gallbladder that does not cast acoustic shadow as can be seen in Figure 6.7. This build up significantly alters the appearance of the gallbladder but was present in a number of subjects to a greater or lesser extent.

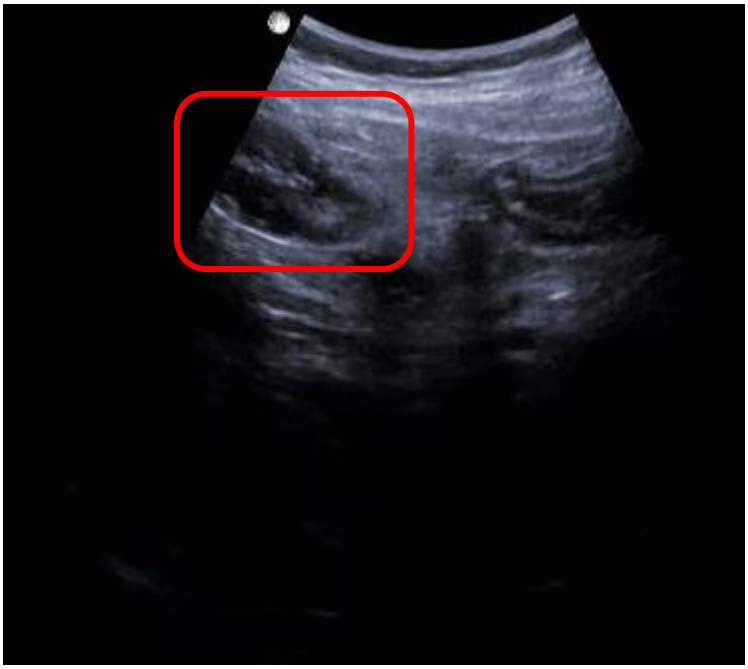


Figure 6.7 - Longitudinal ultrasound scan of the right hypochondrium showing bile that has solidified into a layered sludge inside the gallbladder.

6.4.1.3. Bile duct

As discussed in the literature [452], it is extremely difficult to visualise the bile duct in cadavers. Most subjects did not provide a clear visualisation, but the probe coordinates were saved despite the lack of visibility. The bile duct is only visible in 4 of 11 subjects and did not seem to be related to patient size or factors effecting visualisation of other cross sections as the bile duct could not be found in multiple smaller patients, such as subject 3, where all other cross sections could be visualised except the bile duct. In other subjects where there was difficulty visualising other anatomical structures the bile duct was clearly visible (Figure 6.8).

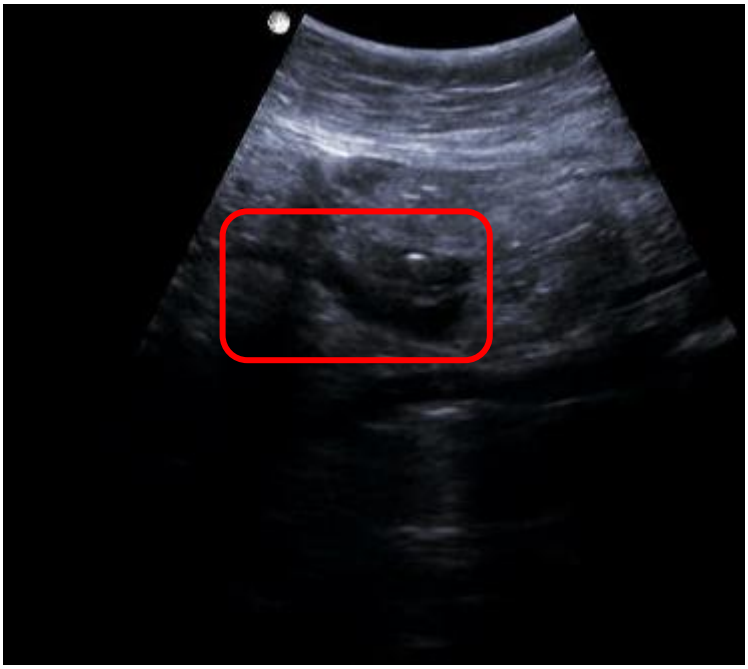


Figure 6.8 – Transverse ultrasound scan of the right hypochondrium showing a clear visualisation of the bile duct above the aorta.

6.4.1.4. Portal Vein

In order to capture the portal vein via the right intercostal approach, the sonographer must take the challenging approach of placing the ultrasound probe between the ribs and sweeping through the anatomical structures beneath. This approach is not always successful due to rib shadows, on assessment there were a number of anatomical anomalies such as an abscess and significant calcification within the subjects that also would affect visual recognition.

Rib Shadows

The portal vein is difficult to visualise due to the ribs. Where visualisation is difficult due to rib shadow, the sonographer is suggested to ask the patient to take a deep breath and hold it while the scan takes place, this expands the diaphragm and rib cage making it

easier to scan between the ribs [459]. This method is not possible in cadaver without specialist equipment to inflate the lungs which was not available at the time of scanning. While only limited visual information was often seen between the ribs, the coordinates of the probe position were still correct.

Figure 6.9 - Right intercostal ultrasound of the portal vein. The scan has been obstructed by the subject's ribs causing shadowing which has obscured the portal vein.

Hepatic Calcification

Dietary changes to the liver can cause the laying down of calcified deposits or steatosis [460] that are visible as abnormal brightness on ultrasound due to the change in hardness as seen in Figure 6.10. This could also be visualised as a diffuse attenuation that in some cases could completely obscure visual details of hepatic structures rendering it impossible to classify [461].

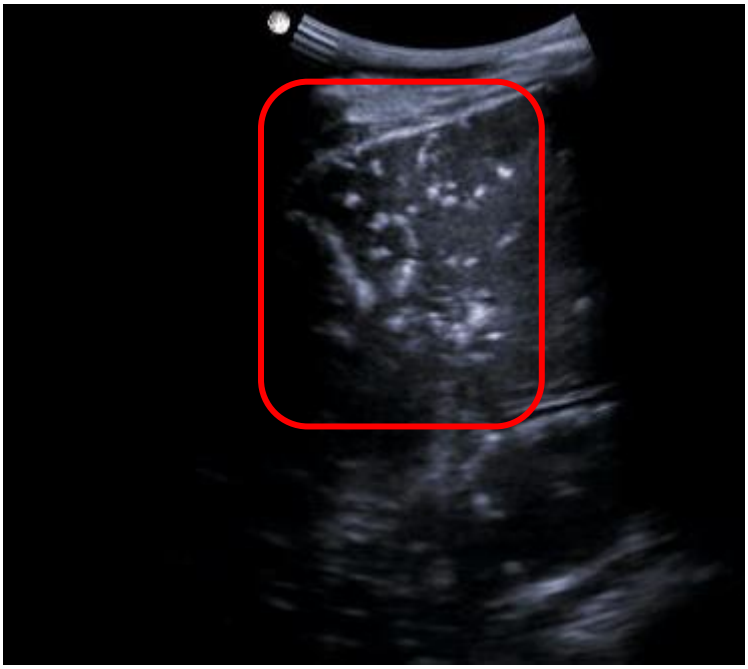


Figure 6.10 - Right intercostal ultrasound scan of the portal vein showing calcified liver steatosis. Severe steatosis is known to completely obscure the portal vein.

Hepatic Abscess

In one patient a large hepatic abscess completely obscures visualisation of the portal vein as seen in Figure 6.11. This apparent complex subhepatic collection of fluid also has deformed the abdominal position of the gallbladder, which required the probe to be moved caudally to visualise. The normal anatomical structures are either substantially displaced and deformed or entirely absent.

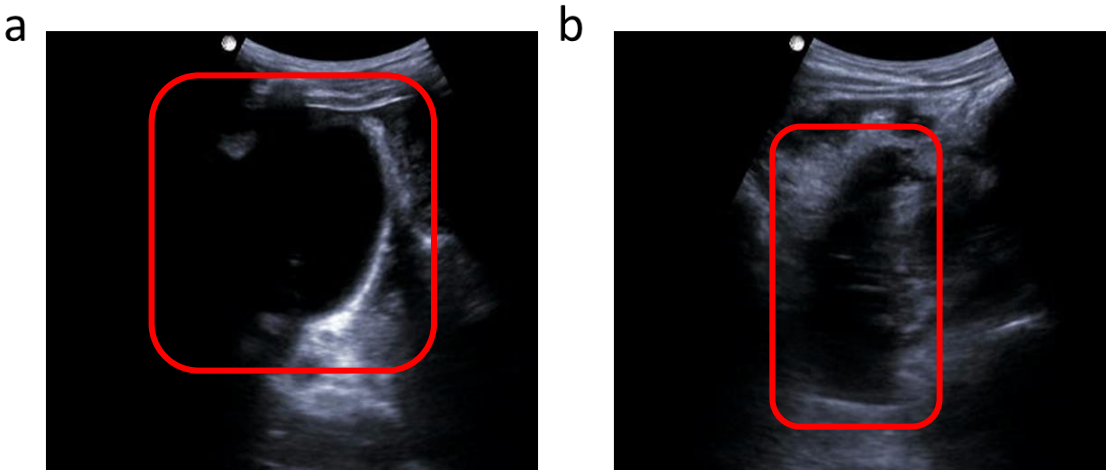


Figure 6.11 – a) Right intercostal ultrasound scan taken from an inferior angle showing a liver abscess completely obscuring visualisation of the portal vein. b) Longitudinal ultrasound scan of the right hypochondrium a liver abscess has deformed the region of interest shifting the position of the gallbladder

6.4.1.5. Left Kidney

The left kidney is located superior to the right, within proximity to the spleen, this makes it more difficult to visualise using a transverse approach [433]. As can be seen in Figure 6.12 a decubitus view provides a clearer image of the kidney but requires the subject to be placed on the side. While in living subjects this is a fairly trivial task, in Thiel cadaver there is not only substantial drainage when adopting this position but requires considerable assistance to turn and maintain the cadaver in this position for scanning.

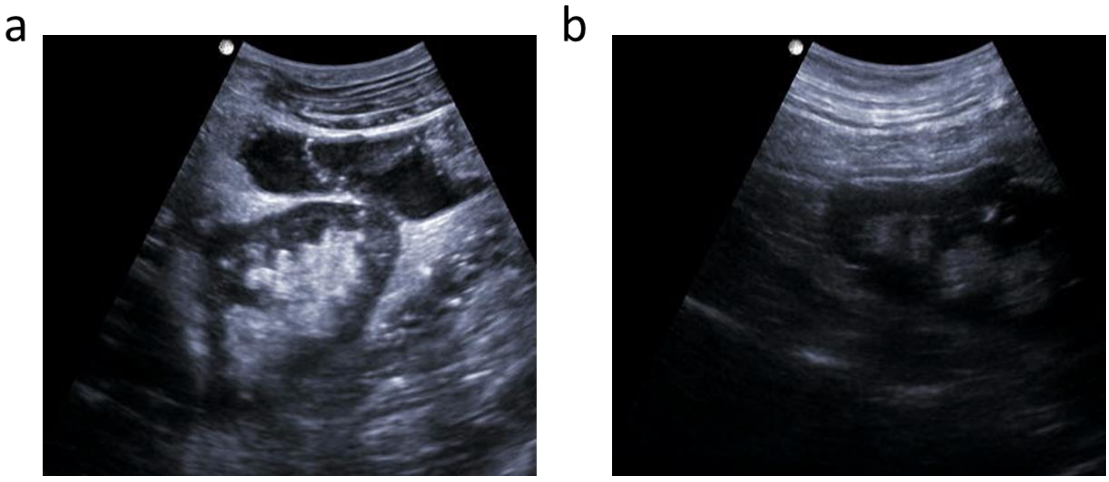


Figure 6.12 – Comparison of Ultrasound probe positions to visualise the kidney: a) Decubital view of left kidney b) Transverse view of the left kidney.

6.4.1.6. Right kidney

While in most subjects the right kidney was easy to visualise, there was a number of anomalous shadowing of the right kidney. In Figure 6.13, the entire abdominal cavity is dark. This shadowing did not improve with additional application of coupling gel or probe movement, it is likely to be a collection of free fluid in the abdominal cavity, but this does not fully explain the lack of detail. This is extremely abnormal as even with fluid in the abdomen, more detail should be visible suggesting another cause for the shadowing.

Figure 6.13 - Unexplained shadowing of right lumbar region and kidney. While sufficient coupling gel applied to the probe no further details could be made out within the lumbar region regardless of probe position.

These results show that image-only based training will encounter significant difficulty during training as there is a large amount of physiological variation due to the documented incidental findings. These difficulties stem from signs of morbidity and disease processes which are more common in older individuals and cadavers, as well as known capture difficulties that could not be easily improved in the same way as in a living subject. This could only be avoided by carefully analysing subject history, scanning more cadavers, and only selecting those cross sections with relatively normal anatomy for training.

6.4.2. Cadaver Discussion

The results of this study have highlighted a number of significant anomalous findings within the ultrasound cross sections of the eleven cadavers that would likely affect training due to the visual deviation caused by the anomaly. However, this does not completely prevent use of these subjects within a machine learning study. A three-step grading system was used when analysing the suitability of each subject (Table 6.4), scans marked as good were given a score of two, those marked moderate a score of one, scans with major anomalies or that did not capture the required anatomy and marked as failed were given a score of zero. Scores were classified as follows:

- Below 5 - poor suitability
- Score 6-7 – moderate suitability.
- Score 8+ - good suitability

Despite all cadavers experiencing some level of visual anomaly or capture failure, four cadavers achieved a good suitability score, with two receiving a moderate score. Although five of the subjects received poor suitability scores valid cross section captures were possible but overall, they did not represent a good use of clinical time as the overall quality of the image was lower than other subjects. The presence of these anomalies within the data will negatively affect potential training outcomes on what is already a small dataset.

Table 6.4 - Suitability of Theil cadaver subject for use in machine learning training for abdominal ultrasound.

Subject	Aorta	Gallbladder	Bile Duct	Portal Vein	Left Kidney	Right Kidney	Subject Suitability
1	minor anomaly	yes	Fail	shadow	shadow	shadow	Moderate (6)
2	Yes	yes	Fail	attenuated	shadow	yes	Good (8)
3	Yes	yes	Fail	shadow	yes	yes	Good (9)
4	major anomaly	yes	Fail	major anomaly	yes	major anomaly	Poor (4)
5	minor anomaly	fail	Yes	shadow	yes	yes	Good (8)
6	minor anomaly	fail	Fail	fail	yes	yes	Poor (5)
7	Fail	fail	Yes	shadow	fail	fail	Poor (3)
8	minor anomaly	minor anomaly	Fail	attenuated	yes	yes	Moderate (7)
9	Yes	minor anomaly	Yes	difficult to visualise	yes	yes	Good (9)
10	Fail	fail	Yes	fail	yes	fail	Poor (4)
11	Fail	fail	Fail	fail	yes	yes	Poor (4)

This study shows that there is a high degree of variability within Thiel cadavers, suggesting that while soft body cadavers can produce suitable imagery, when designing a cadaver-based machine learning study for medical imaging, an increased number of subjects should be scanned to take into account the potential likelihood of anomalies within even the most suitable subject.

6.4.3. Study Limitations

This limited study examines eleven cadavers of Caucasian decent within the elderly age range (70-90) in Scotland. The use of only 11 subjects can provide only a limited view of the potential variation possible within the cadaver population of any one area. While

most cadavers are elderly and therefore have more significant signs of morbidity and degradation, as seen in this study each subject has unique positive and negatives when examining for use in machine learning. The required region of interest and population of the cadaver should be carefully considered when designing a cadaver study to ensure that enough clear examples of the physiology are available to perform classification.

When the large amount of variation within the data set is considered, the size of the dataset is likely not enough to account for the additional variation within the image set. Substantially more subjects would be required to properly analyse image-only machine learning classification accuracy in cadaver. The abnormal anatomy within the scanned cadaver significantly disrupts the normal visual properties of the cross sections, allowing for increased repeatability of scans in the case of monitoring.

6.5. Machine Learning Study

6.5.1. Machine Learning Results

6.5.1.1. Image-only

Initial mean accuracy result for image-only training was 40.9% (Table 6.5), but by the harmonic mean (f-1 score) of just 0.37, this means that although the network correctly identified the specified cross section, it did so by favouring that classification not through actual feature recognition. When accuracy results that achieved a harmonic mean (f-1 score) of below 0.5 are excluded the average accuracy result drops to 22.1% which is just 5.1% above random chance. These results are of insufficient value for a comparative study of IR tracking and therefore are excluded, a significantly larger

sample size with less variation in physiology would be required in order for a full comparison study to be performed.

Table 6.5 - Image-only neural network accuracy showing original and corrected harmonic mean f-1 results.

Cross Section	Raw Accuracy	Raw F-1 Score	Corrected Accuracy	Corrected F-1 Score
Left Kidney	15.4%	0.07	1.5%	0.50
Right Kidney	33.1%	0.22	3.7%	0.58
Aorta	42.1%	0.31	6.5%	0.57
Bile Duct	52.7%	0.60	41.1%	0.70
Gall Bladder	50.4%	0.48	34.6%	0.65
Portal Vein	51.9%	0.57	44.9%	0.59
Average	40.9%	0.37	22.1%	0.65

6.5.1.2. IR positional results

As can be seen in *Table 6.6*, the use of Infrared positional sensing in classification by neural network significantly augmented cross section recognition, when the average accuracy of the trained networks is taken into account a classification accuracy of 88.3% was achieved for one point normalisation, 91.5% for two-point normalisation and 92.8 for 3-point normalisation. Where no normalisation was performed, the network achieved an average accuracy of 82.2%, this means that three-point normalisation achieved results 10.7% higher than that of the average with no normalisation. Each additional normalisation point further reduces the variation in the positional coordinates provided to the CNN increasing the accuracy of the classification through reduced error.

Table 6.6 - Mean normalisation accuracy for classification from 200 neural networks

Cross Section	No Normalisation	1-Point	2-Point	3-Point
Left Kidney	92.9%	96.6%	96.4%	97.1%
Right Kidney	94.9%	98.3%	98.0%	99.2%
Aorta	91.0%	88.5%	92.4%	98.3%
Bile Duct	55.4%	76.0%	82.7%	83.0%
Gall Bladder	64.5%	80.3%	84.2%	84.0%
Portal Vein	94.4%	89.9%	95.4%	95.4%
Average	**82.2%**	**88.3%**	**91.5%**	**92.8%**
Improvement	-	6.1%	9.3%	10.7%

Studying the training variance for classification of the abdominal cross sections (Figure 6.14), shows a clear trend, where additional points of normalisation train neural networks with variation in accuracy result. On average there is an improvement to classification accuracy with the use of an additional point of normalisation with 3-point normalisation being the most accurate with the least variance, this is especially obvious for classification of sweeps of the aorta, despite the potential for visually similar or overlapping image information.

