

UNIVERSITY OF LEEDS

AI-Integrated Treatment Planning and Real-Time Guidance for Ultrasound-Guided High Intensity Focused Ultrasound (USgHIFU)

Lewis James Howell

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds

School of Computer Science & School of Electronic and Electrical Engineering

August 2026

Intellectual Property and Publication Statements

I confirm that the work submitted is my own, except where work which has formed part of jointly authored publications has been included. My contributions and the contributions of others have been explicitly indicated below. I confirm that appropriate credit has been given within the thesis where reference has been made to the work of others.

Papers contributing to this thesis

1. L. Howell, N. Ingram, R. Lapham, A. Morrell, and J. R. McLaughlan, "Deep learning for real-time multi-class segmentation of artefacts in lung ultrasound", Ultrasonics, vol. 140, p. 107251, May 2024, reproduced in Chapter 2 under the terms of the Creative Commons CC BY 4.0 license. The contribution to this work is as follows:

    - **LH**: Conceptualisation, Methodology, Data curation, Software, Investigation, Analysis, Visualisation, Writing manuscript, Review & editing.
    - NI, RL, AM: Data curation.
    - JM: Conceptualisation, Methodology, Supervision, Funding acquisition, Data curation, Review & editing.

2. L. Howell, A. Zarei, T. M. Wah, J. H. Chandler, S. Karthik, Z. Court, H. Ng, and J. R. McLaughlan, "RADEX: a rule-based clinical and radiology data extraction tool demonstrated on thyroid ultrasound reports", European Radiology, vol. 35, pp. 4506–4517, Aug. 2025, reproduced in Chapter 3 under the terms of the Creative Commons CC BY 4.0 license. The contribution to this work is as follows:

    - **LH**: Conceptualisation, Methodology, Data curation, Software, Investigation, Analysis, Visualisation, Writing manuscript, Review & editing.
    - AZ: Methodology, Data curation (Equal), Review & editing.
    - ZC, NH: Data curation, Review & editing.‘
    - SK: Conceptualisation (Supporting).

- JC, JM: Conceptualisation, Methodology (Supporting), Supervision, Funding acquisition, Data curation, Review & editing.

3. L. Howell, M. Waterston, T. M. Wah, J. H. Chandler, and J. R. McLaughlan, "A Unified Platform and Quality Assurance Framework for 3D Ultrasound Reconstruction with Robotic, Optical, and Electromagnetic Tracking". Reproduced in Chapter 4. This work has been submitted to IEEE Transactions on Medical Robotics and Bionics for possible publication.

   - **LH**: Conceptualisation, Methodology (Lead), Data curation, Software, Investigation, Analysis, Visualisation, Writing manuscript, Review & editing.
   - MW: Methodology, Review & editing.
   - TW, JC, JM: Conceptualisation, Methodology, Supervision, Funding acquisition, Data curation, Review & editing.

4. L. Howell, M. Waterston, A. Zarei, T. M. Wah, J. H. Chandler, and J. R. McLaughlan, "AI-Assisted Co-Robotic Ultrasound for Three-Dimensional Thyroid Imaging and Volumetry", reproduced in Chapter 5. This work has been submitted to IEEE Transactions on Medical Robotics and Bionics for possible publication.

   - **LH**: Conceptualisation, Methodology (Lead), Data curation, Software, Investigation, Analysis, Visualisation, Writing manuscript, Review & editing.
   - MW, AZ: Methodology, Review & editing.
   - TW, JC, JM: Conceptualisation, Methodology, Data curation, Supervision, Funding acquisition, Review & editing.

Other publications not included in the thesis:

1. L. J. Howell, V. W.-S. Chan, J. Zhong, H. H.-L. Ng, J. H. Chandler, J. R. McLaughlan, and T. M. Wah, "Histotripsy: New Frontiers for Noninvasive Ablation", The Arab Journal of Interventional Radiology, vol. 9, p. 61, Dec. 2025.

2. H. H. L. Ng, V. W.-S. Chan, L. Howell, T. Shiwani, J. Zhong, J. Brandon, A. Samson, J. Chandler, J. Mclaughlan, and T. M. Wah, "Early Adoption of Image-Guided Histotripsy Therapy in Interventional Oncology: Challenges and Opportunities in the UK", British Journal of Radiology, p. tqag047, Mar. 2026.

3. A. Syversen*, O. Umney*, L. Howell*, J. Breen, E. Briggs, Z. Hancox, S. Khan, O. Mills, V. Moglia, M. Paterson, and R. Stephens, " "How can we involve Patients?" - Students' perspectives on embedding PPIE into a doctoral training centre for AI in medical diagnosis and care", Research Involvement and Engagement, vol. 11, p. 77, July 2025. *Aron Syversen, Oliver Umney and Lewis Howell contributed equally to this work.

Datasets:

1. J. R. McLaughlan, L. Howell, and N. Ingram, "Lung ultrasound COVID phantom dataset used for training machine learning model".

This thesis is submitted as an *alternative style of doctoral thesis*, based around a coherent body of published/publishable peer-reviewed articles (Chapters 2-5) that constitute the primary research outputs. Each study contributes to the overarching research objectives, demonstrating originality, critical ability, and research quality. This reflects the author's continued desire to contribute to open science through the development of open-source software and dissemination in open-access peer-reviewed publications.

The thesis includes an introductory chapter that outlines the research context and objectives; four original articles; and a final discussion and conclusion chapter that brings together the findings and critically evaluates the contributions, limitations, and directions for future work.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement

The right of Lewis James Howell to be identified as Author of this work has been asserted by them in accordance with the Copyright, Designs and Patents Act 1988.

Lewis James Howell

August 2026

Acknowledgements

Firstly, I would like to thank my supervisors, Jim McLaughlan, James Chandler, and Tze Wah, for all their support, advice, and encouragement throughout the project. You have been the best supervisors I could have hoped for. I have learned so much from you, I admire you, and I am grateful for everything you have done to help me through the PhD.

I would like to thank the CDT Directors, David Hogg, Vania Dimitrova, and Owen Johnson for making the CDT programme so great. I am particularly grateful for the opportunities to travel and present at conferences (Les Houches, Monte Carlo, Vienna, and Banff) and would like to acknowledge UKRI and the British taxpayer for funding my research.

I would also like to thank the clinical collaborators who gave up so much of their valuable time to work with me on the papers in this thesis. In particular, Amir Zarei, with whom I spent many hours at LGI and St James' working on the LTUS dataset, as well as Shishir Karthik, Jim Zhong, Amy Barnes, Vinson Chan, Helen Ng, and Zara Court. I would also like to thank Richard Stephens and all the patients who shared their thoughts on my project, as well as those whose data were used in this research.

I would like to thank my peers in the CDT, particularly Ben Keel for raising me up to new heights, Gerado Loza for long walks and good food, and Chris Winder for all the games of Azul. Thanks to Bailey Andrew and Taz Moglia for making me laugh and bringing joy to the office over the last few months. Thanks also to the other Cohort 3 heroes, Allan Pang, Aron Syverson, Morgan Thomas, Oliver Umney, and Zuzanna Wojcik. Thanks to Jack Breen for giving solid PhD advice and being a great board-game antagonist, and Mary Paterson for being a great friend with a keen eye for Canadian deer. I would also like to thank Siân Carey for walking with me at Eccup during my first week, and Alex Coles for his enthusiasm about dinosaurs and other such topics. Finally, a big thanks to Manisha Waterston for being a great PhD little sister - you've got this!

To my Mum, Dad, and Laura, thanks for all the love, encouragement, and kindness throughout my life. I am so grateful for everything. To my Grandpa, thank you for always being kind and inspiring me to keep going. Special thanks to my close friends Matt Field, Cam Haddow, Eleanor Osborne, and Alice Franklin - you guys are the absolute best. Finally, to my enduring partner, Zara Court. Thank you so much for always being there. Your love and support mean more than you know.

Abstract

High Intensity Focused Ultrasound (HIFU) is a non-invasive, non-ionising therapeutic technology that uses acoustic energy to ablate tumours at depth in the body. Compared with traditional surgery, it has the potential to reduce recovery time, minimise side-effects, and improve patient safety. However, challenges in treatment planning and imaging guidance have limited its widespread adoption.

This thesis presents an AI-integrated, ultrasound-guided HIFU system to address these challenges, combining real-time multi-organ segmentation with collaborative robotics (co-robotics) for patient-specific planning and guidance. These advances are demonstrated in the context of benign symptomatic thyroid nodules; however, the core methods and results developed could be broadly applied to a variety of treatment planning and ultrasound imaging tasks.

Initial work with lung ultrasound phantoms established lightweight Deep Learning (DL) models for live B-mode segmentation (mDSC = 0.74, 33 FPS), showing promise for future diagnostic systems. These methods were then extended to enable a co-robotic 3D ultrasound system, which achieved state-of-the-art geometric reconstruction accuracy on a purpose-designed quality assurance phantom (DSC-3D = 0.94 ± 0.01, HD95 = 1.17 ± 0.12 mm), approaching the spatial resolution limits imposed by ultrasound imaging.

In thyroid ultrasound, the co-robotic 3D ultrasound system allowed live multi-organ reconstruction and thyroid volumetry which closely matched the CT ground-truth (37.7 ± 1.3 mL vs. 37.9 ± 1.8 mL). This outperformed the current clinical standard ultrasound technique, with implications for radioiodine dosimetry and future semi-automated HIFU planning.

Finally, to support clinical translation, a new UK ultrasound dataset was curated, enabled by a novel data extraction tool which extracted key diagnostic information from free-text radiology reports (micro-average sensitivity, specificity, and F1-score of 0.97, 0.96, and 0.94).

Overall, this work represents progress towards addressing key barriers to wider HIFU adoption, including long treatment times, complex manual planning, and inter-operator dependence. Addressing these will be key to HIFU's role in improving patient outcomes in the future.

Contents

Intellectual Property and Publication Statements i

Acknowledgements iv

Abstract v

Contents vi

List of Figures ix

List of Tables xii

List of Acronyms xiii

1 Introduction 1

1.1 Plain English Summary . . . 1
1.2 Thesis structure . . . 1
1.3 Contributions . . . 2
1.4 Overview . . . 3
1.5 Background . . . 6
1.5.1 Ultrasound imaging . . . 6
1.5.2 Ultrasound therapy . . . 6
1.5.3 Treatment planning and guidance for HIFU . . . 7
1.5.4 Clinical application: Thyroid Nodules . . . 9
1.5.5 State of Clinical Practice . . . 10
1.5.6 AI for ultrasound guidance . . . 11
1.5.7 3D ultrasound for treatment planning . . . 12
1.5.8 Metrics for geometric similarity . . . 14

2 AI for real-time ultrasound segmentation 16

2.1 Introduction . . . 18
2.2 Methods . . . 21
2.2.1 Image Datasets . . . 21
2.2.2 Models and Training . . . 22

2.3 Results . . . 25
2.4 Discussion . . . 31
2.5 Conclusion . . . 36
2.6 Acknowledgements . . . 36

3 Curating a UK thyroid ultrasound dataset 44
3.1 Introduction . . . 46
3.2 Methods and materials . . . 48
3.2.1 Dataset . . . 48
3.3 RADEX tool . . . 49
3.3.1 Task formulation . . . 50
3.3.2 Develop search strategy . . . 51
3.3.3 Pre-processing . . . 52
3.3.4 Apply search strategy . . . 52
3.3.5 Evaluation . . . 52
3.3.6 Refine search strategy, application, and testing . . . 53
3.4 RADEX for thyroid ultrasound reports . . . 53
3.5 Results . . . 55
3.6 Discussion . . . 57
3.7 Acknowledgements . . . 60
3.8 Funding . . . 60

4 Ultrasound 3D reconstruction and quality assurance 66
4.1 Introduction . . . 68
4.2 Methods . . . 69
4.2.1 Tracked Image Acquisition . . . 69
4.2.2 Calibration and registration . . . 71
4.2.3 Segmentation and 3D reconstruction . . . 72
4.2.4 Evaluation and Quality Assurance . . . 72
4.3 Results . . . 75
4.3.1 System Performance . . . 75
4.3.2 Geometric phantom . . . 75
4.4 Discussion and Conclusion . . . 81
4.5 Acknowledgments . . . 83

5 Co-robotic 3D reconstruction for thyroid ultrasound 88
5.1 Introduction . . . 90
5.2 Methods . . . 93
5.2.1 System overview . . . 93
5.2.2 Segmentation and 3D reconstruction . . . 93
5.2.3 Thyroid volumetry and quantitative analysis . . . 96
5.2.4 Measuring movement smoothness . . . 96

5.3 Results . . . 97
5.4 Discussion and Conclusion . . . 104
5.5 Acknowledgments . . . 107

6 Discussion **114**
6.1 Implications . . . 114
6.2 Limitations . . . 116
6.3 Future directions . . . 118
6.3.1 Exposure placement . . . 118
6.3.2 Thermal dose estimation . . . 121
6.3.3 Robotic-guided ablation . . . 121
6.3.4 Patient motion and real-time feedback . . . 122
6.3.5 Dataset curation . . . 124
6.3.6 Clinical translation . . . 124

7 Conclusions **126**

A Appendix for Chapter 2: AI for real-time ultrasound segmentation **127**
A.1 Comparison with the state-of-the-art . . . 127

B Appendix for Chapter 4: Ultrasound 3D reconstruction and QA **129**
B.1 Supplementary material . . . 129
B.1.1 US-QA-3D phantom fabrication . . . 129

C Ultrasound spatial and temporal calibration **131**
C.1 Experimental setup for 3D reconstruction . . . 134

D Thyroid ultrasound datasets **136**
D.1 Overview of existing datasets . . . 137

E The Leeds thyroid ultrasound dataset **140**
E.1 Thyroid nodule grading . . . 142
E.2 Cytology/pathology . . . 142
E.3 Imaging metadata . . . 146
E.4 Inter-annotator study . . . 147

Bibliography **153**

List of Figures

1.1 Segmented B-mode ultrasound image of the thyroid 5
1.2 Ultrasound-guided High-Intensity Focused Ultrasound (USgHIFU) interface for the Theraclion EchoPulse system 8
1.3 Schematic of robotic ultrasound-guided high intensity focused ultrasound (USgHIFU) for treatment of thyroid nodules. 10
1.4 Dice Similarity Coefficient and Hausdorff Distance 15

2.1 Artefacts in lung ultrasound 19
2.2 Example images and augmentations used to train LUS model 23
2.3 Workflow diagram showing real-time lung ultrasound segmentation with U-Net 24
2.4 Training curves and predictive performance of models trained for lung ultrasound segmentation 26
2.5 Example LUS phantom images and overlaid segmentation masks 27
2.6 Confusion matrix for semantic segmentation of LUS phantom images 28
2.7 Calculation of the B-line Artefact Score (BLAS) 29
2.8 Bland-Altman plot for BLAS, comparing manual and automatic methods . 30
2.9 Real-time segmentation of features in a lung ultrasound phantom (video) . . 31
2.10 Clinical LUS segmentation examples 32

3.1 Typical workflow for the Radiology Data Extraction Tool (RADEX) 51
3.2 Synthetic neck and thyroid ultrasound report, highlighted using a rule-based radiology data extraction tool 55
3.3 F1-score for Radiology Data Extraction Tool (RADEX) predictions on the test clinical dataset, before and after search refinement 56
3.4 Class-wise confusion matrices for multilabel data 57

4.1 Technical diagram showing hardware devices (grey) and software modules (blue) included as part of the experimental system used for real-time 3D US reconstruction 70
4.2 Experimental setup for experiments with US-QA-3D phantom, showing the robotic arm, ultrasound transducer, tracking assembly, water tank with geometric phantom, optical cameras, and electromagnetic field generator . . . 73

4.3 Thresholding-based segmentation of objects in the Ultrasound Quality Assurance in 3-Dimensions (US-QA-3D) phantom . . . 77
4.4 Dual 3D view of live real-time volumetric reconstruction of robotic-tracked US in the US-QA-3D phantom . . . 78
4.5 3D reconstruction of objects in US-QA-3D phantom for different tracking modalities (robotic, optical, and electromagnetic) . . . 80

5.1 System for co-robotic 3D reconstruction in a thyroid anatomical phantom (CIRS, Model 074) using a linear US transducer (GE Healthcare, L4-20T-RS) mounted on a 6-DoF collaborative robot arm (Universal Robots, UR3e). 92
5.2 Lightweight U-Net architecture for multiclass anatomical segmentation in the thyroid phantom. The input image is mapped to the output segmentation map via a convolutional encoder-decoder network . . . 94
5.3 AI-based segmentation of anatomy in the thyroid phantom . . . 98
5.4 Dual Three-dimensional (3D) view of real-time volumetric reconstruction of co-robotic Ultrasound (US) in a thyroid training phantom . . . 99
5.5 AI-assisted co-robotic tracked US 3D reconstruction and visualisation in thyroid phantom . . . 100
5.6 3D render and cross-sectional images of the thyroid and major vessels in an anatomical phantom from MPR . . . 101
5.7 Reference CT imaging of thyroid phantom (CIRS Model 074), showing axial, coronal, and sagittal views . . . 103
5.8 Log dimensionless jerk and spectral arm length for single-sweep scans of thyroid phantom using co-robotic navigation (active and passive) and freehand navigation (optical tracking) . . . 104

6.1 Schematic showing research avenues in Artificial Intelligence (AI)-integrated USgHIFU. . . . 119
6.2 HIFU exposure placement (raster packing) for a thyroid nodule based on 3D reconstruction . . . 120
6.3 Robotic HIFU ablation based on 3D US anatomy . . . 122

A.1 Comparison of the Dice similarity coefficient for various U-Net variants . . . 127
A.2 Comparison of the benchmark framerate for inference of various U-Net variants 128

B.1 US-QA-3D phantom construction . . . 130

C.1 Coordinate definitions for ultrasound spatial calibration . . . 132
C.2 Experimental setup for ultrasound spatial calibration using the N-wire phantom technique . . . 133
C.3 Ultrasound temporal calibration between fixed and moving signals . . . 134
C.4 Experimental setup for US-QA-3D experiments . . . 135

D.1 Data annotation types for thyroid ultrasound 137

E.1 Boxplot for BTA and ACR TI-RADS thyroid nodule gradings. 144
E.2 Total number of nodules and sample-confirmed nodules grouped by BTA grade in the LTUS dataset. 145
E.3 Value counts for cytology and pathology labels for thyroid nodules in the LTUS dataset 145
E.4 TNM staging for malignant thyroid tumours in the LTUS dataset. 146
E.5 Metadata labels for the LTUS dataset 147
E.6 Example segmented images from the inter-annotator study. 149
E.7 Matrices showing paired DSC for inter-annotator study 151

List of Tables

2.1 Summary of model performance for segmentation in the LUS phantom . . . 26

3.1 Search strategy for neck and thyroid ultrasound dataset 54
3.2 Scores for multilabel classification of neck and thyroid ultrasound reports with regular expressions. 56

4.1 System performance, including timing and synchronisation, calibration and registration, and tracking accuracy 75
4.2 Baseline scans of the US-QA-3D phantom, using simultaneous tracking with robotic kinematic, optical and electromagnetic (EM) tracking 79

5.1 CNN structure for real-time multiclass thyroid segmentation 95
5.2 Measurements of the phantom anatomy, comparing baseline values (B-mode and CT) with AI-based 3D US reconstruction under co-robotic (active and passive guidance) and freehand (optical and EM tracked) guidance 102

D.1 Overview of thyroid ultrasound datasets 138
D.2 Suitability of thyroid ultrasound datasets 138

E.1 Summary of the Leeds Thyroid Ultrasound dataset 140
E.2 Characterisation of the Leeds Thyroid Ultrasound dataset 141
E.3 Sonographic features used for calculating Thyroid Imaging Reporting and Data System (TI-RADS) and the RADEX search terms used 143

List of Acronyms

2D Two-dimensional

3D Three-dimensional

ACR American College of Radiologists

AI Artificial Intelligence

API Application Programming Interface

AR Augmented Reality

B-mode Brightness Mode

BERT Bidirectional Encoder Representations from Transformers

BLAS B-line Artefact Score

BTA British Thyroid Association

CAD Computer-Aided Design

FNA Core Needle Biopsy

CNNs Convolutional Neural Networks

CPU Central Processing Unit

CRIS Computerised Radiology Information System

CT Computed Tomography

DH Denavit-Hartenberg

DL Deep Learning

DoF Degrees of Freedom

DRL Deep Reinforcement Learning

DSC Dice Similarity Coefficient

EM Electromagnetic

EMI Electromagnetic Interference

ETA European Thyroid Association

FDA Food and Drug Administration

FN False Negative

FNA Fine Needle Aspiration

FOV Field of View

FP False Positive

FPS Frames per Second

GPU Graphics Processing Unit

grad-CAM Gradient-weighted Class Activation Maps

GUI Graphical User Interface

HD Hausdorff Distance

HD95 95th Percentile Hausdorff Distance

HDMI High-Definition Multimedia Interface

HIFU High-Intensity Focused Ultrasound

ICP Iterative Closest Point

LAT Laser Ablation Therapy

LDLJ Log Dimensionless Jerk

LLMs Large Language Models

LTHT Leeds Teaching Hospital Trust

LTUS Leeds Thyroid Ultraosund Dataset

LUS Lung Ultrasound

MI Mechanical Index

ML Machine Learning

MNG Multinodular Goitre

MPR Multiplanar Reconstruction

MRI Magnetic Resonance Imaging

MRML Medical Reality Modelling Language

MWA Microwave Ablation

NHS National Health Service

NICE National Institute for Health and Care Excellence

NLP Natural Language processing

OCR Optical Character Recognition

PACS Patient Archive and Communication System

PEIT Percutaneous Ethanol Injection Therapy

PINNs Physics Informed Neural Networks

QA Quality Assurance

RADEX Radiology Data Extraction Tool

RAM Random Access Memory

RCT Randomised Control Trial

Regex Regular Expressions

RFA Radiofrequency Ablation

RMS Root mean square

ROI Region of Interest

RUS Robotic ultrasound

SAM Segment Anything Models

SPARC Spectral Arc Length

TGC Time Gain Compensation

TI-RADS Thyroid Imaging Reporting and Data System

TNM Tumour Nodes Metastasis

TP True Positive

UK United Kingdom

UMLS Unified Medical Language System

US Ultrasound

US-QA-3D Ultrasound Quality Assurance in 3-Dimensions

U.S. United States

USB Universal Serial Bus

USgHIFU Ultrasound-guided High-Intensity Focused Ultrasound

VIA VGG Image Annotator

VTK Visualisation Toolkit

Chapter 1

Introduction

1.1 Plain English Summary

High-Intensity Focused Ultrasound (HIFU) is an exciting technology which uses ultrasound waves to treat cancer and other tumours. Like a magnifying glass, HIFU focuses ultrasound waves down onto a small point, where the energy becomes strong enough to destroy the tumour. Because the skin is not cut open, patients usually experience less pain, a lower risk of infection, and faster recovery than with surgery. HIFU is also free of radiation and can be performed without general anaesthesia.

One of the main challenges is in planning and guiding the treatment. This research shows that AI could help the doctor to see the tumour more clearly before and during treatment. This could make planning for HIFU and other treatments faster, safer, and easier. It could also create a 3D view of the patient's organs and blood vessels to help plan treatments. Using collaborative robots that work alongside the doctor could help to move the ultrasound more precisely and keep it stable during scanning.

We hope these improvements will help make HIFU available to more patients in the future. This is only possible thanks to patients who donate their data and take part in research, which is critical to developing HIFU technologies.

1.2 Thesis structure

This thesis includes published material and material under review for publication in the areas of treatment planning and imaging guidance, and AI for ultrasound segmentation and 3D reconstruction. These are critical areas for development to improve diagnostic ultrasound and the delivery of minimally invasive diagnostic and non-invasive therapies such as USgHIFU, as discussed in the context of thyroid nodule treatment.

Chapter 1: Introduces the concepts of treatment planning and imaging guidance for minimally-invasive therapies, such as HIFU, discussing limitations in existing systems that

have affected clinical uptake. Provides background on minimally invasive therapy options to treat benign symptomatic thyroid nodules, and discusses how advances in AI and robotics could enhance ultrasound planning and guidance more generally.

Chapter 2: Establishes the Deep Learning (DL) models and training schemes applied to real-time ultrasound segmentation in subsequent chapters. This work examines segmentation in the context of lung ultrasound, a dynamic artefact-based technique that highlights challenges such as inter-annotator variability and the need for high-quality annotated datasets, informing the following chapters. The paper introduces a method to help address the existing skill gap in lung ultrasound and provides a novel metric for assessing the severity of interstitial lung disease.

Chapter 3: Introduces a tool for clinical data extraction, which was applied to curate a large dataset of thyroid ultrasound images and associated radiological findings. This dataset informed subsequent research methods and provides an avenue for future research. The tool offers a time-saving, freely available option to accelerate structured data collection in research and audit applications, where access to labelled data or computing infrastructure is limited, or where explainability and reproducibility are priorities.

Chapter 4: Describes a flexible platform for tracked ultrasound 3D reconstruction and a Quality Assurance (QA) framework to assess geometric fidelity, comparing robotic kinematics with optical and electromagnetic tracking. Developing this system was a major output of the doctoral research and enabled thyroid ultrasound reconstruction in Chapter 5. The work provides an analysis of the system capabilities, including calibration, synchronisation, tracking accuracy, and 3D reconstruction. A custom-designed phantom with geometric inclusions was proposed for more objectively assessing 3D reconstruction accuracy.

Chapter 5: Presents multi-organ segmentation and 3D reconstruction of the thyroid and major vessels in an anatomical phantom. This is the first co-robotic system for thyroid volumetry, which is shown to be more accurate than the clinical standard approach. This provides a feasible approach to pre-treatment planning for HIFU therapy of thyroid nodules, addressing the challenges set out in Chapter 1.

Chapter 6: Discusses the developments in imaging-guidance technologies presented in previous chapters, including the implications for research, limitations of current work, and future directions towards clinical translation. In particular, other areas where AI could improve HIFU planning and guidance are discussed, including exposure placement, thermal dose estimation, robotic-guided ablation, motion correction, and real-time feedback.

Chapter 7: Concludes the thesis with a discussion of the main contributions and impacts for USgHIFU and beyond.

1.3 Contributions

The main contributions arising from the work are summarised as follows:

- Lightweight DL models for real-time ultrasound multiclass segmentation, demonstrated on lung ultrasound and later on thyroid ultrasound.
- Novel DL training schemes for segmentation, using ultrasound-specific augmentations to improve generalisability and reduce the reliance on large datasets.
- An open-source radiology data extraction tool which enables clinicians to build structured datasets more quickly, demonstrated on thyroid ultrasound reports.
- The largest thyroid ultrasound imaging dataset to be made available for research, with accompanying structured radiological, histological, and pathology information.
- A flexible tracked ultrasound 3D reconstruction system built using open-source tools and applicable to a variety of freehand and robotic tracked ultrasound applications in diagnostic and interventional medicine.
- A novel quality assurance pipeline and phantom designed specifically for assessing 3D ultrasound reconstruction, aimed at improving the quality of reporting in future tracked ultrasound research.
- The first co-robotic ultrasound system for AI-assisted thyroid ultrasound 3D reconstruction, which improves the accuracy of thyroid volumetry compared with the clinical standard technique, and could be applied to treatment planning in HIFU.

1.4 Overview

Minimally- and non-invasive therapies offer a promising modern alternative to traditional open surgery. These procedures rely on accurate pre-operative planning and intra-operative image guidance to safely target diseased tissue while avoiding surrounding healthy anatomy. As image-guided interventions become increasingly precise and less invasive, there is a growing need for imaging systems that can provide accurate spatial information, support clinical decision-making, and enable reliable treatment delivery.

A number of minimally invasive technologies have been used for thermal ablative therapy, including Radiofrequency Ablation (RFA), Microwave Ablation (MWA), and HIFU. In all cases, the safety and efficacy of these therapies depend on accurate localisation of the treatment target and reliable positioning of the ablation zone. Advances in medical imaging and robotic navigation have improved control over therapy delivery [1], but imaging guidance remains one of the principal challenges for many image-guided interventions [2].

To date, Magnetic Resonance Imaging (MRI) has achieved the most clinical success in imaging guidance [3], largely because MRI provides high-contrast, 3D imaging and simpler

registration, in which the patient and robot are automatically registered to a common coordinate frame (the MRI coordinate frame). MRI also provides more advanced thermometry for HIFU monitoring, enabling near-real-time quantitative assessment of tissue temperature, precise control of thermal dose deposition, assessment of off-target heating, and improved understanding of lesion formation and ablation completeness during treatment [1]. However, MR-guided systems present both technical and procedural limitations. Systems generally have a reduced range of motion for therapy navigation, with fewer Degrees of Freedom (DoF), and rely on expensive specialised robotics with pneumatic, hydraulic, or piezoelectric actuators to be MRI compatible [1]. These robots also typically cannot operate simultaneously with imaging, as they would reduce imaging quality [4, 5]. Access to MRI machines also poses a major barrier in terms of the accessibility and widespread use of MR-guided systems, especially in low- and middle-income countries [6].

Therefore, ultrasound imaging guidance provides an attractive alternative owing to its portability, low cost, real-time imaging capability, and widespread clinical availability [7]. However, US-guided interventions present their own challenges. Conventional B-mode imaging provides limited soft-tissue contrast, contains speckle and imaging artefacts, and typically represents only a two-dimensional view of inherently three-dimensional anatomy. As a result, treatment planning and intra-operative guidance often rely heavily on operator experience. Current ultrasound-guided workflows typically require the clinician to manually identify anatomical structures and contour treatment targets across multiple image planes before reconstructing the treatment volume [8]. Treatment trajectories are then generated using relatively simple planning strategies that sequentially position treatment locations while incorporating cooling intervals between exposures to minimise unwanted tissue heating [9]. These workflows are time-consuming [10], increase operator workload, and may result in under-treatment of the target or unnecessary exposure of surrounding healthy tissue [11]. Similar challenges are encountered across ultrasound-guided interventions, where limited anatomical visualisation and the absence of volumetric information make accurate planning and guidance difficult.

AI has the potential to address many of these limitations by improving the interpretation of ultrasound images and supporting treatment planning and guidance. AI has been widely investigated for automating tasks that require human cognition and decision making [12], while DL has achieved particular success in medical image analysis by learning hierarchical feature representations directly from imaging data [13, 14]. Architectures such as U-Net [15] and its variants have become the state of the art for biomedical image segmentation, including ultrasound imaging [16]. Integrating automated segmentation with tracked ultrasound enables volumetric reconstruction of patient anatomy (Figure 1.1), providing richer anatomical information for treatment planning, facilitating the identification of critical structures, and supporting patient-specific intervention strategies.

In particular, HIFU is one example of a non-invasive image-guided therapy that illustrates

these challenges. Over the last decade, several clinical trials have demonstrated its potential for treating benign and malignant tumours in the breast, liver, pancreas, uterus, bone, kidney, prostate, thyroid, and brain [17]. However, despite this promise, wider clinical adoption has been limited by challenges in treatment planning and image guidance, particularly when using ultrasound imaging for planning and guidance.

In this thesis, we investigate technologies to improve ultrasound-based treatment planning and image guidance for minimally invasive interventions, including co-robot-assisted US-gHIFU for benign thyroid nodules as an example of a clinical application with a need for innovation. This work addresses key technical challenges, such as real-time ultrasound segmentation, limited data availability, inter-operator variability, and the need for accurate 3D reconstruction for planning. This research primarily focuses on imaging guidance for thermal-HIFU; however, developments in AI and tracked US are likely to benefit a variety of diagnostic and interventional applications, including volumetry, surgical navigation, and treatment planning for other minimally invasive therapies such as radiofrequency and microwave ablation. Ultimately, the development of ultrasound-guidance technologies will enhance patients' quality of life by improving diagnostic accuracy and delivery of minimally invasive therapies, improving treatment options for challenging tumour sites such as the thyroid, prostate, liver, kidney, and pancreas.

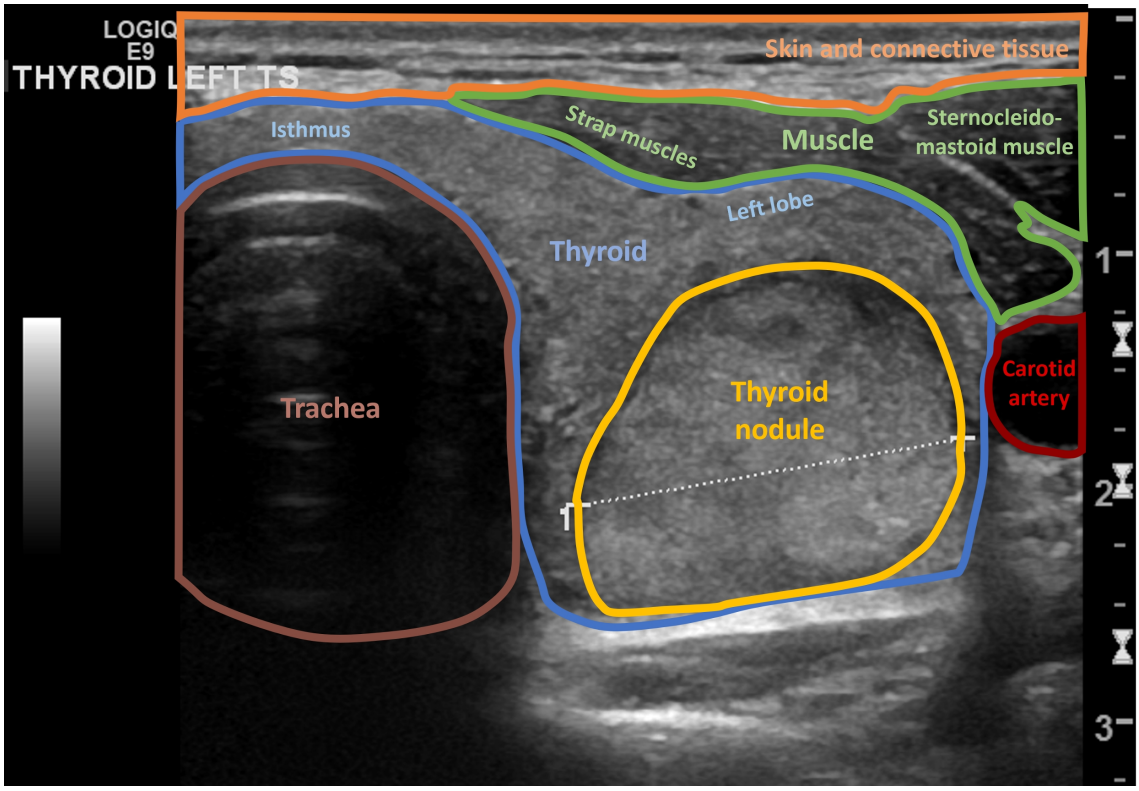


Figure 1.1: Segmented B-mode ultrasound image of the thyroid, showing an isolated isoechoic thyroid nodule in the left posterior thyroid. Also annotated are the trachea, anterior neck muscles, skin, and carotid artery.