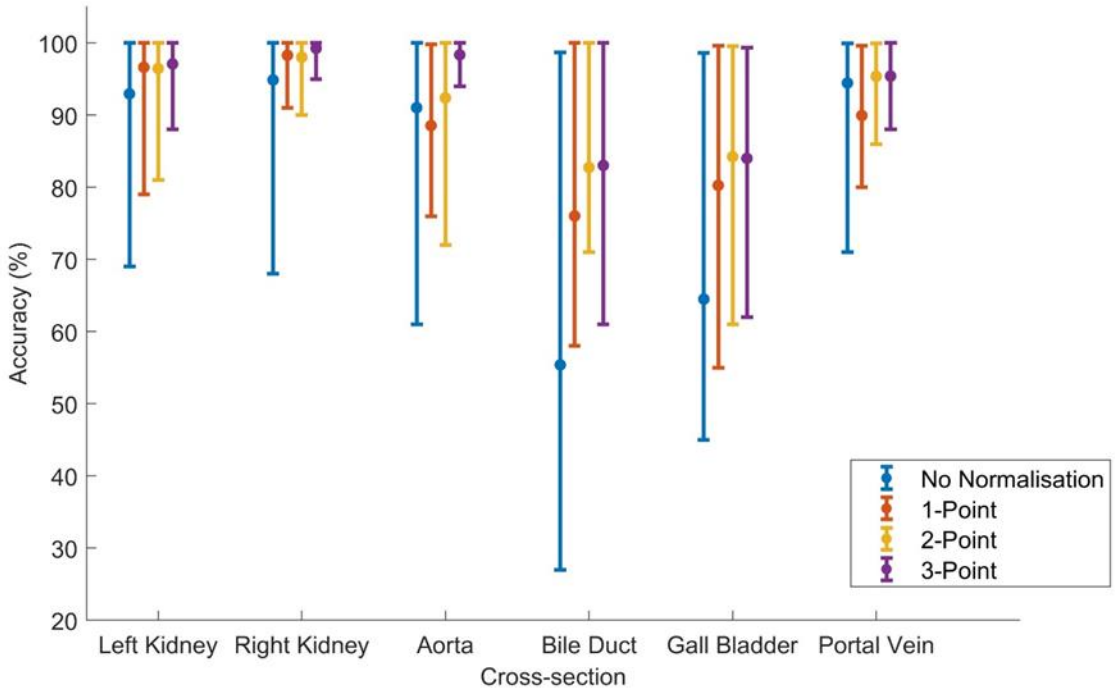


Figure 6.14 - Comparison of the mean classification accuracy of networks trained using ultrasound dataset augmented with positional tracking data. Results show both networks with no normalisation as well as where coordinates have been normalised using 1, 2 & 3 points. Error bars represent the deviation in classification accuracy for each cross section over 250 neural networks.

When examining the networks that achieved the highest accuracy result. No normalisation achieved 88.5%, with 93.5% for 1-point normalisation, 95.5% for 2-point normalisation, 96.8% for 3-point normalisation. As can be seen in Figure 6.15, the largest classification error was in recognition of bile duct and gallbladder regardless of normalisation level. A notable error in classification is that a small number of images of the left kidney were often misclassified as bile duct, these scans do contain incidental captures of the right kidney but with surrounding structures being that of the liver and not the spleen as would be the case with the left kidney.

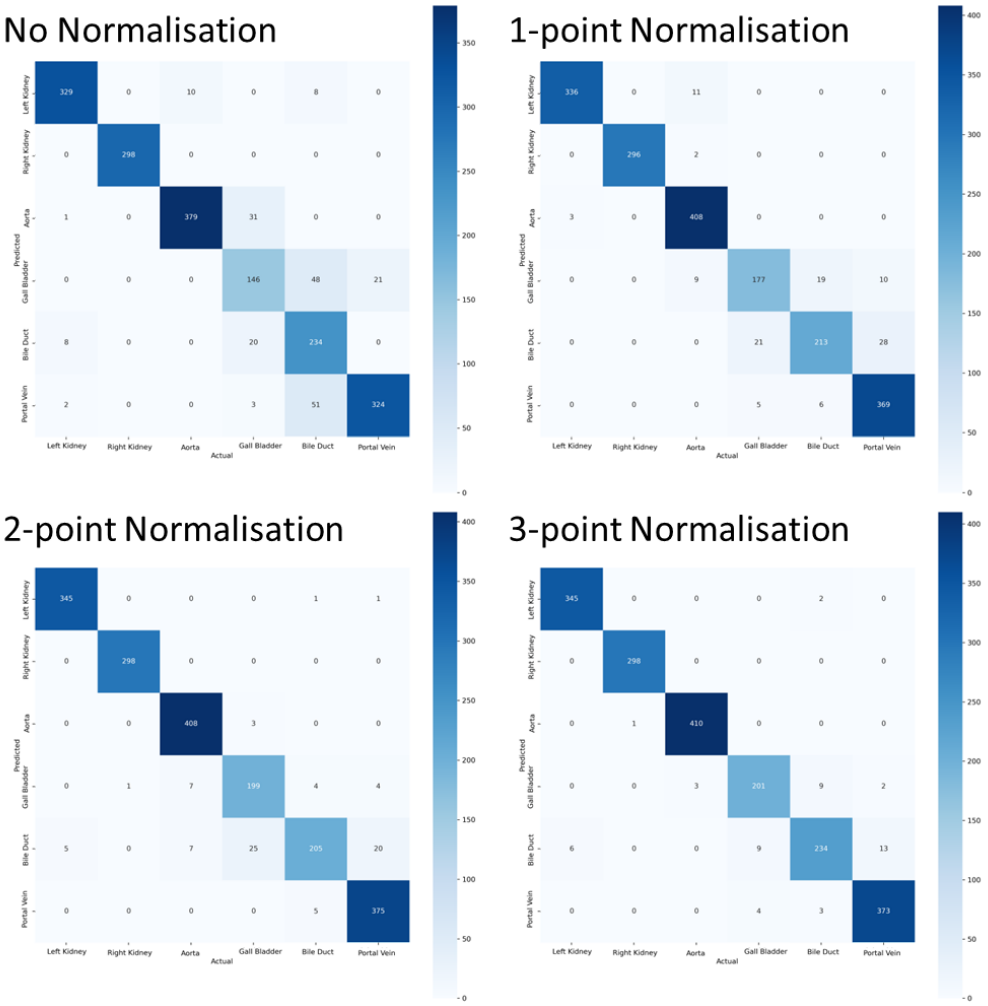


Figure 6.15 - Side by side comparison of confusion matrix examining the effects of normalisation of positional tracking coordinates on classification accuracy.

Unlike the image-only networks, the harmonic mean (F-1 Score) for the positional neural networks as seen in Table 6.7 show that these networks have achieved a robust training result in both precision and recall despite the confounding factors within the image data. No additional adjustment is required for comparative analysis as the produced results are valid.

Table 6.7 – Harmonic mean f1 score for best performing positional neural networks.

Cross Section	No Normalisation	1-Point Normalisation	2-Point Normalisation	3-Point Normalisation
Left Kidney	0.96	0.98	0.99	0.99
Right Kidney	1	1	1	1
Aorta	0.95	0.97	0.98	1
Bile Duct	0.7	0.85	0.9	0.94
Gall Bladder	0.78	0.85	0.86	0.92
Portal Vein	0.89	0.94	0.96	0.97
Average	**0.88**	**0.93**	**0.95**	**0.97**

6.5.2. Machine Learning Discussion

This study further builds upon the previous phantom study by demonstrating the use of IR positional sensor data to improve classification of ultrasound for abdominal ultrasound cross sections in a cadaver study. Classification accuracy using IR positional sensor data with no normalisation on cadavers was 82.2% compared to 92.7%, one point normalisation was 88.3%, two-point normalisation achieved 91.5%, three-point accuracy was 92.8%. While these are lower accuracy results than the phantom, a greater number of subjects were used, introducing additional variables into the classification task such as the difference in abdominal size in the cadaver vs the fixed size in the phantom. The highest accuracy network was a 3-point normalised network that achieved and accuracy

of 96.8% in comparison to 98.8% in the phantom study achieved by a network using single point normalisation.

6.5.2.1. Normalisation

The previous phantom study was unable to fully test the normalisation requirements of the algorithm, the variation in the abdominal cavity size of the cadavers has provided a significant test bed for this study. When examining the standard deviation for the variance in the positional coordinates when normalisation is applied (as seen in Figure 6.16), each additional point of normalisation reduces the size of variance in the coordinates, allowing for better training and validation as the network does have to learn as wide a number of coordinates. There was a greater initial reduction in variance in the x and y axis with 1-point of normalisation, with smaller improvements with subsequent additional points of normalisation. The z axis achieved a greater improvement with 2-points of normalisation but very little subsequent improvement with 3-points. The variation in the rotational axis [a, b, c] was less than 2mm and was not affected by the number of normalisation points. This is due to the fact that the relative angle of the IR sensor attached to the probe to the base station remained the same throughout the collection session (as seen in Figure 6.2), while there was some variation in subject size and position, the probe remained well within expected normal range for angle data.

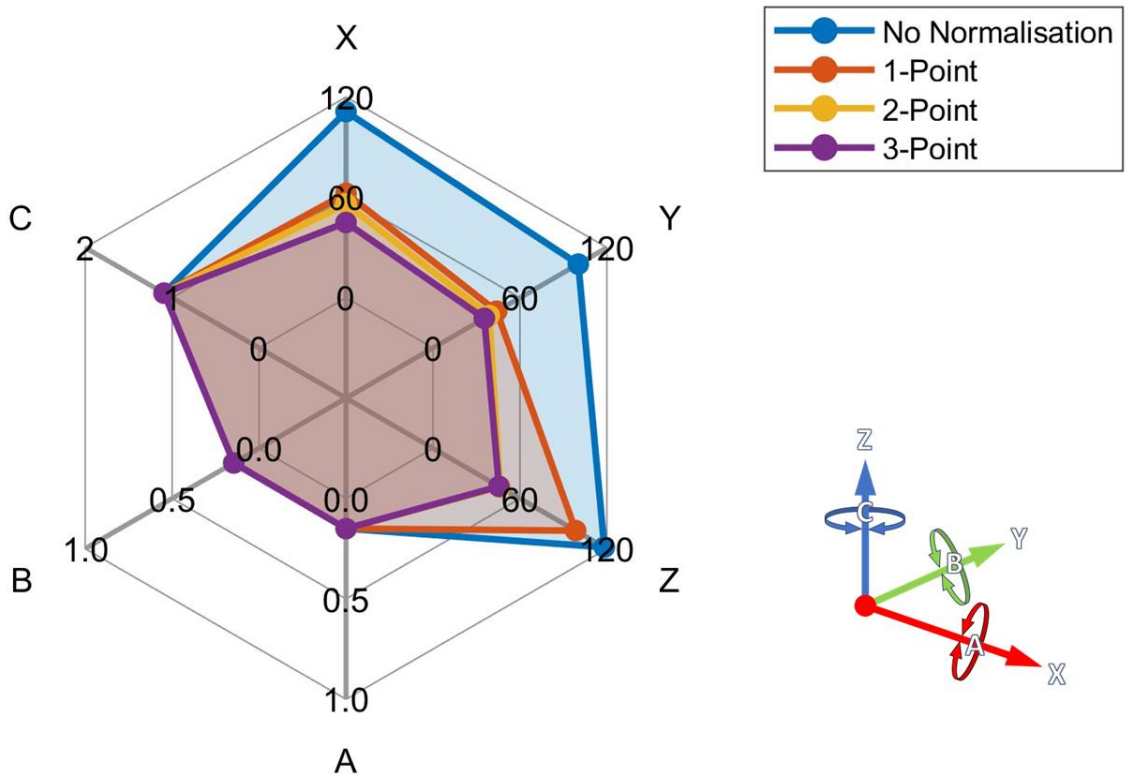


Figure 6.16 – Spider chart visualising the differences in standard deviation caused by normalization of positional coordinates for each individual axis within the ultrasound dataset. Results are displayed in millimetres.

It is possible to see the impact of normalisation by mapping the coordinates in a scatter graph (Figure 6.17), when the data is segmented into distinct areas that correspond to the cross sections a clear trend is visible. Each additional point of normalisation makes the coordinate grouping more distinct and creates a tighter fit, which would allow for neural networks to better generalise on the positional data even when using a smaller data set such as the one presented in this chapter.

Coordinates with no normalisation are subject to two sources of variance: the abdominal cavity size and the position within the IR sensor field. Use of raw coordinate data

increases overall accuracy in comparison to image-only. This method requires careful placement of sensor equipment in relation to the patient to reduce variance in the coordinate data and potentially would require a much larger dataset to improve neural network accuracy to account for the additional sources of variance.

The use of 1-point normalisation allowed for the positional coordinates to be fitted into the same feature space which resulted in a 6.1% improvement to average accuracy. This still does not account for differences in patient anatomy and is therefore subject to variation caused by differences in abdominal size. This is potentially why 1-point normalisation underperformed in the aorta and portal vein scans, as these are moving sweeps scan that would require accurate coordinates on multiple axis.

Applying 2-points of normalisation sets an origin point and reduces coordinate variance between patients by controlling for variation in the width of the abdomen. This allowed for a significant improvement in average classification accuracy of 9.3%. Accounting for width was particularly important in improving accuracy in cross sections that require probe movement such as sweeps of the aorta, portal vein, gallbladder and bile duct.

Normalisation using 3-points, scales the coordinates to three fixed points on the abdomen, effectively making the abdominal cavity a set size. Reducing variance in abdominal height and width provided only a small 1.4% improvement to average accuracy in comparison to two points of normalisation but significantly improved training variance in all but the bile duct classification training results.

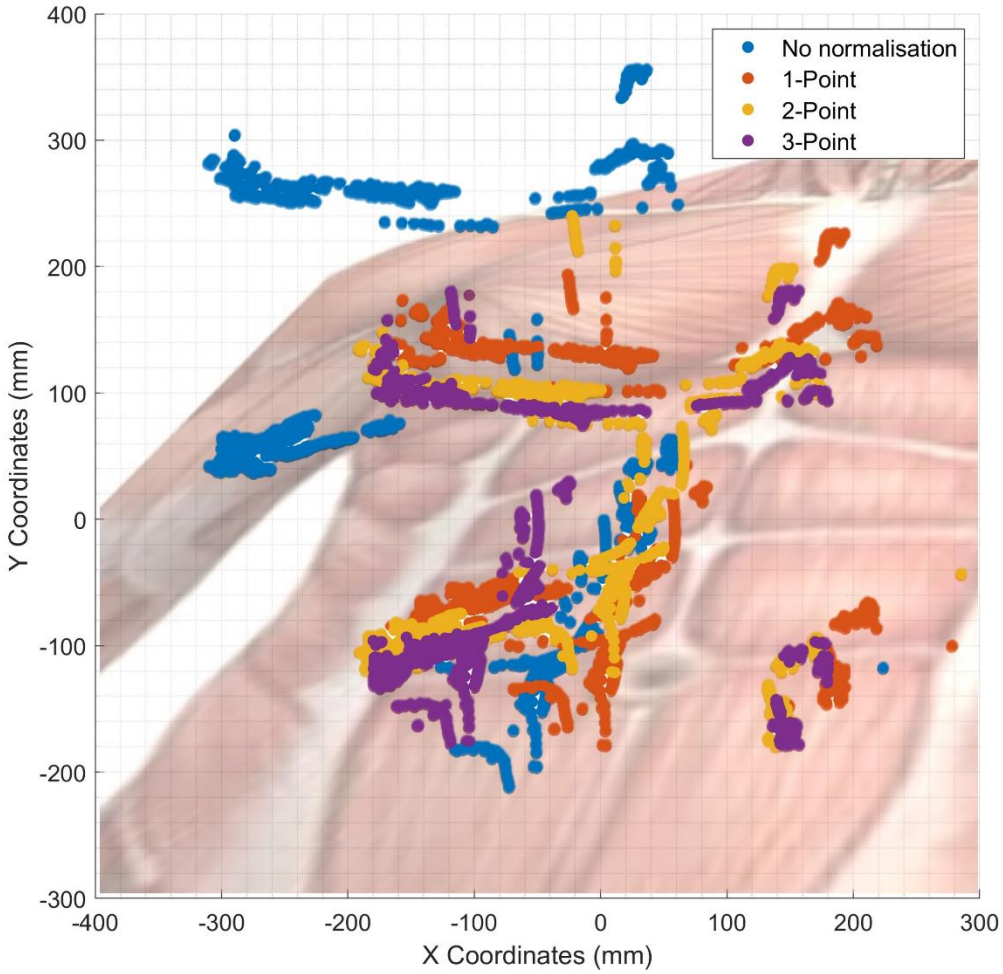


Figure 6.17 – Visual representation of the point cloud of the ultrasound probe X, Y angle during cross sectional capture. The effect of normalization on these coordinates has been transposed onto an anatomical representation of the human body.

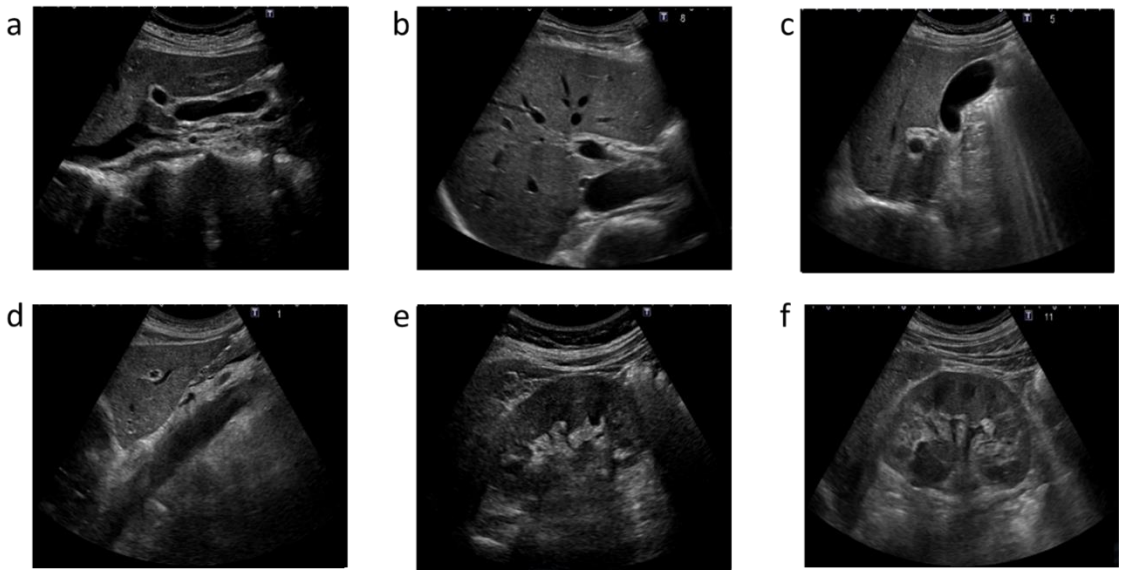
6.5.2.2. Study Limitations

The use of cadavers has expanded the sample size substantially, but as reported in section 4.1, it has also introduced substantial variation into the image set causing image-only recognition accuracy to collapse, as seen in the harmonic mean results in *Table 6.6*. The use of only 11 subjects is not enough to account for the additional variation within

the image set and therefore should be excluded from study. Substantially more subjects would be required to properly analyse image-only accuracy in cadaver, in comparison to harmonic mean in positional trained networks (Table 6.7) which shows stable precision and recall despite the poor image-only results. This dataset is not yet large enough, to use the power curve method described in chapter 4 to estimate the required dataset size, nor would it be possible to accurately estimate the number of cadavers required due to the significant amount of physiological variation present in cadavers. It is not possible to estimate This suggests that positional information could be used to assist in the positioning of the ultrasound probe even in cases where abnormal anatomy significantly disrupts the normal visual properties of the cross section, allowing for increased repeatability of scans in the case of monitoring.