1.5 Background

1.5.1 Ultrasound imaging

Ultrasound imaging, or sonography, is a non-invasive diagnostic technique that uses high-frequency sound waves to generate real-time images of internal structures. Its core applications include examination of internal organs, fetal development, the heart and blood vessels using Doppler imaging, superficial soft tissues, and musculoskeletal structures. Ultrasound is also widely used to guide biopsies, injections, vascular access, and fluid aspiration [18].

The core advantages of ultrasound include its absence of ionising radiation, relatively low cost, portability, and ability to provide dynamic imaging at the point of care. These characteristics make it suitable for repeated examinations and for use in pregnancy, paediatrics, emergency medicine, and resource-limited settings. In addition, Doppler ultrasound provides information about blood flow, while real-time imaging enables the guidance of minimally invasive procedures [19].

Despite these benefits, ultrasound has important limitations. Image quality and correct interpretation are strongly dependent on operator expertise, and sound waves are attenuated or reflected by air and bone, limiting assessment of gas-filled organs and tissues located behind bone [19]. Ultrasound images may also contain artefacts and generally provide lower spatial and contrast resolution than Computed Tomography (CT) or MRI, especially for deep structures [1].

1.5.2 Ultrasound therapy

Beyond imaging, ultrasound can also be used in therapeutic applications. In HIFU, ultrasound waves are transmitted through the skin and intermediate tissue and focused onto a precise target location, where the energy is converted into thermal or mechanical bioeffects. In thermal-HIFU, the predominant mechanism for ablation is coagulation necrosis [20], in which tissue temperatures of approximately 60-80 °C are reached within 1-4 s, causing localised cell death [10]. More recently, histotripsy has emerged as a mechanical form of focused ultrasound therapy that uses controlled acoustic cavitation to destroy targeted cells [21]. In both approaches, the focal region is typically only a few millimetres in size, requiring multiple precisely positioned exposures to completely treat the target volume.

As the only truly non-invasive ablation technique, HIFU offers several unique patient benefits, including reduced risk of infection, shorter recovery time, less pain, and improved cosmetic outcomes compared with conventional surgery [22]. The therapy is also non-ionising and can often be performed without general anaesthesia, making it suitable for patients at high surgical risk [23] or for treating radio-recurrent tumours [24]. However, these benefits can only be fully realised when treatment planning and guidance are sufficiently accurate to ensure safe and effective energy delivery [25].

Unlike surgery, where the tumour is physically removed, thermal ablation techniques destroy the tissue *in situ*, whereby it can then be absorbed by the body [26]. At the focal point of thermal-HIFU, the tissue is irreversibly damaged by extreme hyperthermia (increased temperature), resulting from the conversion of mechanical ultrasound energy to heat by viscous absorption, where pressure fluctuations cause tissue shearing at the macroscopic level [27].

These physical effects translate into biological effects in tissue, depending on the type and duration of exposure. Below 41°C, the damage caused to tissue is mostly reversible regardless of duration, where vasodilation and increased blood perfusion protect the cells from injury [28]. Between 42-60°C, irreversible cellular damage can occur given a great enough exposure time [29]. Temperatures exceeding 60 °C cause cell death almost instantly from protein denaturation, membrane rupture, cell shrinkage, pyknosis, hyperchromasia, and coagulation necrosis [28, 30]. Above 100 °C, undesirable physical effects such as vaporisation and carbonisation can occur, which limit the propagation of ultrasound and reduce the efficiency of energy delivery [26, 31]. For these reasons, temperatures of 60-80° C are usually targeted, and HIFU treatment requires careful planning.

Clinically, the most widely used measure of tissue destruction by thermal ablation is the thermal dose in cumulative equivalent minutes [29]. This measure reduces the temperature-time history to an equivalent exposure at 43° C, CEM_{43}, given by the equation

$$\mathrm{CEM}_{43}(t) = \int_{t=t_0}^{t=t_{\mathrm{end}}} R^{43-T(t)} dt \tag{1.1}$$

where R = 0.5 above 43° C and R = 0.25 otherwise [30]. The lethal thermal dose required for ablation is defined as the thermal dose equivalent of a 240-minute exposure at 43° C [29]. This model provides a useful estimate of the time needed for each HIFU exposure, taking 10 s at 53°C, 1 s at 57°C, and 0.1 s at 60°C [30].

1.5.3 Treatment planning and guidance for HIFU

The development of HIFU systems has been a topic of significant research, with multiple commercial thermal-HIFU systems available [9, 32], including Sonablate (Focus Surgery, Inc., Indianapolis, United States (U.S.)), Ablatherm (EDAP, Lyon, France), ExAblate (InSightec Ltd., Tirat Carmel, Israel), and Haifu JC (Chongqing (HAIFU) HIFU Technology Co. Ltd., China). These have gained clinical approval for treatment sites including the liver, prostate, pancreas, and brain, as well as uterine fibroids. For liver tumours, the Edison system (Histosonics, U.S.) has been clinically approved and is currently the only commercial system using histotripsy [33]. At present, the only commercially approved device for the treatment of thyroid nodules is the EchoPulse (Theraclion, Paris, France) system [8, 32], which has proven safe and effective in a limited number of clinical trials [7, 10, 34]. However, several key challenges have limited the widespread clinical uptake

of these systems, namely long treatment times, operator-dependent efficacy, and difficulty verifying outcomes.

In HIFU, the focal zone for thermal ablation is small (about the size of a grain of rice), enabling precise delineation of treated tissue, but at the cost of requiring many HIFU exposures to treat the entire tumour region [35]. In current USgHIFU systems, including Sonablate, the pre-treatment process is tedious and time-consuming as it requires an experienced physician to manually segment the tumour target, safety margins, and vital structures in several planes [8]. This must be precise and anatomically accurate to avoid undertreatment of the tumour or overtreatment of healthy tissue. The computer then defines the proposed locations for each HIFU exposure, which the operator then checks (Figure 1.2).

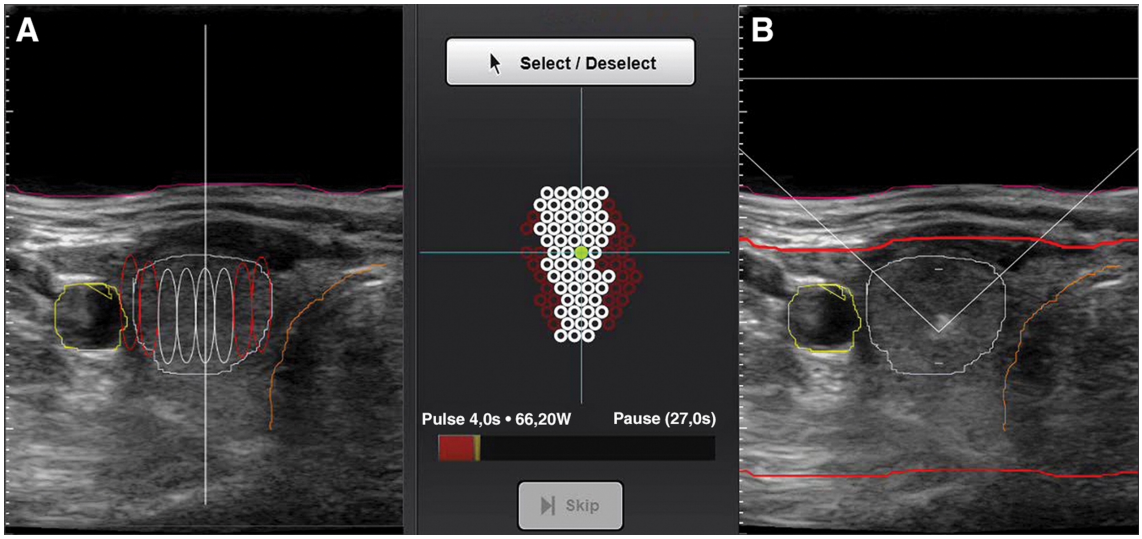


Figure 1.2: USgHIFU interface for the Theraclion EchoPulse system, planning treatment for a thyroid nodule. The treatment plan (central panel) shows a top view of the HIFU lesion locations based on manually segmented ultrasound images. A) Anatomical segmentations are manually defined by the clinician in several planes, including the target nodule (white), carotid artery (yellow), and trachea (orange). Lesion locations from the treatment plan are overlaid (white ovals), and those which fall outside of the safety margins are excluded (red ovals). B) Hyperechoic marks can be seen at the focal point after the first sonication. Figure adapted from Kovatcheva et al., "Benign Solid Thyroid Nodules: US-guided High-Intensity Focused Ultrasound Ablation—Initial Clinical Outcomes", Radiology, 2015, reproduced in accordance with the CC BY 4.0 license. [11]

Another challenge of USgHIFU is the high positioning accuracy and stability required to deliver the therapy plan, making outcomes of freehand systems highly operator-dependent [36]. Previous work has shown that robotics can increase the stability and positioning accuracy of USgHIFU systems, account for patient motion, and enable HIFU delivery with sub-mm precision [37]. Early efforts include work by Chauhan and ter Haar, who developed a system called FUSBOT with a reported navigation error of <0.5 mm [38]. More recently, a variety of robotic HIFU systems have been developed with different robot configurations,

degrees of freedom, and tracking methods [37, 39–41].

Each exposure can be triggered automatically or by the operator, and constant monitoring of imaging guidance is needed to ensure safety. Each exposure typically lasts only a few seconds but is followed by a longer cooling period of 20-40 s to reduce the risk of boiling [10]. In practice, these steps lead to long treatment duration, typically 45-60 min for a 3 cm nodule [10]. This presents a significant barrier to clinical acceptance [3], since RFA and MWA offer a simpler clinical workflow and take around 50% less time [42].

Research which addresses these challenges is crucial to the uptake of HIFU therapy in clinical practice, and software improvements in treatment planning and guidance have the potential to reduce the overall treatment duration and increase accessibility, enabling the treatment of larger lesions and enabling HIFU's clinical acceptance in practice [25].

1.5.4 Clinical application: Thyroid Nodules

Of the clinical targets for HIFU, thyroid nodules provide a challenging, but particularly relevant application where improvements in image-guidance and treatment planning could improve outcomes for a large number of patients. Thyroid nodules are very common, with between 19-68% incidence in the general population [7, 43], but around 85-95% of nodules are benign (not cancerous) [44]. Nevertheless, approximately 15-30% of benign nodules increase in volume over time [45], which can lead to enlargement of the thyroid and neck (known as goitre) [46]. Large nodules or multiple nodules (multinodular goitre) can cause local compressive symptoms or cosmetic problems [7, 47], requiring treatment to reduce their volume[25]. Since no pharmacological treatments have been proven to reduce the size of thyroid nodules, surgery is the primary treatment [44, 47], with common complications including bleeding, infection, damage to the recurrent laryngeal nerve, and hypoparathyroidism [48]. Furthermore, surgery is undesirable for the patient as it leaves a visible scar on the neck and a total thyroidectomy necessitates lifetime treatment with levothyroxine to treat hypothyroidism [8].

Since most thyroid nodules are benign (and not usually life-threatening), surgery is generally only justifiable as a last option for patients with symptomatic benign nodules. Hence, there has been significant interest in the development of organ-sparing technologies which reduce surgical risks while still providing the necessary symptomatic and cosmetic relief to patients [34]. Several minimally invasive interventional radiology techniques have been applied to volume reduction in benign thyroid nodules, including Percutaneous Ethanol Injection Therapy (PEIT), Laser Ablation Therapy (LAT), RFA, MWA, and HIFU [44]. Of these techniques, HIFU is the only non-invasive ablation method, with all other technologies requiring the clinician to precisely insert one or more needle-like devices into the treatment site. For extracorporeal USgHIFU, focused ultrasound energy is delivered from outside the body using a bowl-shaped therapy transducer, with co-axially aligned ultrasound imaging (Fig. 1.3). This transducer assembly is positioned using a robot arm and

creates sound waves which pass harmlessly through the skin and intermediate tissue before converging onto a focal point in the nodule to treat.

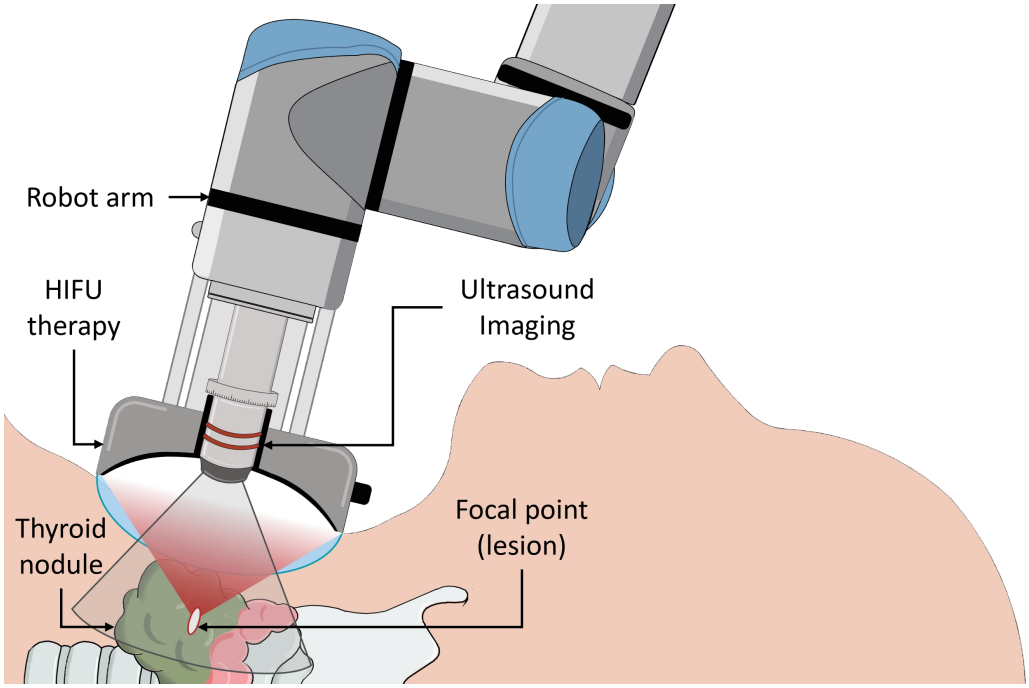


Figure 1.3: Schematic of robotic ultrasound-guided high intensity focused ultrasound (US-gHIFU) for treatment of thyroid nodules.

1.5.5 State of Clinical Practice

HIFU for treating symptomatic benign thyroid nodules is usually performed on an outpatient basis, with discharge 1–2 hours after treatment [8]. Generally, to be eligible for ablation, malignancy must be ruled out using Fine Needle Aspiration (FNA), and the nodule must be solid or primarily solid since fluid-filled (cystic) nodules reflect ultrasound. During therapy, the patient lies in a supine position with moderate neck extension and is typically given painkillers and procedural sedation [34, 49]. Focused ultrasound is delivered with simultaneous imaging of the treatment area. In clinical studies, the primary outcome measures for efficacy are usually the nodule volume reduction percentage and improvement in nodule-related symptoms, while safety is measured by the visual analogue scale for pain, and the incidence of other side effects is recorded [25].

Existing evidence on the efficacy of HIFU for treating benign thyroid nodules is mixed in quality and sample size, but HIFU has shown good efficacy in many patients. In a systematic review and meta-analysis, the thyroid nodule volume reduction after a single HIFU therapy session ranged from 45% to 68%, depending on nodule size and length of follow-up [10, 50]. In a case series study of 108 patients in Hong Kong, the mean ($\pm$ 1 standard deviation) volume reduction ratio was 51.32% $\pm$ 20.71% at 3 months, 68.66% $\pm$ 18.48% at 12 months and 70.41% $\pm$ 17.39% at 24 months [51]. Sequential ablation of multiple

nodules in multinodular goitre has also been considered, with a comparative study of 123 patients reporting efficacy comparable to that of single-nodule ablation [52]. Compared with surgery, HIFU has shown comparable efficacy in some studies, while others have reported higher efficacy for surgery; however, further studies are needed, given the diverse results [53]. HIFU has also been compared with radiofrequency and microwave ablation, finding significant volume reduction (44–50% after 3 months) with all three methods, but no statistically significant difference in efficacy [54].

Despite mixed efficacy results, HIFU therapy is widely considered very safe [25, 50]. Mild to moderate pain is the most frequently reported complication [55], for example, in a study of 146 patients, mean pain scores (visual analogue scale) were measured as 3.5 ± 4.0 during HIFU therapy, reducing to 1.0 ± 1.0 immediately after [56]. Other risks reported across clinical studies included minor skin redness, transient laryngeal nerve palsy, subcutaneous oedema, and, in rare cases, nausea, vomiting and Horner syndrome, where disruption in the sympathetic nerve supply can cause partial ptosis (drooping of the eyelid), miosis (contracted pupil) and facial anhidrosis (absence of sweating) [3, 34]. All these side-effects are non-serious and temporary, with full recovery over the hours and days following the therapy. [7].

HIFU also has a short recovery time, with many patients able to return home within a few hours of treatment [49]. In a study of 146 patients, comparing HIFU (43 patients) with hemithyroidectomy surgery (103 patients), the HIFU group had a significantly shorter hospital duration (0.3 vs. 1.0 day, $p < 0.001$), lower incidence of subclinical hypothyroidism (2.3% vs, 20.4%, $p = 0.008$), higher symptom improvement score (67% vs. 39%, $p = 0.009$), lower risk of voice quality problems ($p < 0.05$) and lower cost (USD 1923.1 vs. USD 5384.6) [56].

Despite promising results from early clinical trials, the most recent review by the United Kingdom (UK) National Institute for Health and Care Excellence (NICE) found that 'the evidence on the safety of high-intensity focused ultrasound for symptomatic benign thyroid nodules raises no major safety concerns; however, the current evidence on its efficacy is limited in quantity and quality' [25], recommending further research be conducted in thermal ablative technologies. At present, there have been no Randomised Control Trial (RCT)s, and limitations of existing studies include overlapping patient cohorts, limited follow-up duration, and small sample sizes. The same HIFU device (Echopulse, Theraclion) was also used in all studies, suggesting a need for new, innovative systems which address treatment challenges.

1.5.6 AI for ultrasound guidance

DL, a subset of AI and Machine Learning (ML), is a promising alternative to manual ultrasound segmentation for ultrasound planning and guidance [57]. DL models use hierarchical combinations of learnable feature extractors to automatically build high-level representa-

tions of data [14]. The most widely applied DL techniques for vision problems are based on convolutional neural networks (CNNs) [16], such as the U-Net architecture [15], which is often used for segmenting biomedical images. Recently, variants of the original U-Net, such as those that incorporate attention gates or vision transformers, have been shown to achieve strong results in semantic image segmentation tasks [58, 59].

Several recent studies have investigated the application of deep learning to segmentation of the thyroid gland [60–65] and thyroid nodules on ultrasound [61, 62, 64–67]. However, most work has focused on the diagnostic applications of such models, such as classification of benign and malignant nodules [67, 68]. None have attempted multi-object classification beyond these classes, and the real-time performance of these models is also rarely considered. To fully model HIFU propagation and heating in tissue, segmentation of the intermediate skin, muscle, and soft tissue layers is required. To define safety margins, delineation of vital structures such as the carotid artery, trachea, and tracheoesophageal groove (containing the recurrent laryngeal nerve) is also useful [8]. Ultrasound tissue mapping has been previously attempted. For example, in 2019, Cunningham *et al.* showed segmentation of 10 muscles, the ligamentum nuchae (ligament), vertebra, and skin in ultrasound images of the posterior cervical neck [69]. In 2022, Huang *et al.* [70] used CNNs with fuzzy logic blocks for breast ultrasound segmentation of fat, mammary tissue, muscle, and tumour. However, to date, this has not been attempted in thyroid images. This is likely due to a lack of annotated training data and other challenges, such as uncertainty in the ground-truth segmentation, which make such granular segmentation difficult.

1.5.7 3D ultrasound for treatment planning

Unlike MR-guided interventions, which provide 3D imaging, conventional US-guided interventions have relied on 2D imaging for planning and guidance. This limits options for volumetric analysis and makes US-guided applications dependent on the operator's skill, scanning technique, and geometric reasoning [71, 72]. One approach is fusion imaging with pre-operative CT or MRI [73]; however, this adds procedural complexity and poses challenges for image registration, including accounting for deformation and motion.

3D US could address this challenge [74], allowing cross-sectional imaging to be acquired immediately prior to therapy. This could be collected either using a specialised matrix-array transducer or using a 2D-tracked transducer with volumetric reconstruction.

While advances in matrix-array transducers have improved their feasibility and expanded their applications in obstetrics, gynaecology, and cardiovascular medicine, they have not yet seen widespread uptake in clinical practice due to system complexity, data throughput demands, and limitations in real-time rendering and analysis [75]. Compared with conventional linear or curvilinear 2D probes, they are also substantially higher in cost and often require compromise on either spatial resolution, Field of View (FOV), or frame rate and are widely incompatible with standard clinical systems [74].

On the other hand, tracked 3D US can leverage established, widely available high-frequency 2D transducers. The transducer is swept across the anatomy, and pose tracking is combined with the imaging stream to reconstruct a volumetric image [76]. This approach enables high-resolution, real-time 3D imaging over a large area without the need for specialist US systems [77, 78]. 3D US based on tracked imaging can also be readily integrated into treatment planning workflows using the same transducers and imaging systems as those used for real-time guidance, thereby avoiding registration challenges and simplifying the workflow. This approach combines conventional Two-dimensional (2D) ultrasound imaging with position and orientation tracking data to reconstruct 3D anatomy over large areas [74, 77], which is important in many planning and guidance settings.

To acquire the tracking data, optical and Electromagnetic (EM) tracking have been used in previous freehand diagnostic applications [77]. For example, Chung *et al.* demonstrated freehand optical-tracked US for carotid artery 3D imaging [79] and Krönke *et al.* applied EM-tracked US to thyroid volumetry [78]. However, optical tracking requires a constant line of sight between cameras and markers on the transducer, and EM tracking is sensitive to Electromagnetic Interference (EMI), which can affect accuracy in clinical settings [80].

More recently, Robotic ultrasound (RUS) has been applied to improve scanning stability, trajectory control, and pose estimation [81, 82]. This robotic approach could be readily applied in USgHIFU without additional cost, since most approaches already use a robotic arm to deliver therapy. In 2020, Daunizau *et al.* proposed a robotic platform for treatment planning in interstitial HIFU for hepatic tumour phantoms, which used manual segmentation of the tumour-mimic volume and rotational scanning with a catheter prototype transducer [41]. This approach showed promise, but the fixed scanning regime is unsuitable for thyroid scans, and manual segmentation would prolong treatment duration. In 2025, Lin *et al.* demonstrated a RUS for thyroid volumetry, with force-based control and deep-learning guidance [83]. This showed the feasibility of the technique, with accurate volumetry (34.3 $\pm$ 0.3 mL vs. 34.5 mL on CT) and good force reproducibility (2.1 $\pm$ 0.5 N) [83], but did not evaluate the effect of scanning speed on 3D reconstruction or include 3D renderings of the reconstructed anatomy.

To date, most of the research into 3D reconstruction of head and neck ultrasound has investigated reconstruction of either the carotid artery [79, 84, 84–90], or thyroid gland [68, 83, 91–95], but no study has yet demonstrated a unified method for diagnostics and treatment planning in the neck. Further, most RUS thyroid studies focus on robotic control strategies for automation; however, automated RUS raises safety, regulatory, and ethical challenges [96], and a large skill gap remains between human experts and autonomous RUS [97].

Cooperative robotic (co-robot) tracking and manual guidance may offer an attractive alternative for use with extracorporeal ultrasound. In this mode, the sonographer manipulates the probe by hand with support from the robot arm while maintaining full manipulability

and control of the transducer [98]. This offers a middle ground between freehand and automated RUS, allowing the operator to control the scanning while still benefiting from robotic stability, speed regulation, and kinematic tracking [99–101]. To our knowledge, co-robotic ultrasound 3D reconstruction has not been attempted in thyroid ultrasound.

1.5.8 Metrics for geometric similarity

To assess the accuracy of treatment planning and guidance models, the Dice Similarity Coefficient (DSC) and Hausdorff Distance (HD) can be used to quantify geometric similarity between shapes and volumes. These are key metrics that can be used to evaluate segmentation, 3D reconstruction, and treatment-planning accuracy.

The DSC quantifies the overlap between a predicted shape (X) and a reference shape (Y), where DSC is given by twice the intersecting area divided by the sum of the individual areas/volumes. It can also be formulated in terms of the True Positive (TP), False Negative (FN), and False Positive (FP) regions:

$$\mathrm{DSC} = \frac{2\,|X \cap Y|}{|X| + |Y|} = \frac{2TP}{FN + 2TP + FP} \tag{1.2}$$

The DSC may be applied in segmentation as a measure of the overlap between manually segmented (ground truth) regions and automatically segmented regions [57]. It can be applied in 2D to determine the pixel-wise agreement between a ground-truth annotation and a predicted segmentation, or in 3D to determine the voxel-wise agreement between two volumes, such as a reconstructed thyroid from ultrasound and the same thyroid on CT. It has also been used to quantify treatment success for planned versus delivered HIFU therapy [41]. The DSC is useful for evaluating treatment planning steps in HIFU since it penalises both regions which may be under-treated (false negatives) and over-treated (false positives) [102]. A pictorial representation of this is shown in Fig. 1.4a.

The DSC is easy to interpret (0 = no overlap, 1 = perfect overlap) and is the most widely applied metric for set similarity in medical research. However, it should be noted that the DSC is insensitive to boundary errors, has a strong dependence on object size, and is insensitive to distinguishing between translation errors and boundary errors, and between false positive and false negative predictions. Most importantly, two segmentations can have a high Dice score while still showing noticeable differences in their boundaries. This is particularly relevant in therapeutic applications, where the accuracy of treatment margins is critical, especially near sensitive structures.

Therefore, other metrics should also be considered, such as the HD, which gives the boundary accuracy as the difference between the surfaces of a predicted shape (X_s) and a reference shape (Y_s). The HD is useful in HIFU because treatment accuracy depends on the precise targeting of tissue boundaries. In some cases where the DSC shows high overlap, the HD highlights small but clinically relevant boundary errors that may indicate overtreatment

near critical structures. For each surface point ($x \in X_s$) and ($y \in Y_s$), the set of shortest distances to the corresponding shape are calculated ($d_{X_s \to Y_s}$ and $d_{Y_s \to X_s}$):

$$d_{X_s \to Y_s} = \left\{ \min_{y \in Y_s} \|x - y\| \ : \ x \in X_s \right\} \tag{1.3}$$

$$d_{Y_s \to X_s} = \left\{ \min_{x \in X_s} \|y - x\| \ : \ y \in Y_s \right\} \tag{1.4}$$

These are used to find the (bidirectional) maximum Hausdorff Distance ($\mathrm{HD}_{\max}$), defined by the maximum of both sets of shortest distances, and the the 95th-percentile HD (HD_{95}), which is less sensitive to outliers. This is shown in Fig. 1.4b.

$$\mathrm{HD}(X_s, Y_s) = \max\left(\max(d_{X_s \to Y_s}),\ \max(d_{Y_s \to X_s})\right) \tag{1.5}$$

$$\mathrm{HD}_{95}(X, Y) = \max\left(P_{95}(d_{X \to Y}),\ P_{95}(d_{Y \to X})\right) \tag{1.6}$$

The HD provides a clinically intuitive and direct indication of the geometric similarity between objects, where 0 mm indicates a perfect prediction. However, the HD is sensitive to noise and depends on image resolution; therefore, HD is best interpreted alongside other similarity metrics, such as the DSC, volume difference, precision, and recall, where a combination of metrics provides the most comprehensive assessment of uncertainty.

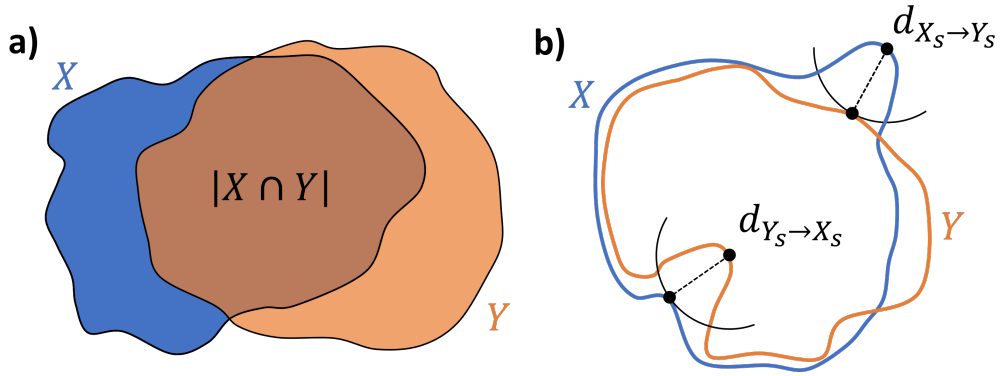


Figure 1.4: a) Dice Similarity Coefficient (DSC) gives the similarity between two shapes b) The Hausdorff Distance (HD) gives the surface distance between two shapes.

Chapter 2

AI for real-time ultrasound segmentation

Medical image segmentation has been widely explored as a result of the AI revolution; however, ultrasound presents unique challenges compared to other imaging modalities. It is a real-time imaging modality, meaning models must be both fast and accurate for imaging-guidance and live diagnostic applications, and speckle noise and poor boundary contrast make multi-organ segmentation challenging.

This is particularly evident in lung ultrasound (LUS), which remains challenging for new clinicians due to its reliance on interpreting artefacts. During the COVID-19 pandemic, LUS gained significant interest, and this paper investigated deep learning for multi-object segmentation in LUS, aiding interpretation through a live imaging overlay. This was a timely example of a challenging ultrasound use case with a clinical dataset available, enabling development of models that would later be applied to thyroid ultrasound.

This paper investigated deep learning architectures and training regimes for real-time segmentation, defining a lightweight U-Net model and a novel ultrasound-specific augmentation pipeline to improve robustness across imaging settings and reduce the annotation burden. The model was implemented with a point-of-care ultrasound system, further developed in Chapter 4 and applied to thyroid ultrasound in Chapter 5.

Importantly, the paper highlighted the need for large, high-quality datasets in ultrasound research. The available clinical dataset was small and homogeneous, comprising only images from patients with advanced COVID-19 pneumonia (with lung consolidation), thereby limiting model performance. This was indicative of UK practice during the pandemic, when lung ultrasound was largely reserved for severely ill patients, and scans were not routinely uploaded to Patient Archive and Communication System (PACS). The study also emphasised the subjectivity of ultrasound annotation and the need for inter-annotator agreement studies to define realistic performance benchmarks.

Deep learning for real-time multi-class segmentation of artefacts in lung ultrasound

Lewis Howell[1,2], Nicola Ingram[3], Roger Lapham[4], Adam Morrell[4], James R. McLaughlan[2,4], Tze Min Wah[2,3]

[1]School of Computing, University of Leeds, UK
[2]School of Electronic and Electrical Engineering, University of Leeds, UK
[3]Leeds Institute of Medical Research, University of Leeds, UK
[4]Radiology Department, Leeds Teaching Hospital Trust, Leeds General Infirmary, Leeds, UK

Ultrasonics 140:107251, May 2024.
DOI: 10.1016/j.ultras.2024.107251
Part of special issue: Artificial Intelligence for Ultrasonics

---

Abstract

Lung Ultrasound (LUS) has emerged as a safe and cost-effective modality for assessing lung health, particularly during the COVID-19 pandemic. However, interpreting LUS images remains challenging due to its reliance on artefacts, leading to operator variability and limiting its practical uptake. To address this, we propose a deep learning pipeline for multi-class segmentation of objects (ribs, pleural line) and artefacts (A-lines, B-lines, B-line confluence) in ultrasound images of a lung training phantom. Lightweight models achieved a mean Dice Similarity Coefficient (DSC) of 0.74, requiring fewer than 500 training images. Applying this method in real-time, at up to 33.4 frames per second in inference, allows enhanced visualisation of these features in LUS images. This could be useful in providing LUS training and helping to address the skill gap. Moreover, the segmentation masks obtained from this model enable the development of explainable measures of disease severity, which have the potential to assist in the triage and management of patients. We suggest one such semi-quantitative measure called the B-line Artefact Score, which is related to the percentage of an intercostal space occupied by B-lines and in turn may be associated with the severity of a number of lung conditions. Moreover, we show how transfer learning could be used to train models for small datasets of clinical LUS images, identifying pathologies such as simple pleural effusions and lung consolidation with DSC values of 0.48 and 0.32 respectively. Finally, we demonstrate how such Deep Learning (DL) models could be translated into clinical practice, implementing the phantom model alongside a portable point-of-care ultrasound system, facilitating bedside assessment and improving the accessibility of LUS.

Keywords: Lung Ultrasound, Deep Learning, Artificial Intelligence, Image Segmentation, Point of Care Ultrasound, COVID-19

2.1 Introduction

Artificial Intelligence (AI) has significant potential in the healthcare industry, where data could be used to reduce costs, whilst improving patient care [1]. Early uses of AI in healthcare were aimed at supporting clinical decisions [2], before rapidly expanding into a wide range of applications, such as management and interpretation of patient data, predictive medicine, and allocation of health service provision [3]. DL, a subset of AI and Machine Learning (ML), uses hierarchical combinations of learnable feature extractors to automatically build high-level representations of data [4]. The most common DL techniques for vision problems are based on Convolutional Neural Networks (CNNs), such as the U-Net architecture [5], which are often used for segmenting biomedical images. Recently, variants of the original U-Net, such as those which incorporate attention gates or vision transformers have been demonstrated to achieve positive results in semantic image segmentation tasks, where each pixel in an image is assigned a class [6–8]. Additionally, transfer learning may help DL models to achieve better performance when there are limited images in a dataset [9, 10], leveraging the knowledge gained by a model trained on one dataset to inform the training of a model on a related dataset [11]. In the medical imaging domain, semantic segmentation networks have mainly been used with radiography datasets from Magnetic Resonance Imaging (MRI) or X-ray Computed Tomography (CT) scanners, with significantly fewer from ultrasound [12]. The primary function of these networks is to efficiently segment features to aid in diagnosis, prognosis, or assist with image-guided therapies and/or interventions.