6.5.2.3. Comparative Study

While the image-only results from the cadavers should be excluded, these cross sections use the same clinical protocol as the previous study in Chapter 3 and therefore results from the cadaver study can be compared to networks resulting from the transfer learning methodology. When comparing the six chosen cross sections (as seen in Figure 6.18) in the top performing inception-based architectures from the baseline study to those of the cadaver positional IR study (in Table 6.8), there is more than a 10% improvement in the average classification from 85.72% and 84.52% in the inception networks to 96.83% for a network trained with 3-points of normalisation.



a) *Bile duct. b) Portal vein. c) Gallbladder.* d) *Aorta.* e) *Left kidney*. f) *Right kidney*

Figure 6.18 - Examples of ultrasound scans of cross sections captured on the Canon Aplio i800 high resolution ultrasound scanner.

The trend of Bile duct and Gall bladder being the weakest classification can be seen in both inception and cadaver network accuracy results with Gallbladder and Bile duct being joint lowest in the inception networks, at 71% and 79% respectively with this trend is also present in the cadaver results. There was a substantial improvement to classification of left and right kidney in positional networks, even non-calibrated networks augmented achieved only 3% error, substantially better than that achieved with image-only classification. Classification of the aorta was lower in positional networks, but it should also be noted that training was performed on a significantly smaller set of data with visualisation of the aorta much more difficult in the cadaver dataset.

Table 6.8 - Comparison of accuracy of transfer learnt networks trained on the Canon dataset and the positional trained networks trained on Cadaver with positional sensor data.

Dataset	Canon (image-only)		Theil (image and positional)			
Model Type	**GoogLeNet**	**InceptionV3**	**No Normalisation**	**1-Point Normalisation**	**2-Point Normalisation**	**3-Point Normalisation**
Left Kidney	93%	86%	97%	99%	99%	98%
Right Kidney	79%	79%	100%	100%	100%	100%
Aorta	100%	100%	97%	95%	97%	99%
Bile Duct	71%	79%	73%	87%	88%	94%
Gall Bladder	71%	79%	69%	89%	95%	94%
Portal Vein	100%	86%	94%	91%	94%	96%
	85.72%	**84.52%**	**88.33%**	**93.50%**	**95.50%**	**96.83%**

The results from this cadaver study show that there is a clear improvement in classification accuracy of these difficult cross sections when augmenting the dataset using positional information even when compared to image-only networks trained on a much larger scale dataset. These results also are in line with those seen in the phantom study with a 3-point calibrated network achieving an accuracy of 96.8%. The accuracy of these networks suggests this method warrants additional exploration in future.

6.6. Conclusion

Cadavers are a major source of medical research data and could potentially be extremely useful in machine learning medical imaging research. Theil cadavers are a potential

source of imaging data for machine learning but contain significant visual variation within the examined cross-sectional data. The size of collection must take into account the increased amount of visual variation within the subject matter.

While increasing the number of subjects in future studies should not be ruled out. These studies should target more specific machine learning subject areas, such as anomaly detection, where the disease process is known, and therefore the medical history of the cadaver could be an advantage. It would also be easier to perform quality control with a smaller region of interest that could account for the significant variation seen in just eleven subjects.

This cadaver study further underlines the effectiveness of sensor-based positional and coordinate information for localising and classifying ultrasound cross sections. In providing an additional source of data, these networks are significantly more accurate than their image-only counterparts, reducing the misclassification of cross sections where there is visual similarity or shared anatomical features. Positional tracking is a-kin to the sonographer remembering the position of their hand during the scan. This additional information completely changes how the neural network understands the provided data by adding additional frame of reference, not just for the individual cross sections but also in relation to one another. This result is especially significant as it required limited additional investment in ground truth development, whereas improving an image-only approaches would require significant additional investment to both increase dataset size and potentially to increase the complexity of the annotation.

Further study of positional tracking requires additional testing using significantly larger sample sizes to further test method to combine image recognition and positional normalisation. Additional methods of probe tracking should be explored, especially those not requiring line of sight such as electromagnetic tracking, a method already used in 3D ultrasound imaging, as this would simplify the tracking procedure. Further normalisation points should be tested to correctly size the abdomen and ensure the most effective normalisation point has been used. There is also potential to apply ensemble learning in future iterations using teacher models on clinical ultrasound datasets to augment the smaller cadaver dataset size.

The use of positional tracking in ultrasound has many potential future clinical applications. At its most simplistic, this system will allow a neural network to effectively detect and classify abdominal cross sections within a scan with reduced error. This could be used to automatically select the best examples of each cross section as the sonographer sweeps the probe over the target region of interest, greatly reducing the time required to perform each scan. This could also be used in the training of new sonographers, with on screen guidance being provided to improve or correct the clinicians positioning using both coordinates and image as a guide. Novice operators could be guided using ‘imageless’ imaging devices with gamified user interfaces with the algorithm feeding backs probe guidance information ensuring that the correct images are collected with no specialist knowledge required from the operator. This same technology potentially could drive a robotic arm, allowing for fully automated collection.

Chapter 7

Conclusion and Future Work

7.1. Research Conclusions

Diagnostic medical imaging has become a crucial component of the diagnostic process, offering clinicians invaluable evidential support for confirming differential diagnoses. Among these imaging modalities, medical ultrasound stands out as one with the most potential for worldwide increase in uptake. Its appeal lies in its safety, portability, real-time imaging capabilities, and cost-effectiveness. However, a significant impediment to further progress in this modality is the heavy reliance on operator expertise. Currently, clinical diagnostics places a substantial burden on increasingly busy clinicians, who must manually perform every scan in order to produce high-quality images. The work presented in this thesis examined the practicalities of machine learning neural networks to assist in the collection and classification of abdominal ultrasound cross sections and was successful in achieving the primary objectives defined in the industrial partners project specification:

1. Identify what planes are acquired in an image – automatic annotation and adherence to guidelines.
2. Automatically identify the correct planes during a continuous sweep – simplification of screening procedure.
3. Improve classification of hard to differentiate edge cases within the Japanese abdominal protocol.

This was first tested as part of a proof of concept on an abdominal phantom and then expanded upon to test calibration on the abdominal cavity within the context of a cadaver study. Ultimately, this work provides a new method with the potential to greatly improve the accuracy of automation processes of numerous routine ultrasound tasks and lower the skill threshold required to capture clinically relevant medical cross sections, thereby enhancing the efficiency and accuracy of medical imaging.

Chapter 2 reviewed the literature in three major areas:

- The increasing use of medical imaging in clinical practice examining the fundamentals of ultrasound and alternative modalities.
- Contextual anatomical and physiological reasoning behind collection of the Japanese Abdominal Ultrasound protocol.
- An overview of foundational machine learning and computer vision techniques and history that underpin this project.

Medical imaging has become increasingly indispensable in clinical practice, offering a dependable means of diagnosis while reducing reliance on subjective clinical judgments

and minimising diagnostic errors. However, the surging demand for imaging, particularly ultrasound and CT scans, has strained healthcare systems, resulting in prolonged wait times for patients. This chapter evaluates the advantages and challenges associated with different imaging modalities. Ultrasound, with its non-invasive nature, provides real-time visualisation but faces challenges like operator reliance and image artifacts. In contrast, X-rays and CT scans employ ionising radiation, which raises concerns about cumulative exposure. Magnetic Resonance Imaging (MRI) is a safer alternative, despite limitations related to scan duration and equipment. The Japanese abdominal ultrasound protocol encompasses 16 cross-sectional views and is a critical tool for diagnosing and monitoring abdominal conditions. It enables non-invasive assessment of vital organs such as the aorta, liver, kidneys, gallbladder, spleen, and pancreas. Specific conditions like aneurysms, liver diseases, gallbladder issues, renal problems, splenic abnormalities, and pancreatic disorders can be detected using this protocol. Ultrasound's safety, portability, and affordability make it invaluable for early detection and monitoring. Machine learning is set to revolutionise healthcare by enabling computers to learn and predict without explicit programming. Its evolution, from early methodologies to deep learning, is explored. Data processing techniques such as image enhancement, standardisation, feature extraction, and segmentation were pivotal in enhancing and standardising medical imaging data for machine learning purposes. Training methodologies are explained including supervised, unsupervised, reinforcement-based, and transfer learning. Finally, this chapter acknowledges criticisms and challenges of machine learning, including concerns about poor practice, bias, and the need for practical improvements in clinical problem-solving.

Chapter 3 examined the effectiveness of nine neural networks utilising transfer learning on a dataset comprising sixteen abdominal ultrasound cross sections from sixty-four patient sets. The primary objective was to establish this baseline response accurately. GoogLeNet and InceptionV3 achieved the highest validation accuracy at 83.9% through transfer learning on a sample set of 26,294 images. InceptionV3 exhibited a top-2 accuracy of 95.1%. For a smaller sample set of 800 images, Alexnet secured the highest accuracy at 79.5% (with a top-2 accuracy of 91.5%). The evaluation of these neural networks allowed the identification of challenging cross sections and edge cases that confounded traditional image only classification, such as between right and left kidneys. A case study involving mobile and small-sized networks, demonstrated the effectiveness of compact networks in ultrasound classification. This chapter extends the findings of previous studies in the literature, highlighting the accuracy potential of various neural network architectures in classifying standard abdominal cross sections. The depth of neural networks had only a marginal impact on classification accuracy, with a 2.2% difference between the top-performing networks among the nine tested. Dataset size emerged as a pivotal factor, indicating that more complex neural networks excel with larger datasets, while simpler linear networks outshine others in smaller datasets.

Chapter 4 addressed the challenge of data acquisition in the absence of additional support from the industrial partner. This chapter introduces a biphasic framework designed to evaluate the cost of data collection by iteratively predicting accuracy concerning sample size. This framework incorporates active learning techniques to guide and optimise human annotation, specifically tailored for machine learning applications

in medical ultrasound imaging. The chapter highlights the potential cost reduction through the use of publicly available breast, foetal, and lung ultrasound datasets, with a focus on the practical case study of breast ultrasound data. This study revealed a correlation between dataset size and ultimate accuracy, resembling patterns observed in clinical trials. Substantial improvements in accuracy are achievable with just 40-50% of the data, depending on the applied tolerance metric. The integration of active learning further reduces the need for manual annotation, resulting in a noteworthy cost reduction of approximately 66%, while maintaining a permissible accuracy deviation of around 4% from theoretical maximums. The significance of this work lies in its ability to quantify the additional data and annotation required to achieve specific research objectives. These methods align with the understanding of clinical funders, providing an effective framework for feasibility and pilot studies with fixed budgets, optimising predictive gains, and informing resource allocation for further clinical studies.

Chapter 5 documented a proof-of-concept study where positional tracking information was introduced as an additional element into the neural network's input to provide the necessary context for recognising these otherwise complex edge case abdominal cross sections. Previous studies showed that distinguishing between multiple liver or left and right kidney cross sections based solely on images can be challenging. This chapter explores the utilisation of optical infrared (IR) and sensor-based infrared tracking to monitor the position of an ultrasound probe while collecting clinical cross sections on an abdominal phantom. Convolutional neural networks were trained using both image-only and image with positional data inputs, with the results comparing their classification

accuracy. The incorporation of positional information led to a substantial enhancement in average classification results, elevating accuracy from approximately 90% for image-only to 95% with optical IR position tracking and 93% with sensor-based IR for six common abdominal cross sections. The application of low-cost positional tracking for machine learning-based ultrasound classification, not only promises increased accuracy in identifying critical diagnostic cross sections, but also holds the potential to validate protocol adherence and provide navigational prompts. This will greatly users in capturing cross sections more effectively in the future. These results were limited by the use of a single phantom which led to overfitting, meaning that results are indicative only.

Chapter 6 discussed the suitability of cadavers for an abdominal image-based ultrasound cross-sectional study, an important step in building upon a previous study involving phantom data. In the previous study, the addition of infrared positional information improved machine learning classification accuracy by 4.3%, but overfitting due to a single subject and not being able to fully test calibration requirements limited the resultant outcome. To address these limitations, this chapter explores the use of cadavers, which offer more variability in anatomy and abdominal cavity size. The study collected abdominal scans and calibration points from eleven cadavers using an ultrasound probe. The results of the study revealed several challenges in using ultrasound imaging of cadavers for training machine learning models to recognise anatomical structures. Cadavers, being diseased individuals, often exhibit deviations from normal anatomical structures. These variations can result from advanced age, injuries, or underlying diseases. These deviations need to be quantified, especially when

working with a small sample size. The aorta can be fully compressed during scanning in cadavers due to the absence of pulsatile flow, making visualisation challenging. Some cadaver subjects exhibited significant deviations in the appearance of the aorta, potentially indicating conditions like abdominal aortic aneurysms. Aortic enlargement, without anomalous deviation of the aortic wall, was observed in some cases. It was not always possible to visualise the gallbladder in cadaver subjects. Gallbladders can contain gallstones, which can affect their appearance and cause acoustic shadowing. Bile sludge within the gallbladder can alter its appearance without causing acoustic shadowing. Visualising the bile duct in cadavers is extremely difficult, with the bile duct being visible in only a subset of subjects. The portal vein is challenging to visualise due to rib shadows, and expanding the rib cage by asking the patient to take a deep breath is not feasible in cadavers. Calcified deposits in the liver can affect visual details and render some structures unrecognisable. Large hepatic abscesses can completely obscure the visualisation of the portal vein and cause deformation of nearby structures. The left kidney, located near the spleen, is more challenging to visualise using a transverse approach, requiring special positioning of the cadaver. The right kidney, while generally easy to visualise, exhibited anomalous shadowing in some cases, which was not fully explained. Overall, training machine learning models solely on image data from cadaver subjects presents significant challenges due to the physiological variations, signs of morbidity, and disease processes commonly found in older individuals and cadavers. To mitigate these challenges, it is recommended to carefully select cadaver subjects with relatively normal anatomy for training and analysis.

Chapter 7 built upon the previous phantom study where the impact of infrared positional information on machine learning classification accuracy for six challenging-to-differentiate ultrasound abdominal cross sections was demonstrated to be ~4.3%. However, this previous study had limitations, notably overfitting due to a single subject and incomplete calibration algorithm testing. To address these limitations, this chapter leverages the variability in anatomy visibility and abdominal cavity size among cadavers to validate the positional tracking and calibration system for machine learning classification. The study collected six common abdominal scans and performed three calibration point scans on eleven cadavers using an ultrasound probe with an attached infrared sensor. Neural networks were trained using image-only and position-augmented datasets through transfer learning. Notably, an image-only approach using transfer learning from previous phantom-trained models failed due to the substantial variation in the cadaver image sample set. However, the inclusion of positional inferred sensor data led to average classification accuracies of 88.3% for one-point calibration, 91.5% for two-point calibration, and 92.8% for three-point calibration. These findings suggest that positional tracking can significantly enhance the recognition of challenging and hard-to-identify diagnostic ultrasound cross sections. Furthermore, the application of machine learning to facilitate the collection of ultrasound diagnostic cross sections holds the potential to streamline clinical workflows by automating image capture and supporting decision-making. It also paves the way for the automation of the entire collection process.

This thesis documents a baseline response of neural networks to the Japanese abdominal ultrasound screening protocol and provided a potential solution to the difficulties surrounding overlapping cross sections and edge cases that confounded image-only classification. A two-phase method to cost effectively collect and annotate data for small scale trials was developed, showing that in many cases costs could potentially be reduced by as much as 50% with only limited reduction in accuracy. A successful phantom study of the use of infrared tracking as an additional source of data for machine learning classification provided a limited proof for further study showing potential for accuracies above 95%, although these results were limited by the small sample size. The efficacy of Thiel cadavers for machine learning of abdominals cross sections was then examined to understand the potential use of Theil cadavers in a small trial, with results suggesting that while cadavers provided multiple deviations from normal physiology the anatomy within the images themselves were fairly clear in subjects without physiological deviation due to morbidity or old age. A cadaver study of the infrared tracking system confirmed that infrared tracking successfully increased the classification accuracy of abdominal cross sections but also showed it could be used in cases where positioning was correct, but the image was unclear. This also provided a larger, more diverse sample size, providing the opportunity to test the calibration of the algorithm on the abdominal cavity showing that additional points of calibration greatly improved classification accuracy.

In order to achieve these objectives, it was necessary to solve increasingly complex, interconnected tasks and use cases, these include:

- The classification of static imagery
- The sorting and classification of ultrasound video sweeps
- Localisation of probe in relation to each abdominal within a protocol
- Improving positional identification by normalising the positional localisation to the patient

Initial work developed a method to sort and classify abdominal images from a static dataset. Convolution neural networks are designed for this type of classification task but the lack image detail within the dataset, coupled with the overlapping regions of interest (ROI) complicate the classification task. From the initial baseline response, this work was progressed of identifying the correct cross sections from full videos of ultrasound abdominal sweeps as part of an industry protype with Canon Medical in Japan. This image-only method is highly successful in collecting the correct cross-sectional imagery when the sonographer is adept at placing the probe in the correct position and angle, potentially speeding up collection for expert users as they do not have to press a button to collect the correct cross sections but simply sweep the ultrasound probe over the correct area and the neural network will automatically collect the best imagery from that available, but this method is severely limited in functionality especially in large scale protocols with multiple overlapping ROI, edge cases and cannot provide assistance in positioning the probe without a significantly larger dataset.

The significant limitations of an image-only approach, coupled with the lack of a fixed positional localisation as frame of reference as is available to other medical imaging modalities such as CT and MRI where the patient is moved as required within the

scanner. This required the development of a much more complex system than previously envisaged, the initial infrared positional prototype showed that positional information could significantly improve classification accuracy by localising the ultrasound probe in relation to each cross section. This fixed positional method was potentially very limited in its use as it did not fit to the patient, leading to a high potential for error.

In order to further progress the proof of concept, the positional data received by the neural network must be normalised to the size of the abdominal cavity to improve accuracy. This additional task was achieved by stipulating a number of fixed points on the abdomen and using these points to adjust the positional data to normalise it to that already seen by the network. The results of the cadaver study suggest significant future potential of this system to enhance ultrasound scanning in the future.

The achievement of these tasks represents a number of significant milestones towards a machine learning system for classification of ultrasound that has significant practicality within clinical practice. While image-only has significant drawbacks and limitations that would limit future clinical use, the development of a positional system could not only speed up classification and collection of ultrasound scans by expert users but also could be used to provide navigational assistance and adherence to collection protocols.

7.1.1. Future Works

Based on the findings and limitations identified in this thesis, there are several future works and areas of study that would further advance the field of abdominal ultrasound imaging and machine learning:

The development of automated and guided ultrasound scanning comes at a time where there is not only an increased dependency on imaging modalities as the primary focus of confirming a differential diagnosis, but also a worldwide shortage of Sonographers capable of performing advanced ultrasound scans. Once productised this technology will assist operators with lower levels of training and experience in the collection of high-quality ultrasound scans. Reducing the skill floor for this workflow, allowing it to be undertaken more cost effectively and also decreasing the burden on senior clinicians.

There are still many restrictions surrounding the use of AI in medical imaging and diagnostics, with regulations still evolving, as machine learning technology continues to be refined. As part of this there is a push towards ensuring that algorithms produce results that are clear and explainable to a human observer. Removing the uncertainty as to what features a neural network is using to classify an image, adding much needed transparency to the process.

The development of real-time feedback systems for sonographers during ultrasound scans can assist in improving image quality. These systems could provide guidance on probe positioning, pressure, and settings to optimise image capture. This level of integration should examine the efficacy of sensor tracked machine learning for educational programs and resources for healthcare students and professionals to learn how to effectively utilise ultrasound scanners.

The potential for automating the collection process of medical ultrasound using robotics would allow it to be utilised similar to that of gold standard high resolution imaging modalities such as CT and MRI, allowing for increased uptake throughout diagnostic

medicine or even designation as a first line diagnostic modality within a clinical protocol which would see mandated widespread adoption. In order for this to happen, this technology will require extensive validation studies involving both living patients and cadaver to assess the real-world performance of machine learning models. This will be critical for ensuring the clinical relevance and safety of these systems.

Expanding dataset size is crucial to improve the robustness and generalisation of machine learning models. Collecting more cadaveric or patient data, ideally from a diverse range of sources and demographic backgrounds, would help address the limitations of the small dataset used in this thesis. The use of ensemble learning techniques should be considered for leveraging neural networks trained on different sources of data such as patient and cadaver. This might also include combining information from different imaging modalities (e.g., ultrasound, MRI, CT) and sensor data (e.g., positional tracking) may lead to more comprehensive and accurate diagnostic models. Use of generative AI has the potential to expand the sample size of existing ultrasound datasets by leveraging the anatomical features within this data to produce new ultrasound images.

Further research into advanced positional tracking technologies and methods can help overcome challenges related to probe positioning. This could include the development of specialised tracking devices or algorithms for better accuracy in both living patients and cadaveric models. This would include examining electromagnetic tracking systems which have submillimetre accuracy and prior clinical use as part of 3D ultrasound systems. Machine learning augmented with positional coordinate data has the potential

to also enhance the segmentation of human anatomy, with organs identified within the cross sections providing more granular information for clinicians.

Developing automated systems that can identify and flag anomalies or challenging cases during ultrasound scans can aid sonographers and improve diagnostic accuracy. This should form part art of a broader strategy to investigate potential opportunities for effective collaboration between healthcare professionals and machine learning systems. This includes studying how clinicians can integrate AI tools into their diagnostic workflow and make informed decisions based on AI-generated insights.

It is also important to consider how this technology fit into broader paradigms within machine learning. Federated learning is increasingly popular as it allows for multiple sites to contribute to the training of a central algorithm while maintaining local processing and storage of source data. Edge computing could also allow for lightweight algorithms to be integrated directly into the probe.

Overall, future research should aim to leverage the potential of machine learning while addressing the practical and ethical considerations required for the successful integration of AI into clinical practice with the potential of fully automating collection using robotics in the future.

References

[1] N. England and N. Improvement, "Diagnostic imaging dataset annual statistical release 2019/20," ed, 2022.

[2] K. Doi, "Diagnostic imaging over the last 50 years: research and development in medical imaging science and technology," *Physics in Medicine & Biology,* vol. 51, no. 13, p. R5, 2006.

[3] A. Chiolero, "Rising demand for medical imaging: what happened to evidence based medicine?," *bmj,* vol. 379, 2022.

[4] W. R. Hendee *et al.*, "Addressing overutilization in medical imaging," *Radiology,* vol. 257, no. 1, pp. 240-245, 2010.

[5] B. Allen, M. Chatfield, J. Burleson, and W. T. Thorwarth, "Improving diagnosis in health care: perspectives from the American College of Radiology," *Diagnosis,* vol. 4, no. 3, pp. 113-124, 2017.

[6] D. G. Fryback and J. R. Thornbury, "The efficacy of diagnostic imaging," *Medical decision making,* vol. 11, no. 2, pp. 88-94, 1991.

[7] R. Weissleder, J. Wittenberg, M. Harisinghani, and J. Chen, "Primer of diagnostic imaging," *August,* vol. 17, pp. 220-21, 2011.

[8] J. N. Itri, R. R. Tappouni, R. O. McEachern, A. J. Pesch, and S. H. Patel, "Fundamentals of diagnostic error in imaging," *Radiographics,* vol. 38, no. 6, pp. 1845-1865, 2018.

[9] J. N. Itri and S. H. Patel, "Heuristics and cognitive error in medical imaging," *American Journal of Roentgenology,* vol. 210, no. 5, pp. 1097-1105, 2018.

[10] E. E. Bartlett, "Physicians' cognitive errors and their liability consequences," *Journal of Healthcare Risk Management,* vol. 18, no. 4, pp. 62-69, 1998.

[11] M. T. Beinfeld and G. S. Gazelle, "Diagnostic imaging costs: are they driving up the costs of hospital care?," *Radiology,* vol. 235, no. 3, pp. 934-939, 2005.

[12] G. Juliusson, B. Thorvaldsdottir, J. M. Kristjansson, and P. Hannesson, "Diagnostic imaging trends in the emergency department: an extensive single-center experience," *Acta radiologica open,* vol. 8, no. 7, p. 2058460119860404, 2019.

[13] R. Smith-Bindman, D. L. Miglioretti, and E. B. Larson, "Rising use of diagnostic medical imaging in a large integrated health system," *Health affairs,* vol. 27, no. 6, pp. 1491-1502, 2008.

[14] D. C. Levin and V. M. Rao, "Factors that will determine future utilization trends in diagnostic imaging," *Journal of the American College of Radiology,* vol. 13, no. 8, pp. 904-908, 2016.

[15] M. Richards, G. Maskell, K. Halliday, and M. Allen, "Diagnostics: a major priority for the NHS," *Future healthcare journal,* vol. 9, no. 2, p. 133, 2022.

[16] N. England and N. Improvement, "Diagnostic imaging dataset annual statistical release," *London: Department of Health,* 2023.

[17] J. Caird, K. Hinds, I. Kwan, and J. Thomas, *A systematic rapid evidence assessment of late diagnosis*. Social Science Research Unit, 2012.

[18] G. Harrison and A. Harris, "Work-related musculoskeletal disorders in ultrasound: Can you reduce risk?," *Ultrasound,* vol. 23, no. 4, pp. 224-230, 2015.
[19] C. T. Coffin, "Work-related musculoskeletal disorders in sonographers: a review of causes and types of injury and best practices for reducing injury risk," *Reports in Medical Imaging,* pp. 15-26, 2014.
[20] C. Ionescu, "The Burden of Work Injuries in Sonography with a Focus on Diagnostic Cardiac Sonographers' Lived Experiences," Franklin University, 2023.
[21] C. Naomi, "Strategies for eliminating the sonographer shortage: Recruitment, retention, and educational perspectives," *Journal of Diagnostic Medical Sonography,* vol. 20, no. 6, pp. 408-413, 2004.
[22] P. Parker and G. Harrison, "Educating the future sonographic workforce: Membership survey report from the British Medical Ultrasound Society," *Ultrasound,* vol. 23, no. 4, pp. 231-241, 2015.
[23] R. E. Pow, C. Mello‐Thoms, and P. Brennan, "Evaluation of the effect of double reporting on test accuracy in screening and diagnostic imaging studies: a review of the evidence," *Journal of medical imaging and radiation oncology,* vol. 60, no. 3, pp. 306-314, 2016.
[24] H. Geijer and M. Geijer, "Added value of double reading in diagnostic radiology, a systematic review," *Insights into imaging,* vol. 9, no. 3, pp. 287-301, 2018.
[25] W. H. Organization, "Diagnostic errors," 2016.
[26] D. Khullar, A. K. Jha, and A. B. Jena, "Reducing diagnostic errors—why now?," *The New England journal of medicine,* vol. 373, no. 26, p. 2491, 2015.
[27] C. S. Lee, P. G. Nagy, S. J. Weaver, and D. E. Newman-Toker, "Cognitive and system factors contributing to diagnostic errors in radiology," *American Journal of Roentgenology,* vol. 201, no. 3, pp. 611-617, 2013.
[28] C. K. Schott *et al.*, "Retention of point-of-care ultrasound skills among practicing physicians: findings of the VA National POCUS Training Program," *The American Journal of Medicine,* vol. 134, no. 3, pp. 391-399. e8, 2021.
[29] B. J. Kimura, S. M. Sliman, J. Waalen, S. A. Amundson, and D. J. Shaw, "Retention of ultrasound skills and training in “point-of-care” cardiac ultrasound," *Journal of the American Society of Echocardiography,* vol. 29, no. 10, pp. 992-997, 2016.
[30] P. Steinmetz, S. Oleskevich, and J. Lewis, "Acquisition and long‐term retention of bedside ultrasound skills in first‐year medical students," *Journal of Ultrasound in Medicine,* vol. 35, no. 9, pp. 1967-1975, 2016.
[31] J. J. Deeks, "Systematic reviews of evaluations of diagnostic and screening tests," *Bmj,* vol. 323, no. 7305, pp. 157-162, 2001.
[32] V. Hasselblad and L. V. Hedges, "Meta-analysis of screening and diagnostic tests," *Psychological bulletin,* vol. 117, no. 1, p. 167, 1995.
[33] R. A. Smith, "Principles of successful cancer screening," *Surgical Oncology Clinics of North America,* vol. 8, no. 4, pp. 587-609, 1999.

[34] J. A. Arter and J. R. Jenkins, "Differential diagnosis—prescriptive teaching: A critical appraisal," *Review of educational research*, vol. 49, no. 4, pp. 517-555, 1979.
[35] P. J. Neumann, G. D. Sanders, L. B. Russell, J. E. Siegel, and T. G. Ganiats, *Cost-effectiveness in health and medicine*. Oxford University Press, 2016.
[36] J. T. Cohen and P. J. Neumann, "The cost savings and cost-effectiveness of clinical preventive care," *POLICY*, vol. 1, p. 6, 2009.
[37] M. Miwa, J. Thomas, A. O'Mara-Eves, and S. Ananiadou, "Reducing systematic review workload through certainty-based screening," *Journal of biomedical informatics*, vol. 51, pp. 242-253, 2014.
[38] J. A. Jensen, "Medical ultrasound imaging," *Progress in biophysics and molecular biology*, vol. 93, no. 1-3, pp. 153-165, 2007.
[39] J. S. Rose and A. E. Bair, "Fundamentals of ultrasound," *Practical guide to Emergency Ultrasound. PA: Lippincott Williams and Wilkins*, pp. 27-41, 2006.
[40] S. J. Patey and J. P. Corcoran, "Physics of ultrasound," *Anaesthesia & Intensive Care Medicine*, vol. 22, no. 1, pp. 58-63, 2021.
[41] C. M. Systems. *Image of ultrasound device*. Available: https://eu.medical.canon/products/ultrasound/aplioi800_imaging
[42] C. M. Systems. *Image of a Convex Probe*. Available: https://uk.medical.canon/products/ultrasound/liver-analysis-package/
[43] K. Martin, "The acoustic safety of new ultrasound technologies," *Ultrasound*, vol. 18, no. 3, pp. 110-118, 2010.
[44] The National Institute for Health and Care Excellence. (2021). *[MIB254], Butterfly iQ+ for diagnostic ultrasound imaging - Medtech innovation briefing [MIB254]*. Available: https://www.nice.org.uk/advice/mib254/chapter/Summary
[45] J. A. Gallego‐Juárez, "Basic principles of ultrasound," *Ultrasound in food processing: Recent advances*, pp. 1-26, 2017.
[46] C. Jacques and C. Pierre, "Development, via compression, of electric polarization in hemihedral crystals with inclined faces," *Bull. Soc. Minérologique Fr*, vol. 3, pp. 90-3, 1880.
[47] L. E. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders, *Fundamentals of acoustics*. 1999.
[48] K. K. Shung, "High frequency ultrasonic imaging," *Journal of medical ultrasound*, vol. 17, no. 1, pp. 25-30, 2009.
[49] P. R. Hoskins, K. Martin, and A. Thrush, *Diagnostic ultrasound: physics and equipment*. Cambridge University Press, 2010.
[50] A. Hindi, C. Peterson, and R. G. Barr, "Artifacts in diagnostic ultrasound," *Reports in Medical Imaging*, pp. 29-48, 2013.
[51] A. Hindi, C. Peterson, and R. G. Barr, "Artifacts in diagnostic ultrasound," *Reports in Medical Imaging*, vol. 6, pp. 29-48, 2013.
[52] F. W. Kremkau and K. Taylor, "Artifacts in ultrasound imaging," *Journal of ultrasound in medicine*, vol. 5, no. 4, pp. 227-237, 1986.
[53] M. K. Feldman, S. Katyal, and M. S. Blackwood, "US artifacts," *Radiographics*, vol. 29, no. 4, pp. 1179-1189, 2009.

[54] A. L. Y. Hung and J. Galeotti, "Good and bad boundaries in ultrasound compounding: preserving anatomic boundaries while suppressing artifacts," *International Journal of Computer Assisted Radiology and Surgery*, vol. 16, no. 11, pp. 1957-1968, 2021.
[55] M. Komatsu *et al.*, "Towards clinical application of artificial intelligence in ultrasound imaging," *Biomedicines*, vol. 9, no. 7, p. 720, 2021.
[56] F. Chen *et al.*, "Deep semi-supervised ultrasound image segmentation by using a shadow aware network with boundary refinement," *IEEE Transactions on Medical Imaging*, 2023.
[57] R. K. Mlosek, B. Migda, and M. Migda, "High-frequency ultrasound in the 21 century," *Journal of Ultrasonography*, vol. 20, no. 83, pp. 233-241, 2021.
[58] V. Damerjian, O. Tankyevych, N. Souag, and E. Petit, "Speckle characterization methods in ultrasound images–A review," *Irbm*, vol. 35, no. 4, pp. 202-213, 2014.
[59] T. Joel and R. Sivakumar, "An extensive review on Despeckling of medical ultrasound images using various transformation techniques," *Applied Acoustics*, vol. 138, pp. 18-27, 2018.
[60] M. L. Oelze and J. Mamou, "Review of quantitative ultrasound: Envelope statistics and backscatter coefficient imaging and contributions to diagnostic ultrasound," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 63, no. 2, pp. 336-351, 2016.
[61] R. Sivakumar, M. Gayathri, and D. Nedumaran, "Speckle filtering of ultrasound b-scan images-a comparative study between spatial and diffusion filters," in *2010 IEEE Conference on Open Systems (ICOS 2010)*, 2010, pp. 80-85: IEEE.
[62] S. K. Pal, A. Bhardwaj, and A. Shukla, "A Review on Despeckling Filters in Ultrasound Images for Speckle Noise Reduction," in *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, 2021, pp. 973-978: IEEE.
[63] L. Zhu, C.-W. Fu, M. S. Brown, and P.-A. Heng, "A non-local low-rank framework for ultrasound speckle reduction," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5650-5658.
[64] R. A. Phillips, M. E. Stratmeyer, and G. R. Harris, "Safety and US Regulatory considerations in the nonclinical use of medical ultrasound devices," *Ultrasound in medicine & biology*, vol. 36, no. 8, pp. 1224-1228, 2010.
[65] G. t. Haar, "Ultrasonic imaging: safety considerations," *Interface focus*, vol. 1, no. 4, pp. 686-697, 2011.
[66] J. W. Hunt, M. Arditi, and F. S. Foster, "Ultrasound transducers for pulse-echo medical imaging," *IEEE Transactions on Biomedical Engineering*, no. 8, pp. 453-481, 1983.
[67] C. K. Holland, C. X. Deng, R. E. Apfel, J. L. Alderman, L. A. Fernandez, and K. J. Taylor, "Direct evidence of cavitation in vivo from diagnostic ultrasound," *Ultrasound in medicine & biology*, vol. 22, no. 7, pp. 917-925, 1996.
[68] B. M. U. Society, "Guidelines for the safe use of diagnostic ultrasound equipment," *Ultrasound*, vol. 18, pp. 52-59, 2010.

[69] R. Williams, A. Windsor, R. D. Rosin, D. V. Mann, and M. Crofton, "Ultrasound scanning of the acute abdomen by surgeons in training," *Annals of the Royal College of Surgeons of England*, vol. 76, no. 4, p. 228, 1994.
[70] B. D. LEWIS and E. M. JAMES, "Current applications of duplex and color Doppler ultrasound imaging: abdomen," in *Mayo Clinic Proceedings*, 1989, vol. 64, no. 9, pp. 1158-1169: Elsevier.
[71] N. J. Khati, J. Gorodenker, and M. C. Hill, "Ultrasound-guided biopsies of the abdomen," *Ultrasound Quarterly*, vol. 27, no. 4, pp. 255-268, 2011.
[72] T. C. Winter, F. T. Lee Jr, and J. L. Hinshaw, "Ultrasound-guided biopsies in the abdomen and pelvis," *Ultrasound quarterly*, vol. 24, no. 1, pp. 45-68, 2008.
[73] H. F. Routh, "Doppler ultrasound," *IEEE Engineering in Medicine and Biology Magazine*, vol. 15, no. 6, pp. 31-40, 1996.
[74] J.-L. Gennisson, T. Deffieux, M. Fink, and M. Tanter, "Ultrasound elastography: principles and techniques," *Diagnostic and interventional imaging*, vol. 94, no. 5, pp. 487-495, 2013.
[75] R. Hampson and G. Dobie, "Towards robust 3D registration of non-invasive tactile elasticity images of breast tissue for cost-effective cancer screening," 2023.
[76] E. Krestel, "Imaging Systems for Medical Diagnosis: Fundamentals and Technical Solutions-X-Ray Diagnostics-Computed Tomography-Nuclear Medical Diagnostics-Magnetic Resonance Imaging-Ultrasound Technology," *ismd*, p. 627, 1990.
[77] R. Hampson, G. West, and G. Dobie, "Tactile, orientation, and optical sensor fusion for tactile breast image mosaicking," *IEEE Sensors Journal*, vol. 23, no. 5, pp. 5315-5324, 2023.
[78] C. M. Systems. *Images of Doppler and Elastography*. Available: https://global.medical.canon/products/ultrasound/General_Imaging
[79] S. Mc Fadden, T. Roding, G. De Vries, M. Benwell, H. Bijwaard, and J. Scheurleer, "Digital imaging and radiographic practise in diagnostic radiography: an overview of current knowledge and practice in Europe," *Radiography*, vol. 24, no. 2, pp. 137-141, 2018.
[80] C. M. Hayre and W. A. Cox, *General Radiography: Principles and Practices*. CRC Press, 2020.
[81] H. E. Martz, C. M. Logan, D. J. Schneberk, and P. J. Shull, *X-ray Imaging: fundamentals, industrial techniques and applications*. CRC Press, 2016.
[82] C. M. Systems. *Image of an X-ray Scan*. Available: https://eu.medical.canon/products/xray/ultimax_i_cg_myelogram
[83] R. L. Perry, "Principles of conventional radiography and fluoroscopy," *Veterinary Clinics of North America: Small Animal Practice*, vol. 23, no. 2, pp. 235-252, 1993.
[84] F. Arfelli *et al.*, "Low-dose phase contrast x-ray medical imaging," *Physics in Medicine & Biology*, vol. 43, no. 10, p. 2845, 1998.
[85] R. A. Lewis, "Medical phase contrast x-ray imaging: current status and future prospects," *Physics in medicine & biology*, vol. 49, no. 16, p. 3573, 2004.