LUS is a technique that can be used to investigate the health of a patient's lungs in a range of clinical settings [13–15]. LUS is straightforward to perform but requires experienced ultrasound practitioners due to a steep learning curve for correct image interpretation [16–18]. This is because, unlike most soft tissue imaging with ultrasound, it relies on the observation and interpretation of ultrasound artefacts [19] (Figure 2.1). A healthy lung would predominantly be air-filled and due to the high acoustic impedance mismatch between tissue and air, it results in near complete back reflection of the ultrasound waves. These reflected waves can reverberate between the ultrasound transducer and the pleural line, which produces B-mode images with characteristic hyperechoic horizontal lines at regular intervals. These horizontal artefacts appear parallel to the pleural line and are commonly referred to as 'A-lines' [20] (Figure 2.1a). As the pathological state of the lung deteriorates, there is a decrease in the presence of air, replaced by fluid or other biological material. It is thought that as the air spaces in the lung are replaced with less attenuating medium the ultrasound field penetrates deeper into the lung, causing the greater

appearance of vertical artefacts including 'B-lines' and 'white lung' artefacts [21]. B-lines are highly heterogeneous in their appearance, however, these vertical artefacts generally appear as discrete, vertical, hyperechoic, artefacts, which originate at the pleural line [20]. Recent work suggests that B-lines may result from the internal reflection of ultrasound by semi-aerated alveoli, acting as acoustic traps [22](Figure 2.1b). In severe pathology cases, discrete B-lines can merge together and become confluent ('confluent B-lines'), even covering the entire intercostal space as a 'white lung' artefact (Figure 2.1c), which may be accompanied by a consolidation, where fluid build-up allows for direct imaging of the lung parenchyma [23]. Other indications include pleural effusions, where fluid built up in the pleural space can be directly imaged with ultrasound. Simple pleural effusions commonly present as a homogeneous anechoic region of fluid, while complex pleural effusions appear heterogeneous, representing a region of turbid fluid which may contain particles, debris, or clotted blood [24].

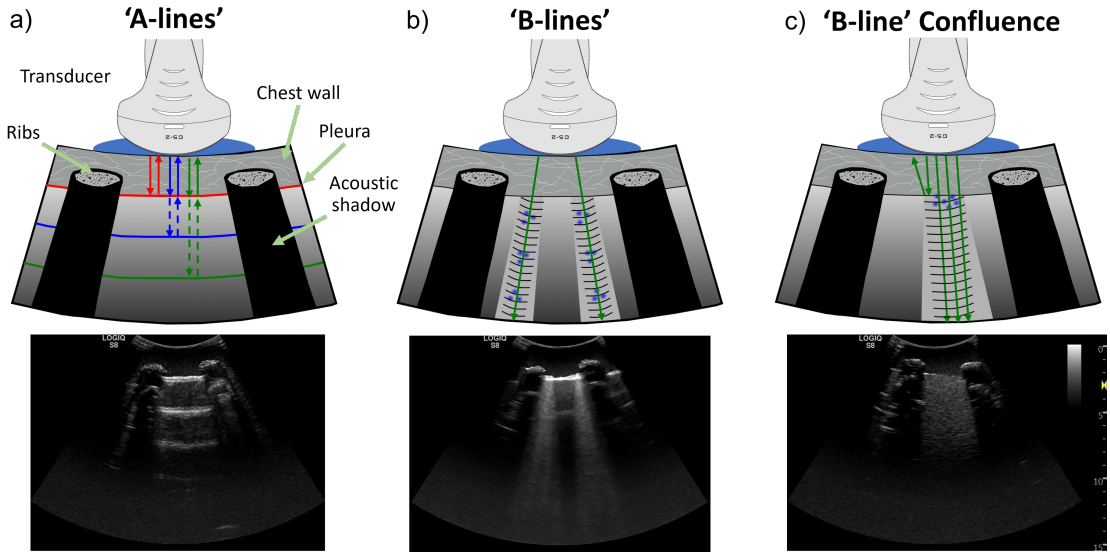


Figure 2.1: A representation of LUS artefacts and corresponding example B-mode images of these artefacts acquired using a lung phantom model. (a) Ultrasound waves are reflected at the pleura, creating horizontal reverberation artefacts ('A-lines') in a healthy lung. (b) As the amount of fluid increases, vertical artefacts ('B-lines') become more prevalent. (c) In more serious cases, B-lines can merge together and become confluent.

To date, a number of studies have investigated the application of DL to LUS, including image classification, artefact detection, and segmentation [10, 25, 26]. Among these, several have investigated the frame-wise classification of B-lines. In 2020, a study classified the presence, or absence of B-lines with 93 % sensitivity and 96 % specificity, additionally measuring B-line severity (multi-class image classification on a scale of 0-4) [27]. While this approach showed promise, it lacks transparency in terms of explaining the model's prediction process, thereby limiting its trustworthiness for computer-assisted diagnosis [28]. Other studies have attempted to count and localise B-lines [29, 30]. In 2020, van Slaun and Demi used DL for automatic detection and localisation of B-lines images from phantoms and

patients, with the potential for real-time implementation at an inference framerate of 276 Frames per Second (FPS) [30]. In this method, localisation was obtained using Gradient-weighted Class Activation Maps (grad-CAM), which lack fine-grained detail compared to segmentation and have been shown to perform poorly in cases with multiple occurrences of the same class [31]. This may be important in the common case of multiple B-lines in a single image. Furthermore, since the appearance of B-lines is highly dependent on ultrasound imaging parameters [32, 33], and B-lines can merge together, B-line counting is unreliable in practice [34].

In 2018, Kulhare *et al.* trained a single-shot object detector for localisation of vertical artefacts (B-lines and confluent B-lines), horizontal artefacts (A-lines), pleural line, pleural effusion, and lung consolidations in B-mode images [35]. While this method achieved >85% sensitivity and specificity in all classes except B-lines, object detection provides only bounding boxes for features, limiting its usefulness for improving visualisation and feature quantification. Conversely, segmentation allows precise delineation of features and may be more useful in severity assessment. In 2020, Roy *et at.* segmented healthy lung features and markers of disease in LUS, using this to produce an early method of severity assessment [36]. Others have segmented specific artefacts and anatomy, such as Xue *et al.* who segmented the pleural line, A-line, B-line, and lung consolidation with a mean DSC of 0.44 [37] and Gare *et al.* who segmented A-line, B-line and pleural line using a pre-trained U-Net model to achieve a mean intersection over union of 0.63, excluding the background class [38].

To date, much of the literature in DL for LUS has focused on classification and secondary localisation of individual features, with few attempting the more challenging and information-rich task of multi-class anatomical and artefact semantic segmentation. Of the studies using segmentation, even fewer consider real-time performance, which is important since LUS is a highly dynamic and investigative imaging technique. Many of the existing DL approaches also rely heavily on large labelled datasets, which can be difficult to obtain in practice [25, 39]. In this work, we train DL models for the semantic segmentation of artefacts (A-lines, B-lines, confluent B-lines) and anatomy (ribs, pleural line) in phantom images, using fewer than 500 images in training. We optimise these models for segmentation performance and inference time, demonstrating the real-time performance of our models in live imaging with a Point-of-Care Ultrasound (PoCUS) system. To supplement this, we investigate the feasibility of transfer learning to train a clinical model with less than 60 images, improving model convergence. We discuss how such models enhance visualisation, enable automatic quantification of features, and may help address the skill gap in LUS interpretation by improving education and minimising operator variability in image interpretation [20, 40–42].

2.2 Methods

2.2.1 Image Datasets

A commercial lung ultrasound model training phantom (CAE Healthcare Inc., Blue Phantom COVID-19 Lung Ultrasound Simulator) was used as a controlled platform to investigate LUS segmentation. This anthropomorphic half-chest model was designed to provide a realistic platform to assess features from healthy to severely damaged lungs and has been suggested as a useful training tool for learning the pathological signs and progression of COVID-19 [43]. Anatomical landmarks of the phantom include the chest wall, ribs, lungs and pleural lining. Under ultrasound examination, it can demonstrate A-lines, B-lines, and B-line confluence in different regions as well as replicating lung sliding with an electric pump.

B-mode ultrasound videos of the lung phantom were acquired using clinical ultrasound systems (GE Healthcare, Logiq S8 & E10) using both a convex curved-array transducer (GE Healthcare, C1-6), and a linear transducer (GE Healthcare, 9L) with a focal depth of 2.5-3 cm (focussed on the pleural line), imaging depth of 10-15 cm, and a Mechanical Index (MI) in the range of 1.2-1.3. A total of 297 two-second video clips were recorded by a senior sonographer, performing sagittal scans through the anterior and lateral lungs, similar to a standard clinical examination [24]. To avoid systematic biases and maximise image diversity, approximately 560 images were randomly sampled from the videos for annotation.

The VGG Image Annotator (VIA) tool [44] was used to label all images, with polygon annotation of objects and artefacts in the image. To make these labels compatible with semantic segmentation, polygons were converted to segmentation masks using contour filling (OpenCV). For the phantom dataset images were annotated for 5 classes of object/artefact: Rib, Pleural line, A-line, B-line, and B-line confluence. Five individuals with varying levels of ultrasound experience (ranging from none to $>$ 15 years of active research in the field) independently labelled 100 images each. Additionally, a senior sonographer labelled a further 64 images, with a peer review of the labels to ensure consistency between groups. This dataset is accessible online.

Additionally, a retrospective clinical dataset of anonymised ultrasound images was requested from Leeds Teaching Hospitals Trust. These images were recorded between March 2020 and March 2021 in hospitals across the Leeds Teaching Hospital Trust (LTHT) and included patients diagnosed with COVID-19 pneumonia. A variety of point-of-care ultrasound systems and transducers were used to collect the images.

A total of 57 images were selected for annotation from 41 patients across 8 different hospitals. As the images were fully anonymised, it was impossible to split data at the patient level, so data was split randomly to minimise bias. Since these images were taken from patients with severe COVID-19 pneumonia, most contained examples of lung consolidation

and/or simple pleural effusion, with no A-lines or individual B-lines present. Additionally, only two images contained clear examples of confluent B-lines which is too few for model training. Therefore, the classes identified for transfer learning were Ribs, Pleural line, Lung consolidation, Simple pleural effusion, and Complex pleural effusion. These images were labelled by two experienced sonographers and cross-verified for accuracy. This anonymised clinical dataset will be made available on reasonable request.

2.2.2 Models and Training

2.2.2.1 Model architecture and augmentations

A lightweight version of the U-Net architecture [5] was developed for image segmentation to provide a balance between model speed and accuracy. This was also compared with other variants of the U-Net including Residual U-Net, Attention U-Net, U-Net++, Inception U-Net, SE U-Net, and Dense U-Net [12]. These models all rely on an encoder-decoder structure for segmentation, using skip connections to improve information flow and better preserve both local detail and global context, improving prediction accuracy [5]. In the encoder, the input image is progressively downsampled, with intermediate convolutional blocks which capture information in the image as an increasing number of feature maps. The decoder then upsamples the feature maps back to the original image size over several steps, collating the information as it passes through more convolutional blocks before mapping it to a segmentation mask in the output layer, where each pixel in the mask is assigned a class.

The lightweight implementation of the U-Net model used in this study, keeps the core structure of the base U-Net, with two convolutional blocks in each of the four layers of the encoder and decoder, with 32, 64, 128, and 256 filters respectively. Alongside 3x3 convolutions, batch normalisation [45] and dropout layers (20%) [46] were added into the convolutional blocks to reduce overfitting and improve generalisability, especially when the dataset size is limited. Additionally, Leaky Rectified Linear Unit (Leaky ReLU) activation functions were employed to introduce non-linearity into the data transforms, and enhance training stability by addressing the issue of vanishing gradients [47]. Finally, bi-linear upsampling (un-pooling) was used instead of deconvolution in the decoder to help minimise computational complexity and reduce the number of learnable parameters.

To improve the model's ability to generalise to unseen data, an ultrasound-specific augmentation pipeline was used during the training phase (Figure 2.2). This included geometric transformations (horizontal axis flip, random rotation) and ultrasound-specific augmentations (gain, time gain compensation, and depth). Geometric transformations increased data diversity, while ultrasound-specific image augmentations were designed to improve model robustness to changes in common imaging settings. To replicate the effect of changing gain, an augmentation was created to adjust the image brightness and contrast by a factor in the interval of +25 to -25 %. For the depth control, an augmentation was applied

to randomly crop, centre, and pad images such that the scale and extent were modified similarly to depth adjustment. Finally, for Time Gain Compensation (TGC) the brightness along the vertical axis was altered using shifted Gaussian functions at eight depths, the amplitude of which was sampled from a normal distribution.

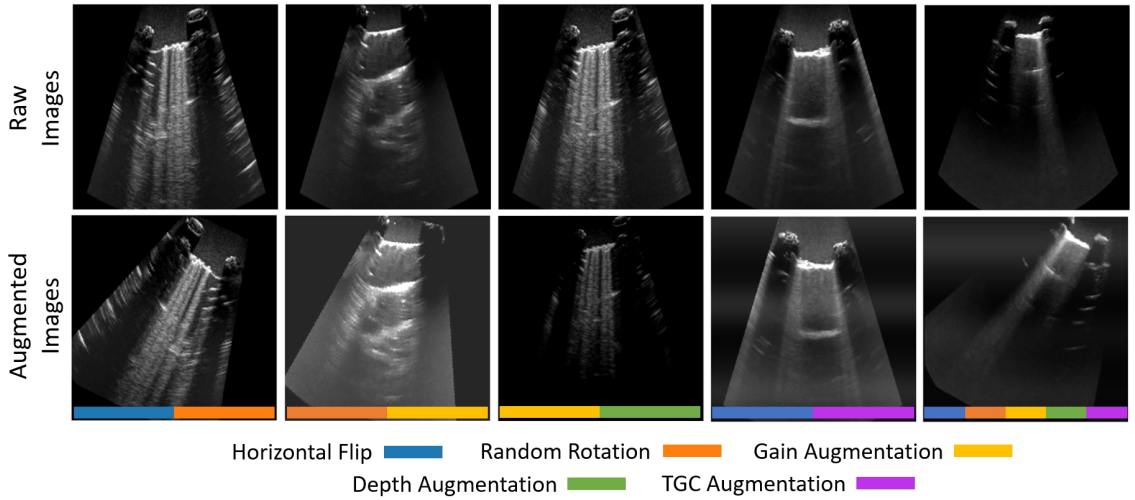


Figure 2.2: Examples of lung ultrasound B-mode images from a model phantom (top row) and the same images with applied augmentations (bottom row) including horizontal flip, random rotation, gain, time gain compensation (TGC), and depth augmentations.

Pre-processing was kept to a minimum to reduce the computation time during model inference. Images and masks were cropped to the approximate scan area (i.e. removing scan descriptors), resized to 256x256x1 and then normalised to the range 0-1.

2.2.2.2 Training and Evaluation

Prior to training with the LUS phantom data, an independent 20% testing dataset was split from the training data to allow a fair assessment of model performance. This split was conducted at the lowest level (video level), reducing the likelihood of similar images appearing in both datasets, hence minimising the risk of information leakage. Additionally, a 20% validation set was used to monitor performance during model training and regulate the learning process. All reported scores are given as the average test set score ± the standard deviation for 5 splits.

A dynamic approach to training was used, utilising the Adam optimiser [48] with an initial learning rate of 1e-4, which decreased upon a plateau of the validation loss, with early stopping to prevent overfitting. A Combo loss function which combined the Dice loss and cross-entropy loss was used, weighted 2:1 in favour of the Dice term [49]. This helps to address the class imbalance (most pixels belong to the background class) whilst maintaining training stability. Augmentations were applied 'on the fly', meaning that each training image had random augmentations applied each epoch, maximising data diversity. All training was conducted on a single Graphics Processing Unit (GPU) (NVIDIA RTX

3080 laptop), using CUDA 11.4, cuDNN 8.4.1 and models implemented in TensorFlow 2.9.1 for Python 3.9. Software implementation for network training and validation is made freely available online.

Models were evaluated using an independent test dataset not seen by the network during training. The per class and average DSC were used to quantify the similarity between the predicted masks (X) and manually labelled masks (Y). The DSC is given by the equation:

$$\mathrm{DSC} = \frac{2\,|X \cap Y|}{|X| + |Y|} \tag{2.1}$$

where $\cap$ represents the intersection operator.

To demonstrate the real-time application of our model, the trained network was used to segment a live image feed from a cart-based point-of-care ultrasound system (GE Healthcare, Venue Fit R3). The LUS phantom was imaged using a convex curved-array transducer (GE Healthcare, C1-5-RS) and frames were streamed to the host computer via an High-Definition Multimedia Interface (HDMI) to Universal Serial Bus (USB) 3.0 capture interface (Magewell, MAG-32060). These frames were pre-processed, segmented by the U-Net model, and the segmentation masks overlaid onto the B-mode image to be displayed on a screen using OpenCV for Python (Figure 2.3).

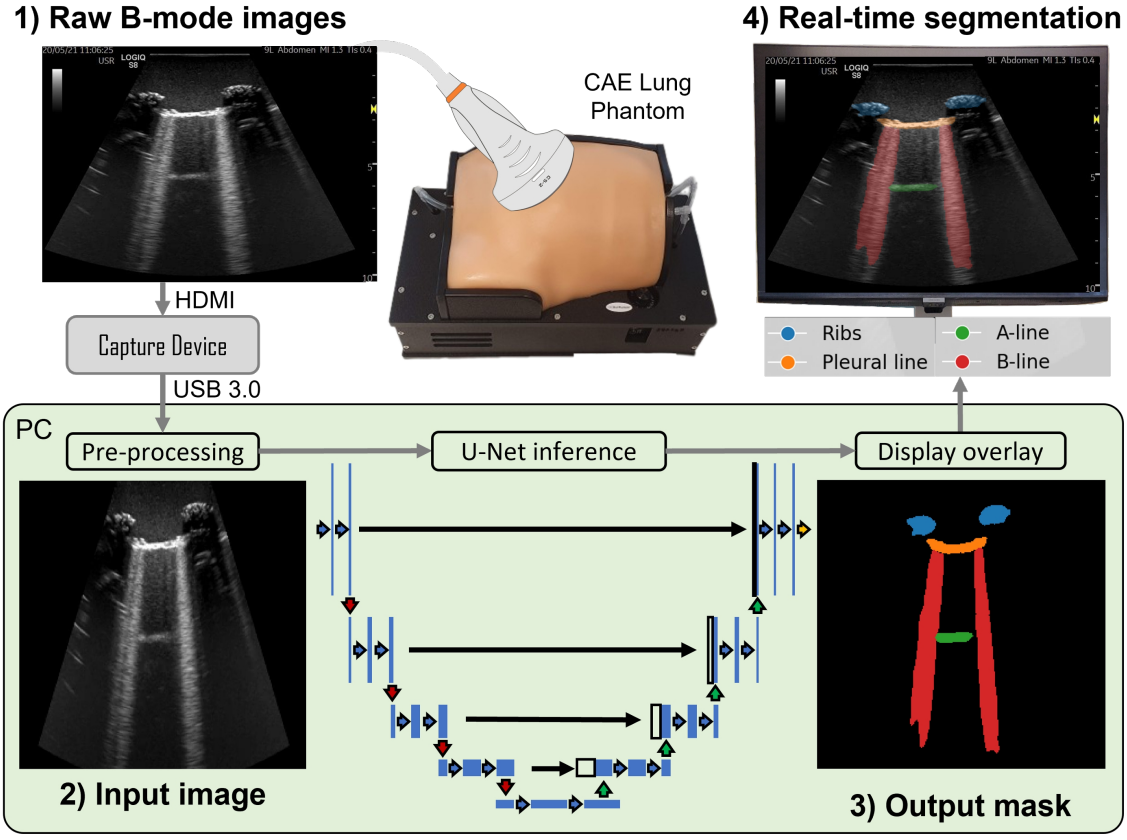


Figure 2.3: Workflow diagram showing real-time lung ultrasound segmentation with U-Net.