[86] B. Wall, G. Kendall, A. Edwards, S. Bouffler, C. Muirhead, and J. Meara, "What are the risks from medical X-rays and other low dose radiation?," *The British journal of radiology,* vol. 79, no. 940, pp. 285-294, 2006.
[87] S. Watson, A. Jones, W. Oatway, and J. Hughes, *Ionising radiation exposure of the UK population: 2005 review*. Health Protection Agency Chilton, Oxon, 2005.
[88] A. J. Maitino, D. C. Levin, L. Parker, V. M. Rao, and J. H. Sunshine, "Nationwide trends in rates of utilization of noninvasive diagnostic imaging among the Medicare population between 1993 and 1999," *Radiology,* vol. 227, no. 1, pp. 113-117, 2003.
[89] R. Smith-Bindman *et al.*, "Use of diagnostic imaging studies and associated radiation exposure for patients enrolled in large integrated health care systems, 1996-2010," *Jama,* vol. 307, no. 22, pp. 2400-2409, 2012.
[90] C. M. Systems. *example of a CT scan*. Available: https://eu.medical.canon/products/computed-tomography/aquilion-exceed-lb-clinical-gallery
[91] L. W. Goldman, "Principles of CT: radiation dose and image quality," *Journal of nuclear medicine technology,* vol. 35, no. 4, pp. 213-225, 2007.
[92] W. R. Webb, W. E. Brant, and N. M. Major, *Fundamentals of Body CT E-Book*. Elsevier Health Sciences, 2014.
[93] B. J. Doyle, L. G. Morris, A. Callanan, P. Kelly, D. A. Vorp, and T. M. McGloughlin, "3D reconstruction and manufacture of real abdominal aortic aneurysms: from CT scan to silicone model," *Journal of biomechanical engineering,* vol. 130, no. 3, 2008.
[94] S. F. PET, "Acquisition protocol considerations for combined PET/CT imaging," *The Journal of Nuclear Medicine,* vol. 45, no. 1, pp. 25S-35S, 2004.
[95] A. Berrington de Gonzalez, E. Pasqual, and L. Veiga, "Epidemiological studies of CT scans and cancer risk: the state of the science," *The British Journal of Radiology,* vol. 94, p. 20210471, 2021.
[96] D. J. Brenner and E. J. Hall, "Computed tomography—an increasing source of radiation exposure," *New England Journal of Medicine,* vol. 357, no. 22, pp. 2277-2284, 2007.
[97] R. Smith-Bindman, "Is computed tomography safe," *N Engl j Med,* vol. 363, no. 1, pp. 1-4, 2010.
[98] C. H. Schultz, R. Fairley, L. S.-L. Murphy, and M. Doss, "The risk of cancer from CT scans and other sources of low-dose radiation: a critical appraisal of methodologic quality," *Prehospital and disaster medicine,* vol. 35, no. 1, pp. 3-16, 2020.
[99] M. D. Cohen, "CT radiation dose reduction: can we do harm by doing good?," *Pediatric radiology,* vol. 42, no. 4, pp. 397-398, 2012.
[100] B. Newman and M. J. Callahan, "ALARA (as low as reasonably achievable) CT 2011—executive summary," *Pediatric radiology,* vol. 41, no. 2, p. 453, 2011.
[101] C. G. Roth and S. Deshmukh, *Fundamentals of Body MRI E-Book*. Elsevier Health Sciences, 2016.

[102] C. M. Systems. *example of a MRI scan*. Available: https://global.medical.canon/products/magnetic-resonance/aice-clinical-gallery-3t
[103] F. T. Wajer, "Non-Cartesian MRI scan time reduction through sparse sampling," 2001.
[104] O. Kraff, A. Fischer, A. M. Nagel, C. Mönninghoff, and M. E. Ladd, "MRI at 7 Tesla and above: demonstrated and potential capabilities," *Journal of Magnetic Resonance Imaging*, vol. 41, no. 1, pp. 13-33, 2015.
[105] T. Niendorf, D. K. Sodickson, G. A. Krombach, and J. Schulz-Menger, "Toward cardiovascular MRI at 7 T: clinical needs, technical solutions and research promises," *European radiology*, vol. 20, pp. 2806-2816, 2010.
[106] S. B. Vos, C. M. Tax, P. R. Luijten, S. Ourselin, A. Leemans, and M. Froeling, "The importance of correcting for signal drift in diffusion MRI," *Magnetic resonance in medicine*, vol. 77, no. 1, pp. 285-299, 2017.
[107] L. Winter, F. Seifert, L. Zilberti, M. Murbach, and B. Ittermann, "MRI‐related heating of implants and devices: a review," *Journal of Magnetic Resonance Imaging*, vol. 53, no. 6, pp. 1646-1665, 2021.
[108] L. P. Panych and B. Madore, "The physics of MRI safety," *Journal of Magnetic Resonance Imaging*, vol. 47, no. 1, pp. 28-43, 2018.
[109] M. F. Dempsey, B. Condon, and D. M. Hadley, "MRI safety review," in *Seminars in Ultrasound, CT and MRI*, 2002, vol. 23, no. 5, pp. 392-401: Elsevier.
[110] M. S. Reisch, "Helium shortage affecting instrument users to extend into 2013," *Chemical and Engineering News*, vol. 90, no. 29, pp. 32-34, 2012.
[111] M. Mahesh and P. B. Barker, "The MRI helium crisis: past and future," *Journal of the American College of Radiology*, vol. 13, no. 12, pp. 1536-1537, 2016.
[112] J. S. o. Sonographers. (2020, 27/11/2023). *Standardized method of abdominal ultrasound*. Available: https://www.jss.org/english/standard/abdominal.html
[113] H. Gray, *Gray's anatomy*. Arcturus Publishing, 2009.
[114] M. Sienz, A. Ignee, and C. Dietrich, "Sonography today: reference values in abdominal ultrasound: aorta, inferior vena cava, kidneys," *Zeitschrift fur Gastroenterologie*, vol. 50, no. 3, pp. 293-315, 2012.
[115] L. Beales, S. Wolstenhulme, J. Evans, R. West, and D. Scott, "Reproducibility of ultrasound measurement of the abdominal aorta," *British journal of surgery*, vol. 98, no. 11, pp. 1517-1525, 2011.
[116] H. Bengtsson, O. Ekberg, P. Aspelin, S. Källerö, and D. Bergqvist, "Ultrasound screening of the abdominal aorta in patients with intermittent claudication," *European Journal of Vascular Surgery*, vol. 3, no. 6, pp. 497-502, 1989.
[117] M. Gürtelschmid, M. Björck, and A. Wanhainen, "Comparison of three ultrasound methods of measuring the diameter of the abdominal aorta," *British Journal of Surgery*, vol. 101, no. 6, pp. 633-636, 2014.
[118] B. Dent, R. Kendall, A. Boyle, and P. Atkinson, "Emergency ultrasound of the abdominal aorta by UK emergency physicians: a prospective cohort study," *Emergency Medicine Journal*, vol. 24, no. 8, pp. 547-549, 2007.

[119] M. Blaivas and D. Theodoro, "Frequency of incomplete abdominal aorta visualization by emergency department bedside ultrasound," *Academic emergency medicine,* vol. 11, no. 1, pp. 103-105, 2004.
[120] W. Derwich *et al.*, "High resolution strain analysis comparing aorta and abdominal aortic aneurysm with real time three dimensional speckle tracking ultrasound," *European Journal of Vascular and Endovascular Surgery,* vol. 51, no. 2, pp. 187-193, 2016.
[121] S. Gisvold and A. Brubakk, "Measurement of instantaneous blood-flow velocity in the human aorta using pulsed Doppler ultrasound," *Cardiovascular research,* vol. 16, no. 1, pp. 26-33, 1982.
[122] C. Couinaud, "Le foie," *Etudes anatomiques et chirurgicales,* 1957.
[123] H. Strunk, G. Stuckmann, J. Textor, and W. Willinek, "Limitations and pitfalls of Couinaud's segmentation of the liver in transaxial Imaging," *European Radiology,* vol. 13, no. 11, pp. 2472-2482, 2003.
[124] N. Baršić, I. Lerotić, L. Smirčić-Duvnjak, V. Tomašić, and M. Duvnjak, "Overview and developments in noninvasive diagnosis of nonalcoholic fatty liver disease," *World Journal of Gastroenterology: WJG,* vol. 18, no. 30, p. 3945, 2012.
[125] Z. F. Lu, J. Zagzebski, and F. Lee, "Ultrasound backscatter and attenuation in human liver with diffuse disease," *Ultrasound in medicine & biology,* vol. 25, no. 7, pp. 1047-1054, 1999.
[126] B. Palmentieri *et al.*, "The role of bright liver echo pattern on ultrasound B-mode examination in the diagnosis of liver steatosis," *Digestive and Liver Disease,* vol. 38, no. 7, pp. 485-489, 2006.
[127] B. Yang *et al.*, "Prospective study of early detection for primary liver cancer," *Journal of cancer research and clinical oncology,* vol. 123, no. 6, pp. 357-360, 1997.
[128] K. Tzartzeva and A. G. Singal, "Testing for AFP in combination with ultrasound improves early liver cancer detection," ed: Taylor & Francis, 2018.
[129] M. Westwood *et al.*, "Contrast-enhanced ultrasound using SonoVue®(sulphur hexafluoride microbubbles) compared with contrast-enhanced computed tomography and contrast-enhanced magnetic resonance imaging for the characterisation of focal liver lesions and detection of liver metastases: a systematic review and cost-effectiveness analysis," *Health technology assessment (Winchester, England),* vol. 17, no. 16, p. 1, 2013.
[130] A. Bonder and N. Afdhal, "Utilization of FibroScan in clinical practice," *Current gastroenterology reports,* vol. 16, pp. 1-7, 2014.
[131] J. C. Cohen, J. D. Horton, and H. H. Hobbs, "Human fatty liver disease: old questions and new insights," *Science,* vol. 332, no. 6037, pp. 1519-1523, 2011.
[132] M. M. Center and A. Jemal, "International trends in liver cancer incidence rates," *Cancer Epidemiology and Prevention Biomarkers,* vol. 20, no. 11, pp. 2362-2368, 2011.
[133] K. D. Reesink, T. Hendrikx, P. J. van Gorp, A. P. Hoeks, and R. Shiri‐Sverdlov, "Ultrasonic Perfluorohexane‐Loaded Monocyte Imaging: Toward a

Minimally Invasive Technique for Selective Detection of Liver Inflammation in Fatty Liver Disease," *Journal of Ultrasound in Medicine,* vol. 37, no. 4, pp. 921-933, 2018.
[134] H. Iijima *et al.*, "Decrease in accumulation of ultrasound contrast microbubbles in non‐alcoholic steatohepatitis," *Hepatology Research,* vol. 37, no. 9, pp. 722-730, 2007.
[135] R. G. Barr, S. R. Wilson, D. Rubens, G. Garcia-Tsao, and G. Ferraioli, "Update to the society of radiologists in ultrasound liver elastography consensus statement," *Radiology,* vol. 296, no. 2, pp. 263-274, 2020.
[136] D. S. Sherman, D. N. Fish, and I. Teitelbaum, "Assessing renal function in cirrhotic patients: problems and pitfalls," *American journal of kidney diseases,* vol. 41, no. 2, pp. 269-278, 2003.
[137] M. Walser, "Assessing renal function from creatinine measurements in adults with chronic renal failure," *American journal of kidney diseases,* vol. 32, no. 1, pp. 23-31, 1998.
[138] O. Helenon, J. Correas, C. Balleyguier, M. Ghouadni, and F. Cornud, "Ultrasound of renal tumors," *European radiology,* vol. 11, pp. 1890-1901, 2001.
[139] J. Schlegel, P. Diggdon, and J. Cuellar, "The use of ultrasound for localizing renal calculi," *The Journal of Urology,* vol. 86, no. 4, pp. 367-369, 1961.
[140] A. V. Zubarev, "Ultrasound of renal vessels," *European radiology,* vol. 11, pp. 1902-1915, 2001.
[141] A. Nilsson, "Contrast-enhanced ultrasound of the kidneys," *European radiology,* vol. 14, p. P104, 2004.
[142] J.-M. Correas, D. Anglicheau, D. Joly, J.-L. Gennisson, M. Tanter, and O. Hélénon, "Ultrasound-based imaging methods of the kidney—recent developments," *Kidney International,* vol. 90, no. 6, pp. 1199-1210, 2016.
[143] D. B. Johnson, D. A. Duchene, G. D. Taylor, M. S. Pearle, and J. A. Cadeddu, "Contrast-enhanced ultrasound evaluation of radiofrequency ablation of the kidney: reliable imaging of the thermolesion," *Journal of endourology,* vol. 19, no. 2, pp. 248-252, 2005.
[144] M. Mahoney, A. Sorace, J. Warram, S. Samuel, and K. Hoyt, "Volumetric Contrast‐Enhanced Ultrasound Imaging of Renal Perfusion," *Journal of Ultrasound in Medicine,* vol. 33, no. 8, pp. 1427-1437, 2014.
[145] E. H. Chang *et al.*, "Diagnostic accuracy of contrast-enhanced ultrasound for characterization of kidney lesions in patients with and without chronic kidney disease," *BMC nephrology,* vol. 18, no. 1, p. 266, 2017.
[146] W.-H. Chow, L. M. Dong, and S. S. Devesa, "Epidemiology and risk factors for kidney cancer," *Nature Reviews Urology,* vol. 7, no. 5, p. 245, 2010.
[147] J. Xie, Y. Jiang, and H.-t. Tsui, "Segmentation of kidney from ultrasound images based on texture and shape priors," *IEEE transactions on medical imaging,* vol. 24, no. 1, pp. 45-57, 2005.
[148] M. L. Robbin, M. E. Lockhart, and R. G. Barr, "Renal imaging with ultrasound contrast: current status," *Radiologic Clinics,* vol. 41, no. 5, pp. 963-978, 2003.

[149] D. R. Elwood, "Cholecystitis," *Surgical Clinics of North America,* vol. 88, no. 6, pp. 1241-1252, 2008.
[150] R. Hundal and E. A. Shaffer, "Gallbladder cancer: epidemiology and outcome," *Clinical epidemiology,* vol. 6, p. 99, 2014.
[151] G. L. Bennett and E. J. Balthazar, "Ultrasound and CT evaluation of emergent gallbladder pathology," *Radiologic Clinics,* vol. 41, no. 6, pp. 1203-1216, 2003.
[152] H. J. Finberg and J. C. Birnholz, "Ultrasound evaluation of the gallbladder wall," *Radiology,* vol. 133, no. 3, pp. 693-698, 1979.
[153] R. J. Gaspari, E. Dickman, and D. Blehar, "Learning curve of bedside ultrasound of the gallbladder," *The Journal of emergency medicine,* vol. 37, no. 1, pp. 51-56, 2009.
[154] S. Bodzioch and M. R. Ogiela, "New approach to gallbladder ultrasonic images analysis and lesions recognition," *Computerized Medical Imaging and Graphics,* vol. 33, no. 2, pp. 154-170, 2009.
[155] R. Lunevicius, I. C. Nzenwa, and M. Mesri, "A nationwide analysis of gallbladder surgery in England between 2000 and 2019," *Surgery,* vol. 171, no. 2, pp. 276-284, 2022.
[156] A. L. Pozo, E. M. Godfrey, and K. M. Bowles, "Splenomegaly: investigation, diagnosis and management," *Blood reviews,* vol. 23, no. 3, pp. 105-111, 2009.
[157] R. E. Mebius and G. Kraal, "Structure and function of the spleen," *Nature reviews immunology,* vol. 5, no. 8, pp. 606-616, 2005.
[158] M. GÜREL, "Why is isolated spleen metastasis a rare entity?," *Turk J Gastroenterol,* vol. 21, no. 4, pp. 452-453, 2010.
[159] R. A. Sardenberg, C. Pinto, C. A. Bueno, and R. N. Younes, "Non-small cell lung cancer stage IV long-term survival with isolated spleen metastasis," *The Annals of Thoracic Surgery,* vol. 95, no. 4, pp. 1432-1434, 2013.
[160] H. H.-S. Lu, C.-M. Chen, Y.-M. Huang, and J.-S. Wu, "Computer-aided diagnosis of liver cirrhosis by simultaneous comparisons of the ultrasound images of liver and spleen," *Journal of Data Science,* vol. 6, no. 3, pp. 429-448, 2008.
[161] P. Lamb, A. Lund, R. Kanagasabay, A. Martin, J. Webb, and R. Reznek, "Spleen size: how well do linear ultrasound measurements correlate with three-dimensional CT volume assessments?," *The British journal of radiology,* vol. 75, no. 895, pp. 573-577, 2002.
[162] M. W. Andrews, "Ultrasound of the spleen," *World journal of surgery,* vol. 24, no. 2, pp. 183-187, 2000.
[163] J. O'Donohue, C. Ng, S. Catnach, P. Farrant, and R. Williams, "Diagnostic value of Doppler assessment of the hepatic and portal vessels and ultrasound of the spleen in liver disease," *European journal of gastroenterology & hepatology,* vol. 16, no. 2, pp. 147-155, 2004.
[164] G. Civardi *et al.*, "Ultrasound‐guided fine needle biopsy of the spleen: high clinical efficacy and low risk in a multicenter Italian study," *American journal of hematology,* vol. 67, no. 2, pp. 93-99, 2001.

[165] D. S. Longnecker, "Anatomy and Histology of the Pancreas," *Pancreapedia: The Exocrine Pancreas Knowledge Base,* 2014.
[166] C. Feig, A. Gopinathan, A. Neesse, D. S. Chan, N. Cook, and D. A. Tuveson, "The pancreas cancer microenvironment," ed: AACR, 2012.
[167] S. Hessel *et al.*, "A prospective evaluation of computed tomography and ultrasound of the pancreas," *Radiology,* vol. 143, no. 1, pp. 129-133, 1982.
[168] C. R. Sidden and K. J. Mortele, "Cystic tumors of the pancreas: ultrasound, computed tomography, and magnetic resonance imaging features," in *Seminars in Ultrasound, CT and MRI*, 2007, vol. 28, no. 5, pp. 339-356: Elsevier.
[169] W. J. Walls, G. Gonzalez, N. L. Martin, and A. W. Templeton, "B-scan ultrasound evaluation of the pancreas: Advantages and accuracy compared to other diagnostic techniques," *Radiology,* vol. 114, no. 1, pp. 127-134, 1975.
[170] A. Martinez-Noguera, E. Montserrat, S. Torrubia, J. Monill, and P. Estrada, "Ultrasound of the pancreas: update and controversies," *European radiology,* vol. 11, no. 9, pp. 1594-1606, 2001.
[171] M. D'Onofrio, A. Gallotti, F. Principe, and R. P. Mucelli, "Contrast-enhanced ultrasound of the pancreas," *World Journal of Radiology,* vol. 2, no. 3, p. 97, 2010.
[172] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal of research and development,* vol. 3, no. 3, pp. 210-229, 1959.
[173] I. El Naqa and M. J. Murphy, *What is machine learning?* Springer, 2015.
[174] J. W. Shavlik and T. G. Dietterich, *Readings in machine learning*. Morgan Kaufmann, 1990.
[175] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 161-168.
[176] A. V. Joshi, *Machine Learning and Artificial Intelligence*. Springer, 2020.
[177] A. L. Fradkov, "Early history of machine learning," *IFAC-PapersOnLine,* vol. 53, no. 2, pp. 1385-1390, 2020.
[178] R. Kaur and R. Gangwar, "A review on Naive Baye's (NB), J48 and K-means based mining algorithms for medical data mining," *Int Res J Eng Technol,* vol. 4, pp. 1664-1668, 2017.
[179] Y. Xu, Y. Zhou, P. Sekula, and L. Ding, "Machine learning in construction: From shallow to deep learning," *Developments in the built environment,* vol. 6, p. 100045, 2021.
[180] N. Nilsson and L. Machines, "Foundations of trainable pattern classifying systems," *McGraw-Hill, New York OBrien RM (2007) A caution regarding rules of thumb for variance ination factors. Qual Quant,* vol. 41, p. 673, 1965.
[181] I. J. Good, "Probability and the Weighing of Evidence," 1950.
[182] C. J. Haug and J. M. Drazen, "Artificial intelligence and machine learning in clinical medicine, 2023," *New England Journal of Medicine,* vol. 388, no. 13, pp. 1201-1208, 2023.
[183] E. B. Hunt, J. Marin, and P. J. Stone, "Experiments in induction," 1966.
[184] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician,* vol. 46, no. 3, pp. 175-185, 1992.

[185] P. Smolensky, "Information processing in dynamical systems: Foundations of harmony theory," 1986.
[186] I. Kononenko, "Experiments in automatic learning of medical diagnostic rules," *Technical report, Jozef Stefan Institute*, 1984.
[187] G. Malathi and V. Shanthi, "Histogram based classification of ultrasound images of placenta," *International Journal of Computer Applications,* vol. 1, no. 16, pp. 0975-8887, 2010.
[188] P. Linares, P. J. McCullagh, N. D. Black, and J. Dornan, "Feature selection for the characterization of ultrasonic images of the placenta using texture classification," in *2004 2nd IEEE International Symposium on Biomedical Imaging: Nano to Macro (IEEE Cat No. 04EX821)*, 2004, pp. 1147-1150: IEEE.
[189] S. Aruna, S. Rajagopalan, and L. Nandakishore, "An empirical comparison of supervised learning algorithms in disease detection," *International Journal of Information Technology Convergence and Services–IJITCS,* vol. 1, no. 4, pp. 81-92, 2011.
[190] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE transactions on systems, man, and cybernetics,* vol. 21, no. 3, pp. 660-674, 1991.
[191] G. Biau and E. Scornet, "A random forest guided tour," *Test,* vol. 25, pp. 197-227, 2016.
[192] W. S. Noble, "What is a support vector machine?," *Nature biotechnology,* vol. 24, no. 12, pp. 1565-1567, 2006.
[193] M. Amer, M. Goldstein, and S. Abdennadher, "Enhancing one-class support vector machines for unsupervised anomaly detection," in *Proceedings of the ACM SIGKDD workshop on outlier detection and description*, 2013, pp. 8-15.
[194] X. Shi, H. D. Cheng, and L. Hu, "Mass detection and classification in breast ultrasound images using fuzzy SVM," in *9th Joint International Conference on Information Sciences (JCIS-06)*, 2006: Atlantis Press.
[195] K. D. Krishna *et al.*, "Computer aided abnormality detection for kidney on FPGA based IoT enabled portable ultrasound imaging system," *Irbm,* vol. 37, no. 4, pp. 189-197, 2016.
[196] V. Ulagamuthalvi and D. Sridharan, "Automatic identification of ultrasound liver cancer tumor using support vector machine," in *International conference on emerging trends in computer and electronics engineering*, 2012, pp. 41-43.
[197] J. M. Alvarez and M. Salzmann, "Learning the number of neurons in deep networks," *Advances in neural information processing systems,* vol. 29, 2016.
[198] J. Hastad, "Almost optimal lower bounds for small depth circuits," in *Proceedings of the eighteenth annual ACM symposium on Theory of computing*, 1986, pp. 6-20.
[199] A. C.-C. Yao, "Separating the polynomial-time hierarchy by oracles," in *26th Annual Symposium on Foundations of Computer Science (sfcs 1985)*, 1985, pp. 1-10: IEEE.
[200] Y. Bengio and O. Delalleau, "On the expressive power of deep architectures," in *International Conference on Algorithmic Learning Theory*, 2011, pp. 18-36: Springer.

[201] R. Eldan and O. Shamir, "The power of depth for feedforward neural networks," in *Conference on learning theory*, 2016, pp. 907-940.
[202] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics,* vol. 5, pp. 115-133, 1943.
[203] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain," *Psychological review,* vol. 65, no. 6, p. 386, 1958.
[204] F. Rosenblatt, *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms*. Spartan books Washington, DC, 1962.
[205] B. Widrow and M. E. Hoff, "Adaptive switching circuits (No. TR-1553-1)," ed: Stanford Univ Ca Stanford Electronics Labs, 1960.
[206] A. H. Klopf, *Brain function and adaptive systems: a heterostatic theory* (no. 133). Air Force Cambridge Research Laboratories, Air Force Systems Command, United …, 1972.
[207] P. J. Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE,* vol. 78, no. 10, pp. 1550-1560, 1990.
[208] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," ed: Institute for Cognitive Science, University of California, San Diego La …, 1985.
[209] K. Fukushima, "Cognitron: A self-organizing multilayered neural network," *Biological cybernetics,* vol. 20, no. 3-4, pp. 121-136, 1975.
[210] S. Grossberg, "Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors," *Biological cybernetics,* vol. 23, no. 3, pp. 121-134, 1976.
[211] S. G. Pauker, G. A. Gorry, J. P. Kassirer, and W. B. Schwartz, "Towards the simulation of clinical cognition: taking a present illness by computer," *The American journal of medicine,* vol. 60, no. 7, pp. 981-996, 1976.
[212] W. B. Schwartz, R. S. Patil, and P. Szolovits, "Artificial Intelligence in Medicine Where Do We Stand?," *Jurimetrics,* vol. 27, no. 4, pp. 362-369, 1987.
[213] J. J. Fenton *et al.*, "Influence of computer-aided detection on performance of screening mammography," *New England Journal of Medicine,* vol. 356, no. 14, pp. 1399-1409, 2007.
[214] C. R. Porter and E. D. Crawford, "Combining artificial neural networks and transrectal ultrasound in the diagnosis of prostate cancer," *Oncology (Williston Park, NY),* vol. 17, no. 10, pp. 1395-9; discussion 1399, 1403, 2003.
[215] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097-1105.
[216] K. Wu, X. Chen, and M. Ding, "Deep learning based classification of focal liver lesions with contrast-enhanced ultrasound," *Optik,* vol. 125, no. 15, pp. 4057-4063, 2014.
[217] H. Ravishankar, R. Venkataramani, S. Thiruvenkadam, P. Sudhakar, and V. Vaidya, "Learning and incorporating shape models for semantic segmentation," in *International conference on medical image computing and computer-assisted intervention*, 2017, pp. 203-211: Springer.

[218] A. Săftoiu *et al.*, "Accuracy of endoscopic ultrasound elastography used for differential diagnosis of focal pancreatic masses: a multicenter study," *Endoscopy,* vol. 43, no. 07, pp. 596-603, 2011.
[219] C.-C. Kuo *et al.*, "Automation of the kidney function prediction and classification through ultrasound-based kidney imaging using deep learning," *NPJ digital medicine,* vol. 2, no. 1, pp. 1-9, 2019.
[220] K. Lekadir *et al.*, "A convolutional neural network for automatic characterization of plaque composition in carotid ultrasound," *IEEE journal of biomedical and health informatics,* vol. 21, no. 1, pp. 48-55, 2016.
[221] D. Meng, L. Zhang, G. Cao, W. Cao, G. Zhang, and B. Hu, "Liver fibrosis classification based on transfer learning and FCNet for ultrasound images," *Ieee Access,* vol. 5, pp. 5804-5810, 2017.
[222] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Transactions on Evolutionary Computation,* vol. 23, no. 5, pp. 828-841, 2019.
[223] D. V. Vargas and J. Su, "Understanding the one-pixel attack: Propagation maps and locality analysis," *arXiv preprint arXiv:1902.02947,* 2019.
[224] D. V. Vargas, "One-Pixel Attack: Understanding and Improving Deep Neural Networks with Evolutionary Computation," in *Deep Neural Evolution*: Springer, 2020, pp. 401-430.
[225] S. A. Taghanaki, A. Das, and G. Hamarneh, "Vulnerability analysis of chest x-ray image classification against adversarial attacks," in *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*: Springer, 2018, pp. 87-94.
[226] S. G. Finlayson, H. W. Chung, I. S. Kohane, and A. L. Beam, "Adversarial attacks against medical deep learning systems," *arXiv preprint arXiv:1804.05296,* 2018.
[227] X. Ma *et al.*, "Understanding adversarial attacks on deep learning based medical image analysis systems," *Pattern Recognition,* p. 107332, 2020.
[228] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access,* vol. 6, pp. 14410-14430, 2018.
[229] R. Paul, M. Schabath, R. Gillies, L. Hall, and D. Goldgof, "Mitigating adversarial attacks on medical image understanding systems," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 2020, pp. 1517-1521: IEEE.
[230] H. Kim, D. C. Jung, and B. W. Choi, "Exploiting the vulnerability of deep learning-based artificial intelligence models in medical imaging: adversarial attacks," *Journal of the Korean Society of Radiology,* vol. 80, no. 2, pp. 259-273, 2019.
[231] M. L. Giger, H. P. Chan, and J. Boone, "Anniversary paper: history and status of CAD and quantitative image analysis: the role of medical physics and AAPM," *Medical physics,* vol. 35, no. 12, pp. 5799-5820, 2008.
[232] D. J. Manning, A. Gale, and E. A. Krupinski, "Perception research in medical imaging," *The British journal of radiology,* vol. 78, no. 932, pp. 683-685, 2005.

[233] S. H. C. Ortiz, T. Chiu, and M. D. Fox, "Ultrasound image enhancement: A review," *Biomedical Signal Processing and Control,* vol. 7, no. 5, pp. 419-428, 2012.
[234] S. L. Bridal, J.-M. Correas, A. Saied, and P. Laugier, "Milestones on the road to higher resolution, quantitative, and functional ultrasonic imaging," *Proceedings of the IEEE,* vol. 91, no. 10, pp. 1543-1561, 2003.
[235] J.-y. Lu, H. Zou, and J. F. Greenleaf, "Biomedical ultrasound beam forming," *Ultrasound in medicine & biology,* vol. 20, no. 5, pp. 403-428, 1994.
[236] W. Wilkening, B. Brendel, H. Jiang, J. Lazenby, and H. Ermert, "Optimized receive filters and phase-coded pulse sequences for contrast agent and nonlinear imaging," in *2001 IEEE Ultrasonics Symposium. Proceedings. An International Symposium (Cat. No. 01CH37263)*, 2001, vol. 2, pp. 1733-1737: IEEE.
[237] B. Mandersson and G. Salomonsson, "Weighted least-squares pulse-shaping filters with application to ultrasonic signals," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control,* vol. 36, no. 1, pp. 109-113, 1989.
[238] J. A. Jensen, "Deconvolution of ultrasound images," *Ultrasonic imaging,* vol. 14, no. 1, pp. 1-15, 1992.
[239] L. Panigrahi, K. Verma, and B. K. Singh, "Ultrasound image segmentation using a novel multi-scale Gaussian kernel fuzzy clustering and multi-scale vector field convolution," *Expert Systems with Applications,* vol. 115, pp. 486-498, 2019.
[240] O. B. Sassi, L. Sellami, M. B. Slima, A. B. Hamida, and K. Chtourou, "Multi-slices breast ultrasound lesion segmentation using multi-scale vector field convolution snake," in *2014 1st International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, 2014, pp. 188-192: IEEE.
[241] P. Singh, R. Mukundan, and R. De Ryke, "Feature enhancement in medical ultrasound videos using contrast-limited adaptive histogram equalization," *Journal of Digital Imaging,* pp. 1-13, 2019.
[242] C. J. Bouma, W. J. Niessen, K. J. Zuiderveld, E. J. Gussenhoven, and M. A. Viergever, "Automated lumen definition from 30 MHz intravascular ultrasound images," *Medical Image Analysis,* vol. 1, no. 4, pp. 363-377, 1997.
[243] M. Ploquin, A. Basarab, and D. Kouamé, "Resolution enhancement in medical ultrasound imaging," *Journal of Medical Imaging,* vol. 2, no. 1, pp. 017001-017001, 2015.
[244] G. Matrone, A. Ramalli, J. D'hooge, P. Tortoli, and G. Magenes, "A comparison of coherence-based beamforming techniques in high-frame-rate ultrasound imaging with multi-line transmission," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control,* vol. 67, no. 2, pp. 329-340, 2019.
[245] M. Cikes, L. Tong, G. R. Sutherland, and J. D'hooge, "Ultrafast cardiac ultrasound imaging: technical principles, applications, and clinical benefits," *JACC: Cardiovascular Imaging,* vol. 7, no. 8, pp. 812-823, 2014.
[246] J.-M. Correas, L. Bridal, A. Lesavre, A. Méjean, M. Claudon, and O. Hélénon, "Ultrasound contrast agents: properties, principles of action, tolerance, and artifacts," *European radiology,* vol. 11, pp. 1316-1328, 2001.

[247] A. Perperidis, "Postprocessing approaches for the improvement of cardiac ultrasound B-mode images: A review," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control,* vol. 63, no. 3, pp. 470-485, 2016.
[248] L. Armi and S. Fekri-Ershad, "Texture image analysis and texture classification methods-A review," *arXiv preprint arXiv:1904.06554,* 2019.
[249] M. Wei *et al.*, "A benign and malignant breast tumor classification method via efficiently combining texture and morphological features on ultrasound images," *Computational and Mathematical Methods in Medicine,* vol. 2020, 2020.
[250] F. Yang, J. Suri, and A. Fenster, "Segmentation of prostate from 3-D ultrasound volumes using shape and intensity priors in level set framework," in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, 2006, pp. 2341-2344: IEEE.
[251] D. Shen, Y. Zhan, and C. Davatzikos, "Segmentation of prostate boundaries from ultrasound images using statistical shape model," *IEEE transactions on medical imaging,* vol. 22, no. 4, pp. 539-551, 2003.
[252] A. K. Mishra, P. Roy, S. Bandyopadhyay, and S. K. Das, "Breast ultrasound tumour classification: A Machine Learning—Radiomics based approach," *Expert Systems,* vol. 38, no. 7, p. e12713, 2021.
[253] X. Ding *et al.*, "A novel wavelet-transform-based convolution classification network for cervical lymph node metastasis of papillary thyroid carcinoma in ultrasound images," *Computerized Medical Imaging and Graphics,* vol. 109, p. 102298, 2023.
[254] V. K. Sudarshan *et al.*, "Application of wavelet techniques for cancer diagnosis using ultrasound images: a review," *Computers in biology and medicine,* vol. 69, pp. 97-111, 2016.
[255] R. Roy, S. Ghosh, and A. Ghosh, "Clinical ultrasound image standardization using histogram specification," *Computers in Biology and Medicine,* vol. 120, p. 103746, 2020.
[256] K. Horsch, M. L. Giger, L. A. Venta, and C. J. Vyborny, "Automatic segmentation of breast lesions on ultrasound," *Medical physics,* vol. 28, no. 8, pp. 1652-1659, 2001.
[257] K. Drukker, M. L. Giger, K. Horsch, M. A. Kupinski, C. J. Vyborny, and E. B. Mendelson, "Computerized lesion detection on breast ultrasound," *Medical physics,* vol. 29, no. 7, pp. 1438-1446, 2002.
[258] W. D. Richard and C. G. Keen, "Automated texture-based segmentation of ultrasound images of the prostate," *Computerized Medical Imaging and Graphics,* vol. 20, no. 3, pp. 131-140, 1996.
[259] Z. Chang-ming, G. Guo-chang, L. Hai-bo, S. Jing, and Y. Hualong, "Segmentation of ultrasound image based on texture feature and graph cut," in *2008 International Conference on Computer Science and Software Engineering*, 2008, vol. 1, pp. 795-798: IEEE.
[260] J. Revell, M. Mirmehdi, and D. McNally, "Applied review of ultrasound image feature extraction methods," in *The 6th Medical Image Understanding and Analysis Conference,* 2002, pp. 173-176: BMVA Press.