2.2.2.3 Transfer learning

We explored the feasibility of transfer learning to train U-Nets for segmentation with clinical LUS. By choosing to refine the pre-trained LUS phantom model instead of training a new model from scratch, we leveraged the generic LUS features learned by the network to help the model learn to recognise new features in clinical LUS images. We trained this clinical model to segment five classes, including ribs and pleural line (common to both the phantom and clinical datasets), lung consolidation, simple pleural effusion, and complex pleural effusion (unique to the clinical dataset).

To this effect, we froze the encoder and bottleneck of the pre-trained phantom model and then fine-tuned the weights in the decoder and output layers. This approach ensured that the representations contained in the encoder and bottleneck were not destroyed during training, allowing the features learned to be used to make predictions on the new dataset. The training pipeline was similar to that of the phantom mode but with a 10% test dataset and no validation set due to the limited dataset size and a smaller learning rate of 5e-5 to mitigate overfitting risks.

2.3 Results

Training on a single GPU took approximately 12 minutes with 450 images. As expected, we observed a decrease in the training and validation losses as the model learns, converging on a point of stability with a negligible generalisation gap between training and validation loss (Figure 2.4a). Model performance improved with an increase in the number of training images until around 300 images, after which we saw no significant improvement (Figure 2.4b). Additionally, we found models trained with ultrasound-specific augmentations consistently outperformed those without, with a DSC increase of 0.04 $\pm$ 0.01. The accuracy and DSC for the model trained on the full dataset (450 images) can be seen in Table 2.1. A comparison with state-of-the-art U-Net variants was also conducted, but we found no benefit to accuracy (Figure A.1) or speed (Figure A.2) when using these more complex architectures with our dataset.

Qualitative examples of U-Net segmentations are presented in Figure 2.5a-l. The ribs and the pleural line were generally segmented accurately, even in cases where the pleural lining was thickened and irregular (a-l). B-lines were detected in the majority of instances (c-h, k); however, the extent to which they extended axially in the predictions was occasionally less than that of the manually-labelled images due to lower signals at greater depths (d, k). Most A-lines, including partially obscured ones, were detected by the model (a-f, h, l), although A-lines further from the pleural line had a lower likelihood of being segmented (d). B-line confluence was segmented accurately in most images (i, j), although, in some cases, confusion between individual B-lines and B-line confluence led to the misclassification of certain regions (l).

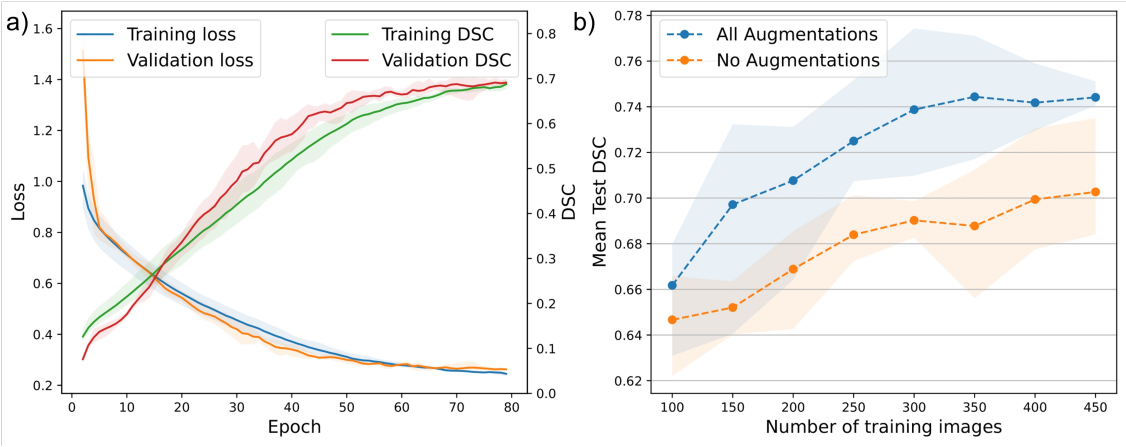


Figure 2.4: a) Lightweight U-Net Training loss and DSC during model training. b) Mean DSC for models trained with 100 to 450 images, with and without image augmentations. Error regions show the best and worst scores over five training runs.

Table 2.1: Scores for segmentation of the LUS phantom are given as the mean ± standard deviation across five training repeats. Timing performance metrics were calculated during real-time experiments and averaged across several hundred frames.

	Metric	**Score**
Dice similarity coefficient (DSC)	Background	0.98 ± 0.001
	Ribs	0.80 ± 0.01
	Pleural line	0.81 ± 0.01
	A-line	0.63 ± 0.03
	B-line	0.72 ± 0.01
	B-line confluence	0.73 ± 0.09
	Mean (ex. background)	0.74 ± 0.02
Accuracy (%)	Pixel-wise accuracy	95.7 ± 0.34
Timing performance (ms)	Inference	30.0 ± 2.77
	Pre-processing	3.36 ± 0.72
	Displaying	16.8 ± 0.39
Framerate (FPS)	Inference framerate	33.4 ± 2.86
	Total framerate	20.0 ± 1.22

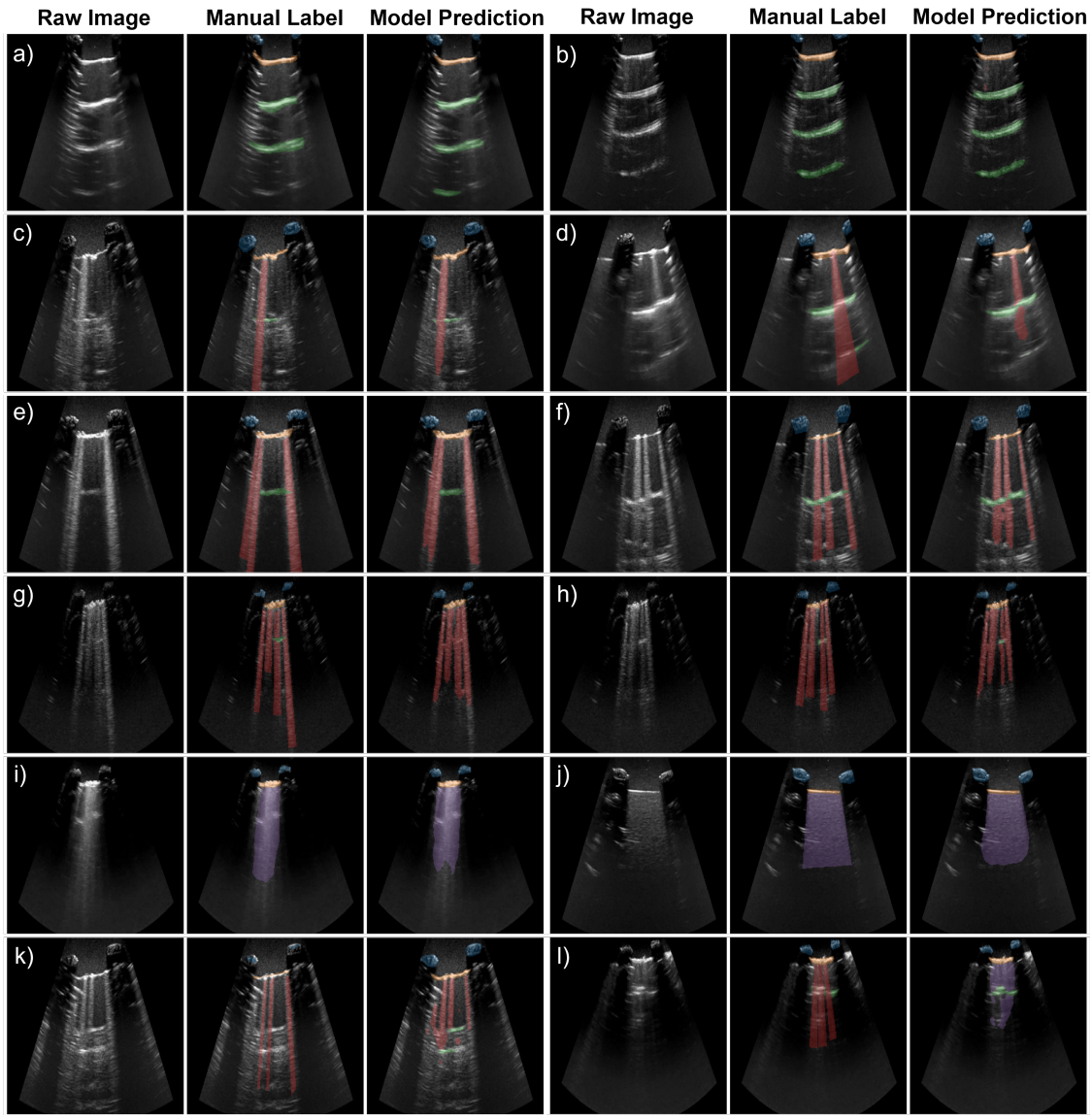


Figure 2.5: Example LUS phantom images, overlaid with their manual label and corresponding model prediction. The colour overlay used is as follows, ribs: blue, pleural line: orange, A-lines: green, B-lines: red, and B-line confluence: purple.

These observations were supported by an assessment of the pixel-wise agreement between the manually-labelled and model-predicted masks. The normalised confusion matrix for precision (Figure 2.6) shows that of the manually-labelled images, 86.8% of pixels labelled as ribs, 85.4% of pleural line, and 72.2% of B-line confluence were predicted correctly by the model. However, only 57.7% of A-line and 57.9% of B-line pixels were correctly predicted. We can see that the majority of pixel-wise errors are false negatives, but some false positive predictions occurred where A-line pixels are predicted as B-line or B-line confluence (likely due to overlapping of A-lines and B-lines) and B-line confluence pixels are predicted as B-line.

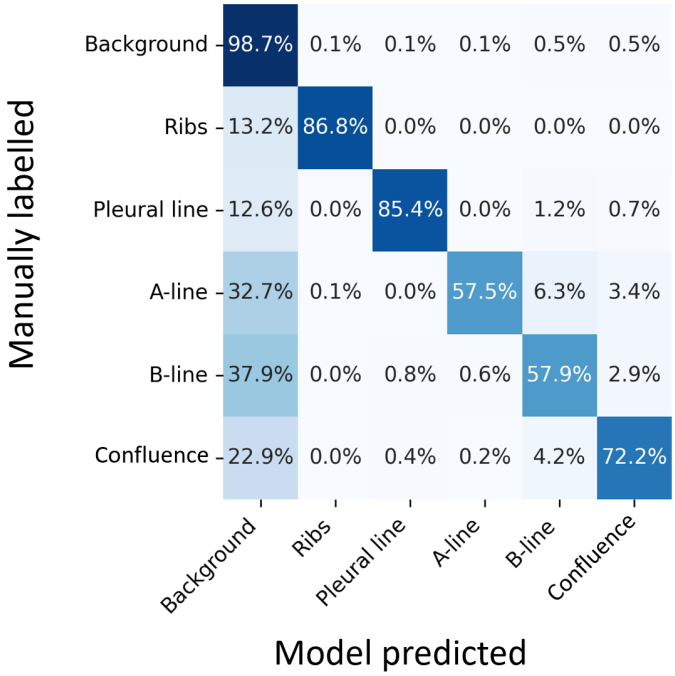


Figure 2.6: Confusion matrix for semantic segmentation of LUS phantom images, normalised to give the pixel-wise precision

Using the predicted segmentation masks, metrics related to the size and shape of objects in the image can also be considered. Here, we defined a semi-quantitative measure for the fraction of the intercostal window taken up by B-lines named the B-line Artefact Score (BLAS), showing how it can be automatically calculated during real-time imaging. To measure this, the intercostal Region of Interest (ROI) was determined as the area between the ribs and bounded by the bottom of the pleural line, extending to the maximum depth of the detected B-lines. The B-line fraction (percentage of pixels classified as B-line or B-line confluence) was then quantified in the ROI, with respect to depth (Figure 2.7). From this, the percentage of the intercostal space occupied by vertical artefacts (B-lines or confluent B-lines) was calculated using Simpson's Rule on the B-line fraction to give a BLAS in the range 0-1. Empirically, BLAS values less than 0.5 generally suggest one or a few B-lines, values between 0.5 and 0.9 suggest multiple B-lines or some B-line confluence, and values greater than 0.9 indicate white lung artefact. Calculating the ROI dynamically for each

image helped to ensure that the BLAS is independent of the imaging angle, depth and extent of the vertical artefacts. This is significant since B-lines cannot be defined by their spatial extent and may not extend to the bottom of the image [20]. Further improvements such as using a trapezoidal or sector-shaped ROI may be more appropriate for curved array transducers.

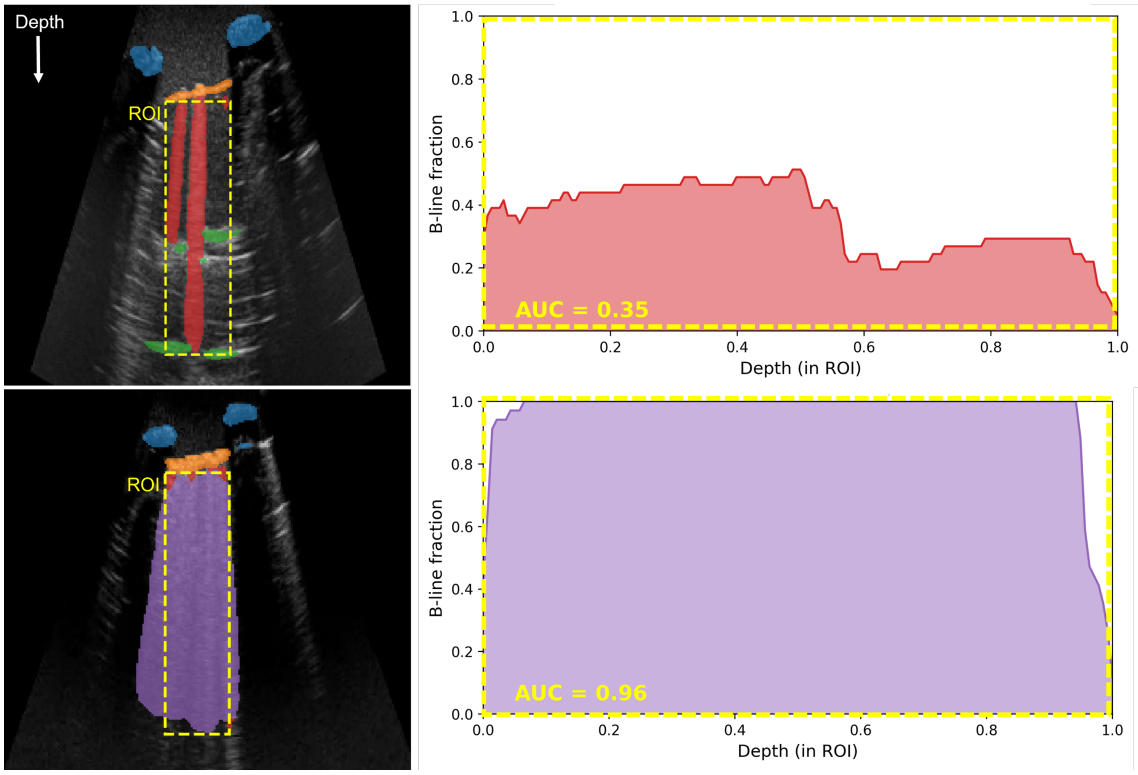


Figure 2.7: Calculation of the B-line Artefact Score (BLAS). (Left) The region of interest (ROI) is determined by the boundaries of the pleural line and extends to the depth extent of the measured B-lines (red) or B-line confluence (purple). (Right) The BLAS is equal to the normalised area under the curve (AUC) for the B-line fraction with depth.

A Bland-Altman plot was used to assess the agreement between measurements of the BLAS by manual segmentation vs automatic segmentation (Figure 2.8). In the phantom test dataset, the mean difference in BLAS was minimal (less than 0.01), indicating little or no systematic bias. The majority of points fall within the limits of agreement (± 0.29), defined by the mean plus or minus 1.96 times the standard deviation of the differences, suggesting agreement between the two methods, particularly for cases where the BLAS was greater than 0.9. Investigation of outliers revealed individual cases where a significant difference existed between the manually-labelled and model-predicted masks. Further research would be needed to evaluate the clinical significance of this approach.

To demonstrate live segmentation, the best model was implemented alongside the PoCUS system to image the LUS phantom and overlay masks in real time during scanning (Figure 2.9). Segmentation showed recognition of most features, despite the domain shift to a different ultrasound scanner and transducer to that used in model training. Images from

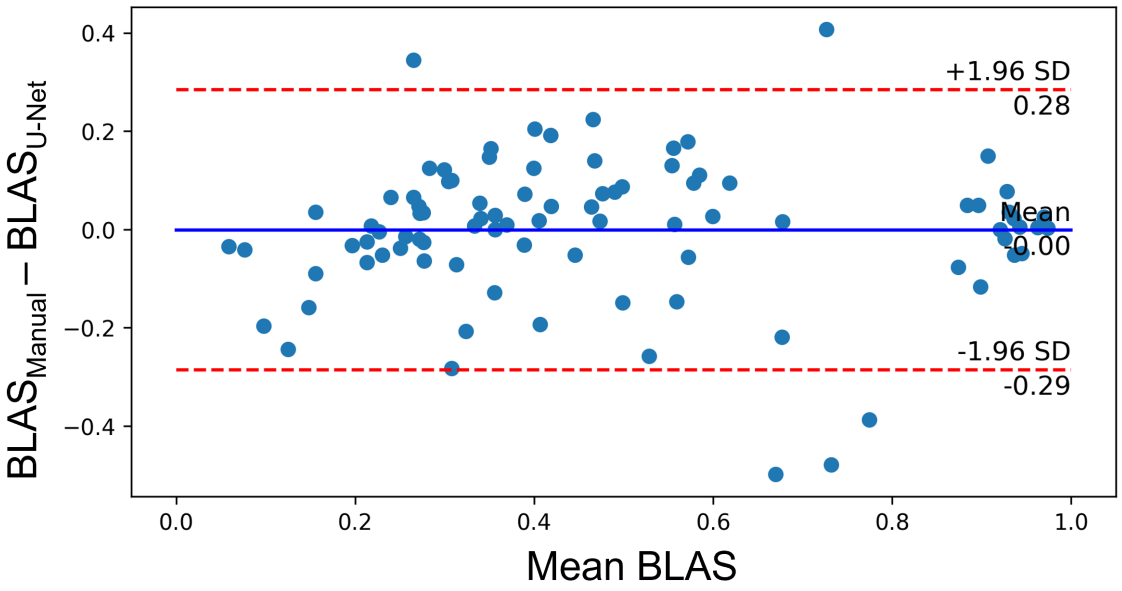


Figure 2.8: Bland-Altman difference plot showing the difference between B-line Artefact Score (BLAS) calculated from manual vs. AI-generated labels of LUS images. The plot shows the score difference plotted against the mean score for the test dataset. The mean difference (solid blue line) and limits of agreement (red dashed lines), defined as the mean difference ±1.96 times the SD of the differences are also displayed.

the PoCUS system showed greater speckle noise than those in the training dataset, yet the model could still accurately segment the ribs, pleural line and most A-lines, B-lines and B-line confluence. In some frames, the model struggled to segment the full extent of B-lines with depth, and some B-lines were misclassified.

In benchmark tests, the lightweight U-Net model achieved an inference framerate of 33.4 ± 2.86 FPS. However, in real-time experiments, the average framerate lowered to 20.0 ± 1.22 FPS. As a result, some frames were dropped, but the overall speed remained acceptable for real-time visualisation and the framerate variance was not noticeable to the eye. This discrepancy stems from the additional time required for pre-processing and image visualisation, where on average, pre-processing took 3.36 milliseconds, model inference took 29.9 milliseconds, and displaying the image took 16.8 milliseconds. To enhance performance, hardware-accelerated rendering could be used to reduce the display time, or alternatively, further optimisation of the network architecture could be explored.

Finally, we investigated the feasibility of transfer learning to re-train our phantom model to segment clinical LUS images. This approach proved valuable as the clinical dataset had very limited images for training, but the domain was similar to that of the phantom model. The limited clinical LUS images were found to be too few to train a U-Net model from scratch, with poor convergence in initial experiments. Transfer learning improved convergence and allowed reasonable feature recognition of the pleural line, lung consolidations, and simple pleural effusions in some examples (Figure 2.10a-d), but ribs and complex pleu-

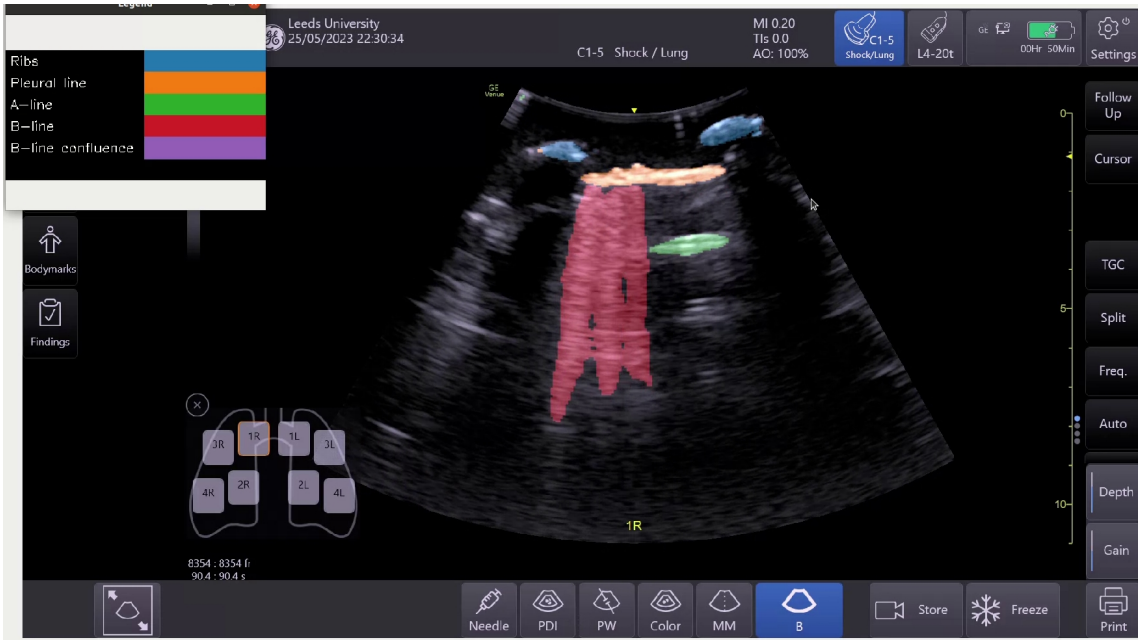


Figure 2.9: Real-time segmentation of anatomy and artefacts in a lung ultrasound phantom. Segmentation masks are overlaid onto B-mode images from a point-of-care ultrasound system during scanning to assist with image interpretation. For a video recording please see the web version of this article.

ral effusions were generally poorly segmented (e-f). Predictions on the clinical dataset were generally of poorer quality than those with the phantom dataset, likely due to the limited size and larger variability in the observed anatomy, pathology and types of scanners in the clinical dataset. Variations in the quality of clinical images also meant that some were reported to be diagnostically ambiguous (f) by the two expert annotators, highlighting the difficulty of the task and introducing additional uncertainty to the segmentation labels. Furthermore, since we could not monitor training with a validation dataset there was a possibility of overfitting and more data would be needed to fully assess the effectiveness of transfer learning. The per class DSC for simple pleural effusion, lung consolidation, and pleural line are 0.48, 0.32, and 0.25 respectively, however, the model was unable to reliably segment the ribs (DSC = 0.01) or complex pleural effusion (DSC = 0.01). This is reflected in the mean DSC (across classes excluding the background) of 0.20 ± 0.08 as measured with the limited test dataset.

2.4 Discussion

ML and DL present an exciting opportunity to assist in the interpretation of LUS [10], and other pathologies imaged using ultrasound [50]. Automatic segmentation of anatomical features and artefacts in LUS B-mode images could aid interpretation, as well as provide new information that could be used to monitor disease severity. To be successful, these models must be both accurate and sufficiently fast to provide feedback for the operator

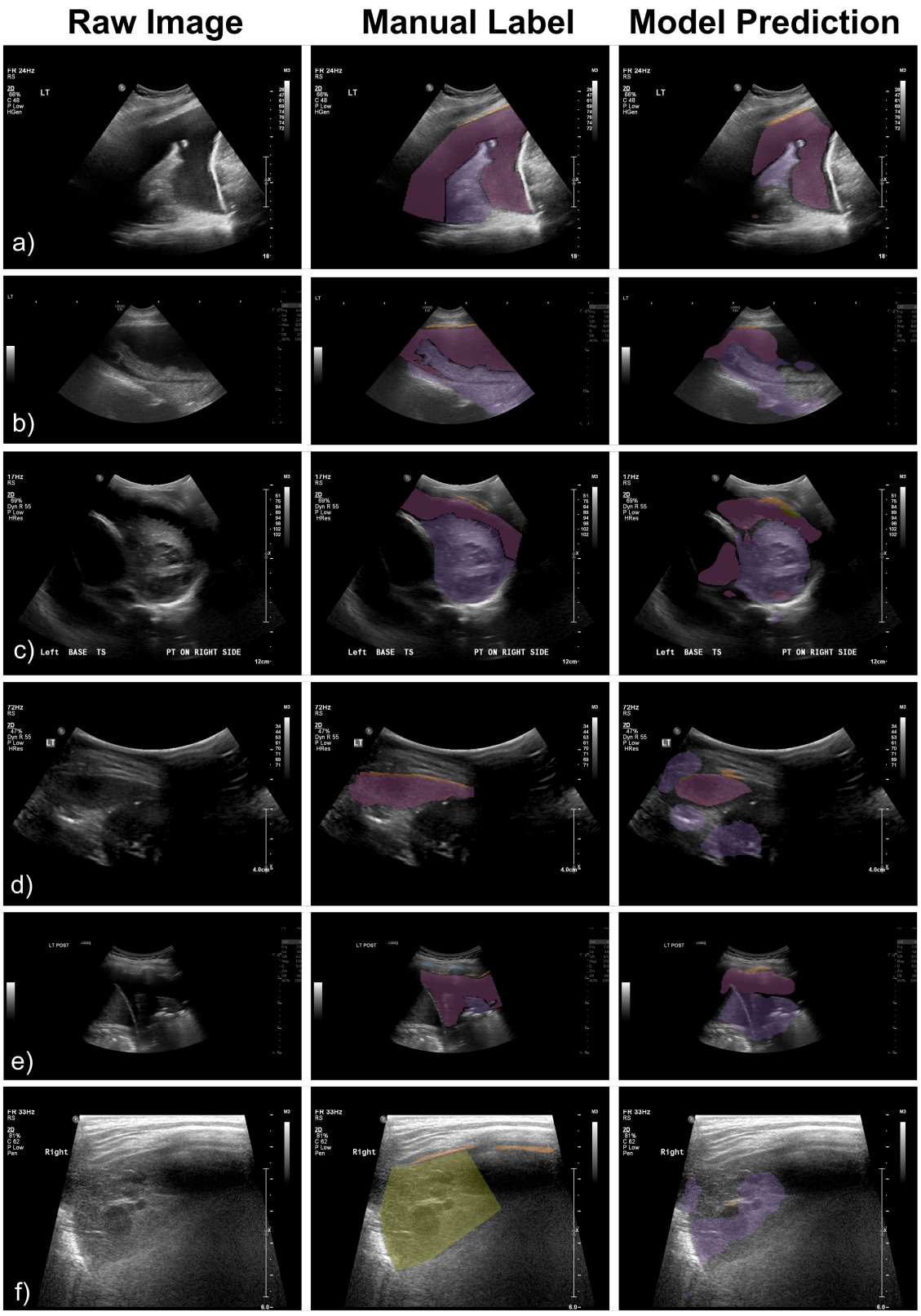


Figure 2.10: Example clinical LUS images, overlaid with their manual labels and corresponding model prediction. The colour overlay used is as follows, ribs: blue, pleural line: orange, lung consolidation: purple, simple pleural effusion: pink, and complex pleural effusion: yellow.

in real-time during scanning. However, due to low signal-to-noise ratio, poor contrast, speckle noise, shadows, and signal dropouts, ultrasound segmentation is a challenging task [51]. Perhaps for this reason, much of the previous work into applying AI with LUS has focussed on image or video classification rather than segmentation [26]. Of these publications, fewer still focus on real-time performance or report the inference time of their models. Additionally, there is an over-reliance in methods which rely on datasets of thousands to hundreds of thousands of images [39]. This is particularly limiting in densely-labelled segmentation, where obtaining large, representative, and high-quality annotated datasets can be prohibitively expensive and time-consuming, so approaches which rely on fewer images are desirable [10, 52].

In this study, multi-class segmentation of anatomical features (ribs and pleural line) and artefacts (A-line, B-line, and B-line confluence) in a LUS phantom was explored, considering both the speed and accuracy of the model. A lightweight U-Net model was trained for this task, which demonstrated a mean DSC >0.7 when trained with as few as 300 images. It was demonstrated that more complex U-Net variants did not yield any significant benefit on this dataset, which could be attributed to several factors including the limited number of training images and the uncertainty introduced by ambiguities in labelling.

Most significantly, the subjective nature of annotation is known to limit model performance in LUS [36, 53]. In ML, data labels provided by human experts are typically regarded as the immutable 'ground truth' [28]. However, objects in ultrasound often cannot be precisely delineated due to speckle noise and limited spatial resolution. This is particularly pronounced for artefacts, which are dynamic, heterogeneous, and lack well-defined boundaries. Consequently, inter-annotator agreement in LUS segmentation may be low [36, 54, 55]. This has impacts on diagnoses, for example, since it is difficult to know if B-lines are separate or confluent, scores based on counting B-lines are ambiguous [34]. For these reasons, we refer to the labelled datasets as 'manually-labelled' rather than ground truth, acknowledging the uncertainty associated with the labels. Nevertheless, this inherent variability may increase the value of DL segmentation in practice, since clinically verified models trained on data from several expert annotators may be able to provide a second opinion without the confirmation bias of any one individual.

A commonly overlooked issue in the literature is the sensitivity of ML models to imaging parameters [20]. To help address this concern, we incorporated ultrasound-specific augmentations into model training, reducing the diversity requirements in the dataset. This helps to improve generalisability and reduce the likelihood of overfitting, which is particularly valuable when training with limited images. As a result, we observed a mean DSC increase of 6.1% compared to training without augmentations.

Finally, many of the existing studies using AI to calculate prognostic scores from LUS lack explainability, relying on 'black box' methods which do not assist in the decision-making process or allow the clinician to independently review the basis for the recommen-

dations [10]. This can form a major barrier to their adoption in practice, where trust can play a significant role in whether a model is useful [56]. Approaches such as grad-CAM [57] have been used previously in LUS to represent areas of the image that are important to an image's classification [36, 42], but do not fully explain model reasoning.

Although segmentation still relies on complex models, its results are inherently intelligible and verifiable when segmentation masks are overlaid with real-time images. Therefore, metrics based on segmentation masks are more explainable and interpretable than those based on image classifiers. Here we present one such metric, named the B-line Artefact Score (BLAS), which is a simple semi-quantitative measure of the percentage of each intercostal space occupied by B-lines. This provides an indication of the number and appearance of B-lines in an image, which is known to be associated with a number of respiratory diseases [58], including COVID-19 [21], where significant B-line confluence marks the presence of pneumonia deterioration. Compared to individual line counting, measuring the percentages of rib space covered by confluent B-lines has been shown to be more reliable [34]. Our score can be calculated automatically over one or more respiratory cycles, with the maximum score displayed immediately on-screen. This facilitates a rapid assessment of patients while eliminating the ambiguity in counting individual B-lines and selecting the frame to measure.

A similar method could potentially be used to assess pleural line irregularities, which are the second most frequent finding associated with COVID-19 diagnosis, or A-lines which are a marker of recovery from COVID-19 [26]. Once validated on clinical data, such methods could allow the automatic calculation of severity scores, such as those proposed for COVID-19 pneumonia [59], allowing improved triage and management of patients in clinical settings, or monitoring disease progression. In this case, it must also be acknowledged that since the appearance of artefacts varies with frequency, bandwidth, and beam angle [60], predictions or diagnoses based on these measurements must consider these as confounding factors [20].

To showcase the practical application of our models, their implementation alongside a point-of-care ultrasound system was demonstrated. This showed the real-time overlay of segmentation masks onto B-mode images at 20 FPS, which highlighted key features to the operator during scanning. A user-friendly graphical interface allowed clinicians to toggle the visibility of segmentation masks or adjust their transparency, providing an alternative representation which aids feature visualisation. Although in this example a separate computer was used for processing the imaging feed, future implementations may consider how DL methods could be integrated onboard or with embedded devices. For example, tools such as TensorFlow Lite that optimise models for mobile devices may allow lightweight models to be deployed with handheld ultrasound, greatly expanding their usefulness at the bedside. Such approaches have already gained commercial interest. For example, in April 2023 an AI-enabled tool for scoring B-lines received Food and Drug Administration (FDA)

510(k) clearance for use with a handheld ultrasound device [61].