[261] D. F. Schneider, "Machine Learning and Artificial Intelligence," in *Health Services Research*: Springer, 2020, pp. 155-168.
[262] L.-M. Fu, *Neural networks in computer intelligence*. Tata McGraw-Hill Education, 2003.
[263] H. Guo and H. L. Viktor, "Learning from imbalanced data sets with boosting and data generation: the databoost-im approach," *ACM Sigkdd Explorations Newsletter,* vol. 6, no. 1, pp. 30-39, 2004.
[264] S. García and F. Herrera, "Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy," *Evolutionary computation,* vol. 17, no. 3, pp. 275-306, 2009.
[265] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD explorations newsletter,* vol. 6, no. 1, pp. 20-29, 2004.
[266] A. Cano, D. T. Nguyen, S. Ventura, and K. J. Cios, "ur-CAIM: improved CAIM discretization for unbalanced and balanced data," *Soft Computing,* vol. 20, no. 1, pp. 173-188, 2016.
[267] Q. Wei and R. L. Dunbrack Jr, "The role of balanced training and testing data sets for binary classifiers in bioinformatics," *PloS one,* vol. 8, no. 7, p. e67863, 2013.
[268] J. Hernandez, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, "An empirical study of oversampling and undersampling for instance selection methods on imbalance datasets," in *Iberoamerican Congress on Pattern Recognition*, 2013, pp. 262-269: Springer.
[269] G. C. Cawley and N. L. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *The Journal of Machine Learning Research,* vol. 11, pp. 2079-2107, 2010.
[270] H. Jabbar and R. Z. Khan, "Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study)," *Computer Science, Communication and Instrumentation Devices,* pp. 163-172, 2015.
[271] R. Roelofs *et al.*, "A meta-analysis of overfitting in machine learning," in *Advances in Neural Information Processing Systems*, 2019, pp. 9179-9189.
[272] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research,* vol. 15, no. 1, pp. 1929-1958, 2014.
[273] S. Lawrence and C. L. Giles, "Overfitting and neural networks: conjugate gradient and backpropagation," in *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, 2000, vol. 1, pp. 114-119: IEEE.
[274] X. Sun, X. Ren, S. Ma, and H. Wang, "meprop: Sparsified back propagation for accelerated deep learning with reduced overfitting," *arXiv preprint arXiv:1706.06197,* 2017.
[275] G. C. Cawley and N. L. Talbot, "Preventing over-fitting during model selection via Bayesian regularisation of the hyper-parameters," *Journal of Machine Learning Research,* vol. 8, no. Apr, pp. 841-861, 2007.

[276] C. Zhang, O. Vinyals, R. Munos, and S. Bengio, "A study on overfitting in deep reinforcement learning," *arXiv preprint arXiv:1804.06893*, 2018.
[277] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.
[278] Z. Ghahramani, "Unsupervised learning," in *Summer School on Machine Learning*, 2003, pp. 72-112: Springer.
[279] F. Locatello *et al.*, "Challenging common assumptions in the unsupervised learning of disentangled representations," in *international conference on machine learning*, 2019, pp. 4114-4124.
[280] J. G. Dy and C. E. Brodley, "Feature selection for unsupervised learning," *Journal of machine learning research*, vol. 5, no. Aug, pp. 845-889, 2004.
[281] F. Zhuang *et al.*, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43-76, 2020.
[282] K. Mendel, H. Li, D. Sheth, and M. Giger, "Transfer learning from convolutional neural networks for computer-aided diagnosis: a comparison of digital breast tomosynthesis and full-field digital mammography," *Academic radiology*, vol. 26, no. 6, pp. 735-743, 2019.
[283] D. Han, Q. Liu, and W. Fan, "A new image classification method using CNN transfer learning and web data augmentation," *Expert Systems with Applications*, vol. 95, pp. 43-56, 2018.
[284] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Advances in neural information processing systems*, 2014, pp. 3320-3328.
[285] M. E. Taylor and P. Stone, "Transfer learning for reinforcement learning domains: A survey," *Journal of Machine Learning Research*, vol. 10, no. 7, 2009.
[286] V. Cheplygina, M. de Bruijne, and J. P. Pluim, "Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Medical image analysis*, vol. 54, pp. 280-296, 2019.
[287] M. Roberts *et al.*, "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans," *Nature Machine Intelligence*, vol. 3, no. 3, pp. 199-217, 2021.
[288] M. Ghassemi and E. O. Nsoesie, "In medicine, how do we machine learn anything real?," *Patterns*, vol. 3, no. 1, 2022.
[289] T. Hoff, "Deskilling and adaptation among primary care physicians using two work innovations," *Health Care Management Review*, vol. 36, no. 4, pp. 338-348, 2011.
[290] A. A. Povyakalo, E. Alberdi, L. Strigini, and P. Ayton, "How to discriminate between computer-aided and computer-hindered decisions: a case study in mammography," *Medical Decision Making*, vol. 33, no. 1, pp. 98-107, 2013.
[291] T. L. Tsai, D. B. Fridsma, and G. Gatti, "Computer decision support as a source of interpretation error: the case of electrocardiograms," *Journal of the American Medical Informatics Association*, vol. 10, no. 5, pp. 478-483, 2003.

[292] G. Varoquaux and V. Cheplygina, "Machine learning for medical imaging: methodological failures and recommendations for the future," *NPJ digital medicine,* vol. 5, no. 1, p. 48, 2022.
[293] A. A. Pulini, W. T. Kerr, S. K. Loo, and A. Lenartowicz, "Classification accuracy of neuroimaging biomarkers in attention-deficit/hyperactivity disorder: effects of sample size and circular analysis," *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging,* vol. 4, no. 2, pp. 108-120, 2019.
[294] S. Saeb, L. Lonini, A. Jayaraman, D. C. Mohr, and K. P. Kording, "The need to approximate the use-case in clinical machine learning," *Gigascience,* vol. 6, no. 5, p. gix019, 2017.
[295] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 1721-1730.
[296] A. Torres, F. Blasi, N. Dartois, and M. Akova, "Which individuals are at increased risk of pneumococcal disease and why? Impact of COPD, asthma, smoking, diabetes, and/or chronic heart disease on community-acquired pneumonia and invasive pneumococcal disease," *Thorax,* vol. 70, no. 10, pp. 984-989, 2015.
[297] J. Mongan, L. Moy, and C. E. Kahn Jr, "Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers," vol. 2, ed: Radiological Society of North America, 2020, p. e200029.
[298] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," *Journal of imaging,* vol. 6, no. 6, p. 52, 2020.
[299] I. R. Mendo, G. Marques, I. de la Torre Díez, M. López-Coronado, and F. Martín-Rodríguez, "Machine learning in medical emergencies: a systematic review and analysis," *Journal of Medical Systems,* vol. 45, no. 10, p. 88, 2021.
[300] A. Kaur, Y. Singh, N. Neeru, L. Kaur, and A. Singh, "A survey on deep learning approaches to medical images and a systematic look up into real-time object detection," *Archives of Computational Methods in Engineering,* pp. 1-41, 2022.
[301] C. Compagnone, G. Borrini, A. Calabrese, M. Taddei, V. Bellini, and E. Bignami, "Artificial intelligence enhanced ultrasound (AI-US) in a severe obese parturient: a case report," *The Ultrasound Journal,* vol. 14, no. 1, p. 34, 2022.
[302] Z. Salahuddin, H. C. Woodruff, A. Chatterjee, and P. Lambin, "Transparency of deep neural networks for medical image analysis: A review of interpretability methods," *Computers in biology and medicine,* vol. 140, p. 105111, 2022.
[303] H. Guan and M. Liu, "Domain adaptation for medical image analysis: a survey," *IEEE Transactions on Biomedical Engineering,* vol. 69, no. 3, pp. 1173-1185, 2021.
[304] G. Smith, Q. Zhang, and C. MacLellan, "Do it Like the Doctor: How We Can Design a Model That Uses Domain Knowledge to Diagnose Pneumothorax," *arXiv preprint arXiv:2205.12159*, 2022.
[305] L. Sun, X. Jiang, H. Ren, and Y. Guo, "Edge-cloud computing and artificial intelligence in internet of medical things: architecture, technology and application," *IEEE Access,* vol. 8, pp. 101079-101092, 2020.

[306] A. A. Abdellatif, A. Mohamed, C. F. Chiasserini, M. Tlili, and A. Erbad, "Edge computing for smart health: Context-aware approaches, opportunities, and challenges," *IEEE Network,* vol. 33, no. 3, pp. 196-203, 2019.
[307] F. Pesapane *et al.*, "Legal and regulatory framework for AI solutions in healthcare in Eu, us, China, and Russia: new scenarios after a pandemic," *Radiation,* vol. 1, no. 4, pp. 261-276, 2021.
[308] A. D. Saenz, Z. Harned, O. Banerjee, M. D. Abràmoff, and P. Rajpurkar, "Autonomous AI systems in the face of liability, regulations and costs," *NPJ digital medicine,* vol. 6, no. 1, p. 185, 2023.
[309] P. N. Wells and H.-D. Liang, "Medical ultrasound: imaging of soft tissue strain and elasticity," *Journal of the Royal Society Interface,* vol. 8, no. 64, pp. 1521-1549, 2011.
[310] K. K. Shung, "Diagnostic ultrasound: Past, present, and future," *J Med Biol Eng,* vol. 31, no. 6, pp. 371-4, 2011.
[311] K. A. Stewart *et al.*, "Trends in ultrasound use in low and middle income countries: a systematic review," *International Journal of Maternal and Child Health and AIDS,* vol. 9, no. 1, p. 103, 2020.
[312] L. J. Salomon *et al.*, "Practice guidelines for performance of the routine mid‐trimester fetal ultrasound scan," *Ultrasound in Obstetrics & Gynecology,* vol. 37, no. 1, pp. 116-126, 2011.
[313] V. Kini *et al.*, "Focused cardiac ultrasound in place of repeat echocardiography: reliability and cost implications," *Journal of the American Society of Echocardiography,* vol. 28, no. 9, pp. 1053-1059, 2015.
[314] L. Chan, T. Fung, T. Leung, D. Sahota, and T. Lau, "Volumetric (3D) imaging reduces inter‐and intraobserver variation of fetal biometry measurements," *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology,* vol. 33, no. 4, pp. 447-452, 2009.
[315] M. A. Maraci, R. Napolitano, A. Papageorghiou, and J. A. Noble, "Searching for structures of interest in an ultrasound video sequence," in *International Workshop on Machine Learning in Medical Imaging,* 2014, pp. 133-140: Springer.
[316] A. Garg and V. Mago, "Role of machine learning in medical research: A survey," *Computer science review,* vol. 40, p. 100370, 2021.
[317] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual review of biomedical engineering,* vol. 19, pp. 221-248, 2017.
[318] L. Cai, J. Gao, and D. Zhao, "A review of the application of deep learning in medical image classification and segmentation," *Annals of translational medicine,* vol. 8, no. 11, 2020.
[319] A. Paszke *et al.*, "Automatic differentiation in pytorch," 2017.
[320] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical image analysis,* vol. 42, pp. 60-88, 2017.

[321] D. Avola, L. Cinque, A. Fagioli, G. Foresti, and A. Mecca, "Ultrasound medical imaging techniques: a survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 3, pp. 1-38, 2021.
[322] S. Han *et al.*, "A deep learning framework for supporting the classification of breast lesions in ultrasound images," *Physics in Medicine & Biology*, vol. 62, no. 19, p. 7714, 2017.
[323] J. Chi, E. Walia, P. Babyn, J. Wang, G. Groot, and M. Eramian, "Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network," *Journal of digital imaging*, vol. 30, no. 4, pp. 477-486, 2017.
[324] M. Guo and Y. Du, "Classification of Thyroid Ultrasound Standard Plane Images using ResNet-18 Networks," in *2019 IEEE 13th International Conference on Anti-counterfeiting, Security, and Identification (ASID)*, 2019, pp. 324-328: IEEE.
[325] D. S. Reddy, R. Bharath, and P. Rajalakshmi, "A novel computer-aided diagnosis framework using deep learning for classification of fatty liver disease in ultrasound imaging," in *2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom)*, 2018, pp. 1-5: IEEE.
[326] D. Sabih and M. Hussain, "Automated classification of liver disorders using ultrasound images," *Journal of medical systems*, vol. 36, no. 5, pp. 3163-3172, 2012.
[327] M. Pesteie *et al.*, "Real-time ultrasound image classification for spine anesthesia using local directional Hadamard features," *International journal of computer assisted radiology and surgery*, vol. 10, no. 6, pp. 901-912, 2015.
[328] P. Zhu and Z. Li, "Guideline-based machine learning for standard plane extraction in 3D cardiac ultrasound," in *Medical Computer Vision and Bayesian and Graphical Models for Biomedical Imaging*: Springer, 2016, pp. 137-147.
[329] Y. Gao, Y. Zhu, B. Liu, Y. Hu, and Y. Guo, "Automated recognition of ultrasound cardiac views based on deep learning with graph constraint," *medRxiv*, 2020.
[330] C. Morioka *et al.*, "Automatic classification of ultrasound screening examinations of the abdominal aorta," *Journal of digital imaging*, vol. 29, pp. 742-748, 2016.
[331] P. M. Cheng and H. S. Malhi, "Transfer learning with convolutional neural networks for classification of abdominal ultrasound images," *Journal of digital imaging*, vol. 30, no. 2, pp. 234-243, 2017.
[332] Z. Xu *et al.*, "Less is more: Simultaneous view classification and landmark detection for abdominal ultrasound images," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, 2018, pp. 711-719: Springer.
[333] D. S. Reddy, P. Rajalakshmi, and M. Mateen, "A deep learning based approach for classification of abdominal organs using ultrasound images," *Biocybernetics and Biomedical Engineering*, vol. 41, no. 2, pp. 779-791, 2021.

[334] X. P. Burgos-Artizzu *et al.*, "Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes," *Scientific Reports*, vol. 10, no. 1, pp. 1-12, 2020.
[335] H. Chen *et al.*, "Ultrasound standard plane detection using a composite neural network framework," *IEEE transactions on cybernetics*, vol. 47, no. 6, pp. 1576-1586, 2017.
[336] P. Sridar, A. Kumar, A. Quinton, R. Nanan, J. Kim, and R. Krishnakumar, "Decision fusion-based fetal ultrasound image plane classification using convolutional neural networks," *Ultrasound in medicine & biology*, vol. 45, no. 5, pp. 1259-1273, 2019.
[337] W. Zhao, "Research on the deep learning of the small sample data based on transfer learning," in *AIP Conference Proceedings*, 2017, vol. 1864, no. 1, p. 020018: AIP Publishing LLC.
[338] P. Zhang, Y. Zhong, Y. Deng, X. Tang, and X. Li, "A survey on deep learning of small sample in biomedical image analysis," *arXiv preprint arXiv:1908.00473*, 2019.
[339] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211-252, 2015.
[340] Z. Xu *et al.*, "Less is more: Simultaneous view classification and landmark detection for abdominal ultrasound images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2018, pp. 711-719: Springer.
[341] C. M. S. Corperation. (2021, 27/11/2023). *Canon Aplio i800*. Available: https://global.medical.canon/products/ultrasound/aplioi800_imaging
[342] P. Mildenberger, M. Eichelberg, and E. Martin, "Introduction to the DICOM standard," *European radiology*, vol. 12, no. 4, pp. 920-927, 2002.
[343] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097-1105, 2012.
[344] A. Krizhevsky, "One weird trick for parallelizing convolutional neural networks," *arXiv preprint arXiv:1404.5997*, 2014.
[345] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[346] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.
[347] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1-9.
[348] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818-2826.
[349] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[350] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,* vol. 6, no. 02, pp. 107-116, 1998.
[351] H. Ide and T. Kurita, "Improvement of learning for CNN with ReLU activation by sparse regularization," in *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017, pp. 2684-2691: IEEE.
[352] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510-4520.
[353] A. Howard *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314-1324.
[354] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, 2019, pp. 6105-6114: PMLR.
[355] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size," *arXiv preprint arXiv:1602.07360,* 2016.
[356] A. Lawley, R. Hampson, K. Worrall, and G. Dobie, "A cost focused framework for optimizing collection and annotation of ultrasound datasets," *Biomedical Signal Processing and Control,* vol. 92, p. 106048, 2024.
[357] A. Lawley, R. Hampson, K. Worrall, and G. Dobie, "Prescriptive method for optimizing cost of data collection and annotation in machine learning of clinical ultrasound," in *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2023, pp. 1-4: IEEE.
[358] B. Luijten *et al.*, "Adaptive ultrasound beamforming using deep learning," *IEEE Transactions on Medical Imaging,* vol. 39, no. 12, pp. 3967-3978, 2020.
[359] L. H. Lee, Y. Gao, and J. A. Noble, "Principled ultrasound data augmentation for classification of standard planes," in *International Conference on Information Processing in Medical Imaging*, 2021, pp. 729-741: Springer.
[360] M. J. Willemink *et al.*, "Preparing medical imaging data for machine learning," *Radiology*, vol. 295, no. 1, pp. 4-15, 2020.
[361] J. Y.-L. Chan, K. T. Bea, S. M. H. Leow, S. W. Phoong, and W. K. Cheng, "State of the art: a review of sentiment analysis based on sequential transfer learning," *Artificial Intelligence Review*, vol. 56, no. 1, pp. 749-780, 2023.
[362] L. Schmarje, M. Santarossa, S.-M. Schröder, and R. Koch, "A survey on semi-, self-and unsupervised learning for image classification," *IEEE Access,* vol. 9, pp. 82146-82168, 2021.
[363] R. Huang, J. A. Noble, and A. I. Namburete, "Omni-supervised learning: scaling up to large unlabelled medical datasets," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2018, pp. 572-580: Springer.
[364] V. Rani, S. T. Nabi, M. Kumar, A. Mittal, and K. Kumar, "Self-supervised Learning: A Succinct Review," *Archives of Computational Methods in Engineering*, pp. 1-15, 2023.

[365] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16000-16009.
[366] M. Karnes, S. Perera, S. Adhikari, and A. Yilmaz, "Adaptive Few-Shot Learning PoC Ultrasound COVID-19 Diagnostic System," in *2021 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 2021, pp. 1-6: IEEE.
[367] A. Patra and J. A. Noble, "Hierarchical class incremental learning of anatomical structures in fetal echocardiography videos," *IEEE journal of biomedical and health informatics*, vol. 24, no. 4, pp. 1046-1058, 2020.
[368] M. J. Campbell, G. Lancaster, and S. Eldridge, "A randomised controlled trial is not a pilot trial simply because it uses a surrogate endpoint," *Pilot and Feasibility Studies*, vol. 4, pp. 1-4, 2018.
[369] K. Hemming, S. Eldridge, G. Forbes, C. Weijer, and M. Taljaard, "How to design efficient cluster randomised trials," *bmj*, vol. 358, 2017.
[370] D. J. Biau, S. Kernéis, and R. Porcher, "Statistics in brief: the importance of sample size in the planning and interpretation of medical research," *Clinical orthopaedics and related research*, vol. 466, no. 9, pp. 2282-2288, 2008.
[371] J. Cohen, "Statistical power analysis," *Current directions in psychological science*, vol. 1, no. 3, pp. 98-101, 1992.
[372] R. L. Figueroa, Q. Zeng-Treitler, S. Kandula, and L. H. Ngo, "Predicting sample size required for classification performance," *BMC medical informatics and decision making*, vol. 12, pp. 1-10, 2012.
[373] A. Rokem, Y. Wu, and A. Y. Lee, "Assessment of the need for separate test set and number of medical images necessary for deep learning: a sub-sampling study," *bioRxiv*, p. 196659, 2017.
[374] J. Cho, K. Lee, E. Shin, G. Choy, and S. Do, "How much data is needed to train a medical image deep learning system to achieve necessary high accuracy?," *arXiv preprint arXiv:1511.06348*, 2015.
[375] B. Widrow and M. E. Hoff, "Adaptive switching circuits," in *IRE WESCON convention record*, 1960, vol. 4, no. 1, pp. 96-104: New York.
[376] E. Baum and D. Haussler, "What size net gives valid generalization?," *Advances in neural information processing systems*, vol. 1, 1988.
[377] S. Haykin, *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1998.
[378] V. N. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," in *Measures of complexity: festschrift for alexey chervonenkis*: Springer, 2015, pp. 11-30.
[379] K. Fukunaga and R. R. Hayes, "Effects of sample size in classifier design," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 8, pp. 873-885, 1989.
[380] A. S. Detsky, "Are clinical trials a cost-effective investment?," *Jama*, vol. 262, no. 13, pp. 1795-1800, 1989.
[381] H. A. Glick, J. A. Doshi, S. S. Sonnad, and D. Polsky, *Economic evaluation in clinical trials*. OUP Oxford, 2014.

[382] D. Neijzen and G. Lunter, "Unsupervised learning for medical data: A review of probabilistic factorization methods," *Statistics in Medicine*, vol. 42, no. 30, pp. 5541-5554, 2023.
[383] B. Settles, "Algorithms for active learning," in *Cost-Sensitive Machine Learning*: CRC Press, 2011, pp. 21-48.
[384] B. Settles, "From theories to queries: Active learning in practice," in *Active learning and experimental design workshop in conjunction with AISTATS 2010*, 2011, pp. 1-18: JMLR Workshop and Conference Proceedings.
[385] S. Budd, E. C. Robinson, and B. Kainz, "A survey on active learning and human-in-the-loop deep learning for medical image analysis," *Medical Image Analysis*, vol. 71, p. 102062, 2021.
[386] Y. Yang, Z. Ma, F. Nie, X. Chang, and A. G. Hauptmann, "Multi-class active learning by uncertainty sampling with diversity maximization," *International Journal of Computer Vision*, vol. 113, no. 2, pp. 113-127, 2015.
[387] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in *SIGIR'94*, 1994, pp. 3-12: Springer.
[388] Z. Guochen, "Four uncertain sampling methods are superior to random sampling method in classification," in *2021 2nd International Conference on Artificial Intelligence and Education (ICAIE)*, 2021, pp. 209-212: IEEE.
[389] L. Venturini, A. T. Papageorghiou, J. A. Noble, and A. I. Namburete, "Uncertainty estimates as data selection criteria to boost omni-supervised learning," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, 2020, pp. 689-698: Springer.
[390] J. Yun, J. Oh, and I. Yun, "Gradually Applying Weakly Supervised and Active Learning for Mass Detection in Breast Ultrasound Images," *Applied Sciences*, vol. 10, no. 13, p. 4519, 2020.
[391] G. Liu *et al.*, "Breast Ultrasound Tumor Detection Based on Active Learning and Deep Learning," EasyChair2516-2314, 2021.
[392] L. Gao *et al.*, "Multi-modal active learning for automatic liver fibrosis diagnosis based on ultrasound shear wave elastography," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 2021, pp. 410-414: IEEE.
[393] L. Liu, W. Lei, X. Wan, L. Liu, Y. Luo, and C. Feng, "Semi-supervised active learning for COVID-19 lung ultrasound multi-symptom classification," in *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, 2020, pp. 1268-1273: IEEE.
[394] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy. (2020). *Dataset of breast ultrasound images*. Available: https://scholar.cu.edu.eg/?q=afahmy/pages/dataset
[395] J. Born *et al.*, "Accelerating detection of lung pathologies with explainable ultrasound image analysis," *Applied Sciences*, vol. 11, no. 2, p. 672, 2021.
[396] J. Born *et al.*, "L2 Accelerating COVID-19 differential diagnosis with explainable ultrasound image analysis: an AI tool," ed: BMJ Publishing Group Ltd, 2021.

[397] J. W. Tsung, D. O. Kessler, and V. P. Shah, "Prospective application of clinician-performed lung ultrasonography during the 2009 H1N1 influenza A pandemic: distinguishing viral from bacterial pneumonia," *Critical ultrasound journal,* vol. 4, no. 1, pp. 1-10, 2012.
[398] D. Malla, V. Rathi, S. Gomber, and L. Upreti, "Can lung ultrasound differentiate between bacterial and viral pneumonia in children?," *Journal of Clinical Ultrasound,* vol. 49, no. 2, pp. 91-100, 2021.
[399] A. Ahmad *et al.*, "Deep-AntiFP: Prediction of antifungal peptides using distanct multi-informative features incorporating with deep neural networks," *Chemometrics and Intelligent Laboratory Systems,* vol. 208, p. 104214, 2021.
[400] S. Akbar, M. Hayat, M. Tahir, S. Khan, and F. K. Alarfaj, "cACP-DeepGram: classification of anticancer peptides via deep neural network and skip-gram-based word embedding model," *Artificial intelligence in medicine,* vol. 131, p. 102349, 2022.
[401] S. Akbar, S. Khan, F. Ali, M. Hayat, M. Qasim, and S. Gul, "iHBP-DeepPSSM: Identifying hormone binding proteins using PsePSSM based evolutionary features and deep learning approach," *Chemometrics and Intelligent Laboratory Systems,* vol. 204, p. 104103, 2020.
[402] S. Akbar, M. Hayat, M. Iqbal, and M. A. Jan, "iACP-GAEnsC: Evolutionary genetic algorithm based ensemble classification of anticancer peptides by utilizing hybrid feature space," *Artificial intelligence in medicine,* vol. 79, pp. 62-70, 2017.
[403] V. L. Parsons, "Stratified sampling," *Wiley StatsRef: Statistics Reference Online,* pp. 1-11, 2014.
[404] C. Liem and A. Panichella, "Run, forest, run? on randomization and reproducibility in predictive software engineering," *arXiv preprint arXiv:2012.08387,* 2020.
[405] F. Ali, S. Ahmed, Z. N. K. Swati, and S. Akbar, "DP-BINDER: machine learning model for prediction of DNA-binding proteins by fusing evolutionary and physicochemical information," *Journal of Computer-Aided Molecular Design,* vol. 33, pp. 645-658, 2019.
[406] A. Ahmad, S. Akbar, M. Tahir, M. Hayat, and F. Ali, "iAFPs-EnC-GA: identifying antifungal peptides using sequential and evolutionary descriptors based multi-information fusion and ensemble learning approach," *Chemometrics and Intelligent Laboratory Systems,* vol. 222, p. 104516, 2022.
[407] S. Akbar, A. Ahmad, M. Hayat, A. U. Rehman, S. Khan, and F. Ali, "iAtbP-Hyb-EnC: Prediction of antitubercular peptides via heterogeneous feature representation and genetic algorithm based ensemble learning model," *Computers in Biology and Medicine,* vol. 137, p. 104778, 2021.
[408] T. A. Mulder *et al.*, "Unravelling the skillset of point-of-care ultrasound: a systematic review," *The Ultrasound Journal,* vol. 15, no. 1, pp. 1-13, 2023.
[409] K. Mohit, R. Gupta, and B. Kumar, "A Survey on the Machine Learning Techniques for Automated Diagnosis from Ultrasound Images," *Current Medical Imaging,* 2023.

[410] L. Dandan, M. Huanhuan, J. Yu, and S. Yi, "A multi-model organ segmentation method based on abdominal ultrasound image," in *2020 15th IEEE international conference on signal processing (ICSP)*, 2020, vol. 1, pp. 505-510: IEEE.
[411] C. Peng, Q. Cai, M. Chen, and X. Jiang, "Recent advances in tracking devices for biomedical ultrasound imaging applications," *Micromachines*, vol. 13, no. 11, p. 1855, 2022.
[412] F. Mohamed and C. V. Siang, "A survey on 3D ultrasound reconstruction techniques," *Artificial Intelligence—Applications in Medicine and Biology*, pp. 73-92, 2019.
[413] P. Beigi, S. E. Salcudean, G. C. Ng, and R. Rohling, "Enhancement of needle visualization and localization in ultrasound," *International journal of computer assisted radiology and surgery*, vol. 16, pp. 169-178, 2021.
[414] R. Summan *et al.*, "Spatial calibration of large volume photogrammetry based metrology systems," *Measurement*, vol. 68, pp. 189-200, 2015.
[415] S. Rana, R. Hampson, and G. Dobie, "Breast cancer: model reconstruction and image registration from segmented deformed image using visual and force based analysis," *IEEE Transactions on Medical Imaging*, vol. 39, no. 5, pp. 1295-1305, 2019.
[416] G. S. Mok, Q. Zhang, J. Sun, D. Zhang, P. H. Pretorius, and M. A. King, "Preliminary investigation of auto-classification of respiratory trace using convolutional neural network for adaptive respiratory gated myocardial perfusion SPECT," in *2019 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, 2019, pp. 1-3: IEEE.
[417] S. Vafadar, L. Gajny, M. Boëssé, and W. Skalli, "Evaluation of CNN-based human pose estimation for body segment lengths assessment," in *VipIMAGE 2019: Proceedings of the VII ECCOMAS Thematic Conference on Computational Vision and Medical Image Processing, October 16–18, 2019, Porto, Portugal*, 2019, pp. 179-187: Springer.
[418] P. Caserman, A. Garcia-Agundez, R. Konrad, S. Göbel, and R. Steinmetz, "Real-time body tracking in virtual reality using a Vive tracker," *Virtual Reality*, vol. 23, pp. 155-168, 2019.
[419] T. Ameler *et al.*, "A comparative evaluation of steamvr tracking and the optitrack system for medical device tracking," in *2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, 2019, pp. 1465-1470: IEEE.
[420] S. M. Van der Veen, M. Bordeleau, P. E. Pidcoe, C. R. France, and J. S. Thomas, "Agreement analysis between vive and vicon systems to monitor lumbar postural changes," *Sensors*, vol. 19, no. 17, p. 3632, 2019.
[421] L. Meszaros-Beller, M. Antico, D. Fontanarosa, and P. Pivonka, "Assessment of spinal curvatures in static postures using localized 3D ultrasound: A proof-of-concept study."
[422] J. Marharjan, B. R. Mitchell, V. W. Chan, and E. Kim, "Guided ultrasound imaging using a deep regression network," in *Medical Imaging 2020: Ultrasonic Imaging and Tomography*, 2020, vol. 11319, pp. 28-36: SPIE.

[423] G. Bradski, "The openCV library," *Dr. Dobb's Journal: Software Tools for the Professional Programmer,* vol. 25, no. 11, pp. 120-123, 2000.
[424] A. Lawley, R. Hampson, K. Worrall, and G. Dobie, "Analysis of neural networks for routine classification of sixteen ultrasound upper abdominal cross sections," *Abdominal Radiology,* pp. 1-11, 2024.
[425] P. M. Cheng and H. S. Malhi, "Transfer learning with convolutional neural networks for classification of abdominal ultrasound images," *Journal of digital imaging,* vol. 30, pp. 234-243, 2017.
[426] Vicon. *Vicon award winning motion tracking systems*. Available: https://www.vicon.com/
[427] P. Caserman, A. Garcia-Agundez, and S. Göbel, "A survey of full-body motion reconstruction in immersive virtual reality applications," *IEEE transactions on visualization and computer graphics,* vol. 26, no. 10, pp. 3089-3108, 2019.
[428] HTC. *Introducing VIVE Tracker (3.0)* Available: https://www.vive.com/uk/accessory/tracker3/
[429] HTC. *SteamVR Base Station 2.0*. Available: https://www.vive.com/uk/accessory/base-station2/
[430] K. Group. *OpenXR Overview*. Available: https://www.khronos.org/openxr/
[431] R. Caruana, S. Lawrence, and C. Giles, "Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping," *Advances in neural information processing systems,* vol. 13, 2000.
[432] N. S. Keskar, J. Nocedal, P. T. P. Tang, D. Mudigere, and M. Smelyanskiy, "On large-batch training for deep learning: Generalization gap and sharp minima," in *5th International Conference on Learning Representations, ICLR 2017*, 2017.
[433] J. Frimann-Dahl, "Normal variations of the left kidney: an anatomical and radiologic study," *Acta radiologica,* no. 3, pp. 207-216, 1961.
[434] W. Thiel, "The preservation of the whole corpse with natural color," *Annals of anatomy= Anatomischer Anzeiger: official organ of the Anatomische Gesellschaft,* vol. 174, no. 3, pp. 185-195, 1992.
[435] W. Thiel, "Supplement to the conservation of an entire cadaver according to W. Thiel," *Annals of anatomy= Anatomischer Anzeiger: official organ of the Anatomische Gesellschaft,* vol. 184, no. 3, pp. 267-269, 2002.
[436] M. Benkhadra *et al.*, "Flexibility of Thiel's embalmed cadavers: the explanation is probably in the muscles," *Surgical and Radiologic Anatomy,* vol. 33, pp. 365-368, 2011.
[437] J. Y. Balta *et al.*, "A comparison of embalming fluids on the structures and properties of tissue in human cadavers," *Anatomia, histologia, embryologia,* vol. 48, no. 1, pp. 64-73, 2019.
[438] R. Eisma and T. Wilkinson, "From "silent teachers" to models," *PLoS biology,* vol. 12, no. 10, p. e1001971, 2014.
[439] F. Waerlop, N. Rashidian, S. Marrannes, K. D'Herde, and W. Willaert, "Thiel embalmed human cadavers in surgical education: Optimizing realism and long-term application," *The American Journal of Surgery,* vol. 221, no. 6, pp. 1300-1302, 2021.

[440] N. Ernesto Ottone, C. A. Vargas, R. Fuentes, and M. del Sol, "Walter Thiel's Embalming Method. Review of Solutions and Applications in Different Fields of Biomedical Research," *International Journal of Morphology,* vol. 34, no. 4, 2016.
[441] M. Benkhadra *et al.*, "Comparison of fresh and Thiel's embalmed cadavers according to the suitability for ultrasound-guided regional anesthesia of the cervical region," *Surgical and radiologic anatomy,* vol. 31, pp. 531-535, 2009.
[442] S. Hayashi *et al.*, "History and future of human cadaver preservation for surgical training: from formalin to saturated salt solution method," *Anatomical science international,* vol. 91, pp. 1-7, 2016.
[443] R. Eisma, C. Lamb, and R. Soames, "From formalin to Thiel embalming: What changes? One anatomy department's experiences," *Clinical anatomy,* vol. 26, no. 5, pp. 564-571, 2013.
[444] M. Rakuša and L. K. Šaherl, "Thiel embalming method used for anatomy dissection as an educational tool in teaching human anatomy, in research, and in training in comparison of different methods for long term preservation," *Folia Morphologica,* 2022.
[445] G. McLeod, R. Eisma, A. Schwab, G. Corner, R. Soames, and S. Cochran, "An evaluation of Thiel-embalmed cadavers for ultrasound-based regional anaesthesia training and research," *Ultrasound,* vol. 18, no. 3, pp. 125-129, 2010.
[446] S. Munirama, R. Eisma, M. Columb, G. Corner, and G. McLeod, "Physical properties and functional alignment of soft-embalmed Thiel human cadaver when used as a simulator for ultrasound-guided regional anaesthesia," *BJA: British Journal of Anaesthesia,* vol. 116, no. 5, pp. 699-707, 2016.
[447] S. Munirama *et al.*, "Trainee anaesthetist diagnosis of intraneural injection—a study comparing B-mode ultrasound with the fusion of B-mode and elastography in the soft embalmed Thiel cadaver model," *BJA: British Journal of Anaesthesia,* vol. 117, no. 6, pp. 792-800, 2016.
[448] J. Joy *et al.*, "Quantitative assessment of Thiel soft-embalmed human cadavers using shear wave elastography," *Annals of Anatomy-Anatomischer Anzeiger,* vol. 202, pp. 52-56, 2015.
[449] R. Duncan, H. McLeod, B. Burton, S. Matthew, and G. Houston, "Thiel-embalmed cadavers as a model for Colour Doppler ultrasound and its applications," 2017: European Congress of Radiology-ECR 2017.
[450] I. Karakitsios, T. Saliev, H. McLeod, S. Ahmad, R. Eisma, and A. Melzer, "Response of thiel-embalmed human liver and kidney to MR-guided focused ultrasound surgery," *Biomedical Engineering/Biomedizinische Technik,* vol. 57, no. SI-1-Track-C, pp. 975-975, 2012.
[451] G. G. R. Schramek, D. Stoevesandt, A. Reising, J. T. Kielstein, M. Hiss, and H. Kielstein, "Imaging in anatomy: a comparison of imaging techniques in embalmed human cadavers," *BMC medical education,* vol. 13, pp. 1-7, 2013.
[452] J. Y. Balta *et al.*, "Assessing radiological images of human cadavers: Is there an effect of different embalming solutions?," *Journal of Forensic Radiology and Imaging,* vol. 11, pp. 40-46, 2017.

[453] D. University. (2023). *Thiel Cadaver Facility*. Available: https://www.dundee.ac.uk/locations/thiel-cadaver-facility
[454] T. Kuroiwa *et al.*, "Deep learning estimation of median nerve volume using ultrasound imaging in a human cadaver model," *Ultrasound in Medicine & Biology*, vol. 48, no. 11, pp. 2237-2248, 2022.
[455] A. Chuan *et al.*, "A randomised controlled trial comparing meat‐based with human cadaveric models for teaching ultrasound‐guided regional anaesthesia," *Anaesthesia*, vol. 71, no. 8, pp. 921-929, 2016.
[456] G. McLeod *et al.*, "Patterns of skills acquisition in anesthesiologists during simulated interscalene block training on a soft embalmed Thiel cadaver: cohort study," *JMIR Medical Education*, vol. 8, no. 3, p. e32840, 2022.
[457] M. Wagner *et al.*, "Implementation and evaluation of training for ultrasound-guided vascular access to small vessels using a low-cost cadaver model," *Pediatric Critical Care Medicine*, vol. 19, no. 11, pp. e611-e617, 2018.
[458] A. McNeill, A. Anstee, and T. Hoare, "Advanced ultrasound-guided biopsy simulation training using cadaveric phantoms: cadaver day," *Clinical Radiology*, vol. 75, no. 10, pp. 798. e23-798. e25, 2020.
[459] F. Macnaught and N. Campbell‐Rogers, "The liver: how we do it," *Australasian Journal of Ultrasound in Medicine*, vol. 12, no. 3, p. 44, 2009.
[460] R. Haddow and R. Kemp-Harper, "Calcification in the liver and portal system," *Clinical Radiology*, vol. 18, no. 3, pp. 225-236, 1967.
[461] Y. N. Zhang *et al.*, "Liver fat imaging—a clinical overview of ultrasound, CT, and MR imaging," *The British journal of radiology*, vol. 91, no. 1089, p. 20170959, 2018.