We also investigate the feasibility of transfer learning to train a model for clinical LUS images. By using transfer learning we leveraged the knowledge acquired from models trained on LUS phantom data to segment pathological features present in patients with severe COVID-19 pneumonia, such as lung consolidation and simple pleural effusions. Previous studies have used similar methods for segmentation of the ribs in LUS images, pre-training U-Net models on a large greyscale dataset of natural images (real-world objects) prior to transfer learning [62]. However, pre-training on ultrasound images may help models to better detect lung-specific patterns [10] and since phantom data is simple to collect and is free of data governance issues, this may present a viable alternative for model pre-training with transfer learning. While our model's DSC indicates it is unlikely to be suitable for implementation in clinical practice, this should be considered a baseline for comparison with future models. Here, training and evaluation of our models is limited by the small size and large heterogeneity of the available data and we are also unable to fully exclude residual information leakage in our model. Nevertheless, this method shows promise for future work in decision support and patient monitoring. With access to a larger and more diverse labelled dataset, future models will improve LUS interpretability in patients with varying disease severity.

Other limitations of this work include a lack of assessment for intra-annotator variability, which could not be conducted since there were no commonly labelled images available. Measuring this variability improves understanding of the data characterisation and subjectivity from annotation [63], and should therefore be a priority in the curation of future datasets. Secondly, phantom segmentation uses images from a single subject with clearly defined features, so translating the methods employed here to clinical data, which typically exhibit a wider range of patients and pathologies, may pose additional challenges and necessitate a larger and more diverse dataset for optimal performance [28]. Transfer learning may assist in training such models, but does not guarantee better performance [9] and still requires good-quality labelled training data. This study found that transfer learning from a model pre-trained with phantom data provided a practical method to improve model convergence, but more clinical data was needed to assess its impact on segmentation performance. Finally, it would be useful to quantify the generalisability of our models to new ultrasound systems and transducers and to assess whether the inclusion of ultrasound-specific augmentations enhances robustness to changes in imaging settings.

Future works could investigate the application of DL for real-time multi-object LUS segmentation of clinical images, with the aim of deploying models with portable or handheld scanners for training and education, and longer-term in clinical decision support systems. This could have an impact on the diagnosis and management of numerous respiratory conditions [26], especially in low- and middle-income countries and rural areas, where cost-effective ultrasound systems are more widely available than CT [64], but the expertise

needed to interpret them may be lacking [65]. More broadly, the methods for training and validation proposed here could be used to develop models for clinical examination of the thyroid, breast, heart or abdomen, among others. This may be especially useful in a variety of settings where real-time feedback is important, such as emergency assessment of trauma with ultrasound [66], ultrasound-guided biopsies [67], or for therapeutic systems such as ultrasound-guided focused ultrasound for tissue ablation [68].

2.5 Conclusion

This study demonstrated the application of deep learning to real-time, multi-class segmentation of objects and artefacts in lung ultrasound. Models trained with as few as 300 images were able to able to segment the ribs, pleural line, A-lines, B-lines, and B-line confluence in B-mode images of a COVID-19 lung tissue-mimicking phantom, with an average DSC of 0.74 across the five classes. When overlaid onto LUS B-mode images, segmentation masks offer an improved visual interpretation of the image. This could be incorporated into training systems for medical education of LUS in specialised programs or as part of training new clinicians [20], addressing the existing skill gap in LUS. To maximise their clinical potential, these models must be deployable on PoCUS systems and suitable for real-time inference. We demonstrated the feasibility of this at 20 FPS, with the potential to achieve $>$ 30 FPS with hardware-accelerated rendering. Furthermore, segmentation masks provide an explainable method for scoring disease severity, which has the potential to assist in the triage and management of patients for a variety of respiratory conditions. On this theme, we propose a B-line artefact score (BLAS) which automatically measures the percentage of the intercostal space occupied by vertical artefacts (B-lines and confluent B-lines). Future work should consider the translation of these methods to clinical data, considering transfer learning as a viable method to build models which can assist in the interpretation of LUS and reduce inter-operator variability associated with this subjective imaging technique.

2.6 Acknowledgements

This work was supported by the following UK Research and Innovation (UKRI) grants, EP/S001069/1 and EP/V043714/1. The authors would also like to thank Stephen Wolstenhulme who aided with the acquisition and labelling of phantom data.

Bibliography

[1] K.-H. Yu, A. L. Beam, and I. S. Kohane, "Artificial intelligence in healthcare," *Nature biomedical engineering*, vol. 2, no. 10, pp. 719–731, 2018.

[2] R. J. Burton, M. Albur, M. Eberl, and S. M. Cuff, "Using artificial intelligence to re-

duce diagnostic workload without compromising detection of urinary tract infections," *BMC medical informatics and decision making*, vol. 19, no. 1, pp. 1–11, 2019.

[3] S. Secinaro, D. Calandra, A. Secinaro, V. Muthurangu, and P. Biancone, "The role of artificial intelligence in healthcare: a structured literature review," *BMC medical informatics and decision making*, vol. 21, pp. 1–23, 2021.

[4] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27–48, 2016.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241, Springer, 2015.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, and others, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020.

[7] L. Karlinsky, T. Michaeli, and K. Nishino, *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*, vol. 13806. Springer Nature, 2023.

[8] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "UNETR: Transformers for 3D Medical Image Segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 574–584, 2022.

[9] K. Sailunaz, T. Özyer, J. Rokne, and R. Alhajj, "A survey of machine learning-based methods for COVID-19 medical image analysis," *Medical & Biological Engineering & Computing*, vol. 61, pp. 1257–1297, June 2023.

[10] T. Yang, O. Karakus, N. Anantrasirichai, and A. Achim, "Current Advances in Computational Lung Ultrasound Imaging: A Review," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 70, pp. 2–15, Jan. 2023.

[11] D. Karimi, S. K. Warfield, and A. Gholipour, "Transfer Learning in Medical Image Segmentation: New Insights from Analysis of the Dynamics of Model Parameters and Learned Representations," *Artificial intelligence in medicine*, vol. 116, p. 102078, June 2021.

[12] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-net and its variants for medical image segmentation: A review of theory and applications," *Ieee Access*, vol. 9, pp. 82031–82057, 2021.

[13] M. A. Pereda, M. A. Chavez, C. C. Hooper-Miele, R. H. Gilman, M. C. Steinhoff, L. E. Ellington, M. Gross, C. Price, J. M. Tielsch, and W. Checkley, "Lung ultrasound for

the diagnosis of pneumonia in children: a meta-analysis," *Pediatrics*, vol. 135, no. 4, pp. 714–722, 2015.

[14] L. Long, H.-T. Zhao, Z.-Y. Zhang, G.-Y. Wang, and H.-L. Zhao, "Lung ultrasound for the diagnosis of pneumonia in adults: a meta-analysis," *Medicine*, vol. 96, no. 3, p. e5713, 2017.

[15] M. Smith, S. Hayward, S. Innes, and A. Miller, "Point-of-care lung ultrasound in patients with COVID-19–a narrative review," *Anaesthesia*, vol. 75, no. 8, pp. 1096–1104, 2020.

[16] D. A. Lichtenstein, "BLUE-protocol and FALLS-protocol: two applications of lung ultrasound in the critically ill," *Chest*, vol. 147, no. 6, pp. 1659–1670, 2015.

[17] S. Moore and E. Gardiner, "Point of care and intensive care lung ultrasound: a reference guide for practitioners during COVID-19," *Radiography*, vol. 26, no. 4, pp. e297–e302, 2020.

[18] S. Wolstenhulme and J. R. McLaughlan, "Lung ultrasound education: simulation and hands-on," *The British Journal of Radiology*, vol. 94, p. 20200755, Mar. 2021.

[19] A. Miller, "Practical approach to lung ultrasound," *BJA Education*, vol. 16, pp. 39–45, Feb. 2016.

[20] L. Demi, F. Wolfram, C. Klersy, A. De Silvestri, V. V. Ferretti, M. Muller, D. Miller, F. Feletti, M. Wełnicki, N. Buda, A. Skoczylas, A. Pomiecko, D. Damjanovic, R. Olszewski, A. W. Kirkpatrick, R. Breitkreutz, G. Mathis, G. Soldati, A. Smargiassi, R. Inchingolo, and T. Perrone, "New International Guidelines and Consensus on the Use of Lung Ultrasound," *Journal of Ultrasound in Medicine*, vol. 42, no. 2, pp. 309–344, 2023.

[21] G. Soldati, A. Smargiassi, R. Inchingolo, D. Buonsenso, T. Perrone, D. F. Briganti, S. Perlini, E. Torri, A. Mariani, E. E. Mossolani, and others, "Is there a role for lung ultrasound during the COVID-19 pandemic?," *Journal of Ultrasound in Medicine*, vol. 39, no. 7, p. 1459, 2020.

[22] M. Demi, R. Prediletto, G. Soldati, and L. Demi, "Physical Mechanisms Providing Clinical Information From Ultrasound Lung Images: Hypotheses and Early Confirmations," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 67, pp. 612–623, Mar. 2020.

[23] E. A. Fischer, T. Minami, I. W. Ma, and K. Yasukawa, "Lung Ultrasound for Pleural Line Abnormalities, Confluent B-Lines, and Consolidation: Expert Reproducibility and a Method of Standardization," *Journal of Ultrasound in Medicine*, vol. 41, no. 8, pp. 2097–2107, 2022.

[24] T. J. Marini, D. J. Rubens, Y. T. Zhao, J. Weis, T. P. O'Connor, W. H. Novak, and

K. A. Kaproth-Joslin, "Lung Ultrasound: The Essentials," *Radiology: Cardiothoracic Imaging*, vol. 3, p. e200564, Apr. 2021.

[25] L. Zhao and M. A. Lediju Bell, "A Review of Deep Learning Applications in Lung Ultrasound Imaging of COVID-19 Patients," *BME Frontiers*, vol. 2022, p. 9780173, Feb. 2022.

[26] J. Wang, X. Yang, B. Zhou, J. J. Sohn, J. Zhou, J. T. Jacob, K. A. Higgins, J. D. Bradley, and T. Liu, "Review of Machine Learning in Lung Ultrasound in COVID-19 Pandemic," *Journal of Imaging*, vol. 8, Mar. 2022.

[27] C. Baloescu, G. Toporek, S. Kim, K. McNamara, R. Liu, M. M. Shaw, R. L. McNamara, B. I. Raju, and C. L. Moore, "Automated Lung Ultrasound B-Line Assessment Using a Deep Learning Algorithm," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 67, pp. 2312–2320, Nov. 2020.

[28] G. Choy, O. Khalilzadeh, M. Michalski, S. Do, A. E. Samir, O. S. Pianykh, J. R. Geis, P. V. Pandharipande, J. A. Brink, and K. J. Dreyer, "Current Applications and Future Impact of Machine Learning in Radiology," *Radiology*, vol. 288, pp. 318–328, Aug. 2018.

[29] X. Wang, J. S. Burzynski, J. Hamilton, P. S. Rao, W. F. Weitzel, and J. L. Bull, "Quantifying lung ultrasound comets with a convolutional neural network: Initial clinical results," *Computers in Biology and Medicine*, vol. 107, pp. 39–46, Apr. 2019.

[30] R. J. G. van Sloun and L. Demi, "Localizing B-Lines in Lung Ultrasonography by Weakly Supervised Deep Learning, In-Vivo Results," *IEEE journal of biomedical and health informatics*, vol. 24, pp. 957–964, Apr. 2020.

[31] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized Gradient-based Visual Explanations for Deep Convolutional Networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 839–847, Mar. 2018.

[32] F. Mento and L. Demi, "Dependence of lung ultrasound vertical artifacts on frequency, bandwidth, focus and angle of incidence: An in vitro study," *The Journal of the Acoustical Society of America*, vol. 150, p. 4075, Dec. 2021.

[33] F. Mento and L. Demi, "On the influence of imaging parameters on lung ultrasound B-line artifacts, in vitro studya)," *The Journal of the Acoustical Society of America*, vol. 148, pp. 975–983, Aug. 2020.

[34] K. L. Anderson, J. M. Fields, N. L. Panebianco, K. Y. Jenq, J. Marin, and A. J. Dean, "Inter-Rater Reliability of Quantifying Pleural B-Lines Using Multiple Counting Methods," *Journal of Ultrasound in Medicine*, vol. 32, no. 1, pp. 115–120, 2013.

[35] S. Kulhare, X. Zheng, C. Mehanian, C. Gregory, M. Zhu, K. Gregory, H. Xie, J. McAn-

drew Jones, and B. Wilson, "Ultrasound-Based Detection of Lung Abnormalities Using Single Shot Detection Convolutional Neural Networks," in *Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation* (D. Stoyanov, Z. Taylor, S. Aylward, J. M. R. Tavares, Y. Xiao, A. Simpson, A. Martel, L. Maier-Hein, S. Li, H. Rivaz, I. Reinertsen, M. Chabanas, and K. Farahani, eds.), Lecture Notes in Computer Science, (Cham), pp. 65–73, Springer International Publishing, 2018.

[36] S. Roy, W. Menapace, S. Oei, B. Luijten, E. Fini, C. Saltori, I. Huijben, N. Chennakeshava, F. Mento, A. Sentelli, E. Peschiera, R. Trevisan, G. Maschietto, E. Torri, R. Inchingolo, A. Smargiassi, G. Soldati, P. Rota, A. Passerini, and L. Demi, "Deep Learning for Classification and Localization of COVID-19 Markers in Point-of-Care Lung Ultrasound," *IEEE Transactions on Medical Imaging*, vol. 39, pp. 2676 – 2687, Aug. 2020.

[37] W. Xue, C. Cao, J. Liu, Y. Duan, H. Cao, J. Wang, X. Tao, Z. Chen, M. Wu, J. Zhang, H. Sun, Y. Jin, X. Yang, R. Huang, F. Xiang, Y. Song, M. You, W. Zhang, L. Jiang, Z. Zhang, S. Kong, Y. Tian, L. Zhang, D. Ni, and M. Xie, "Modality alignment contrastive learning for severity assessment of COVID-19 from lung ultrasound and clinical information," *Medical Image Analysis*, vol. 69, p. 101975, Apr. 2021.

[38] G. R. Gare, A. Schoenling, V. Philip, H. V. Tran, B. P. deBoisblanc, R. L. Rodriguez, and J. M. Galeotti, "Dense Pixel-Labeling for Reverse-Transfer and Diagnostic Learning on Lung Ultrasound for COVID-19 and Pneumonia Detection," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1406–1410, Apr. 2021.

[39] F. Mento, U. Khan, F. Faita, A. Smargiassi, R. Inchingolo, T. Perrone, and L. Demi, "State of the Art in Lung Ultrasound, Shifting from Qualitative to Quantitative Analyses," *Ultrasound in Medicine & Biology*, vol. 48, pp. 2398–2416, Dec. 2022.

[40] C.-H. Tsai, J. van der Burgt, D. Vukovic, N. Kaur, L. Demi, D. Canty, A. Wang, A. Royse, C. Royse, K. Haji, and others, "Automatic deep learning-based pleural effusion classification in lung ultrasound images for respiratory pathology diagnosis," *Physica Medica*, vol. 83, pp. 38–45, 2021.

[41] L. Zhao, T. C. Fong, and M. A. L. Bell, "COVID-19 feature detection with deep neural networks trained on simulated lung ultrasound B-mode images," in *2022 IEEE International Ultrasonics Symposium (IUS)*, pp. 1–3, IEEE, 2022.

[42] U. Khan, S. Afrakhteh, F. Mento, N. Fatima, L. De Rosa, L. L. Custode, Z. Azam, E. Torri, G. Soldati, F. Tursi, V. N. Macioce, A. Smargiassi, R. Inchingolo, T. Perrone, G. Iacca, and L. Demi, "Benchmark methodological approach for the application of artificial intelligence to lung ultrasound data from COVID-19 patients: From frame to prognostic-level," *Ultrasonics*, vol. 132, p. 106994, July 2023.

[43] C. Milius, L. Jepson, D. Maclaskey, V. Pazdernik, and T. Kondrashova, "Development of Hands-On Skills in Diagnostics of Lung Diseases Using Ultrasonography in Undergraduate Medical Education," *Missouri Medicine*, vol. 120, no. 2, pp. 128–133, 2023.

[44] A. Dutta and A. Zisserman, "The VIA Annotation Software for Images, Audio and Video," in *Proceedings of the 27th ACM International Conference on Multimedia*, (Nice France), pp. 2276–2279, ACM, Oct. 2019.

[45] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," Mar. 2015.

[46] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," July 2012.

[47] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical Evaluation of Rectified Activations in Convolutional Network," Nov. 2015.

[48] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," Jan. 2017.

[49] S. A. Taghanaki, Y. Zheng, S. Kevin Zhou, B. Georgescu, P. Sharma, D. Xu, D. Comaniciu, and G. Hamarneh, "Combo loss: Handling input and output imbalance in multi-organ segmentation," *Computerized Medical Imaging and Graphics*, vol. 75, pp. 24–33, July 2019.

[50] S. Liu, Y. Wang, X. Yang, B. Lei, L. Liu, S. X. Li, D. Ni, and T. Wang, "Deep Learning in Medical Ultrasound Analysis: A Review," *Engineering*, vol. 5, pp. 261–275, Apr. 2019.

[51] J. Noble and D. Boukerroui, "Ultrasound image segmentation: a survey," *IEEE Transactions on Medical Imaging*, vol. 25, pp. 987–1010, Aug. 2006.

[52] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Medical Image Analysis*, vol. 63, p. 101693, July 2020.

[53] H. Mason, L. Cristoni, A. Walden, R. Lazzari, T. Pulimood, L. Grandjean, C. A. G. Wheeler-Kingshott, Y. Hu, and Z. M. Baum, "Lung Ultrasound Segmentation and Adaptation between COVID-19 and Community-Acquired Pneumonia," Aug. 2021.

[54] J. L. Herraiz, C. Freijo, J. Camacho, M. Muñoz, R. González, R. Alonso-Roca, J. Álvarez Troncoso, L. M. Beltrán-Romero, M. Bernabeu-Wittel, R. Blancas, A. Calvo-Cebrián, R. Campo-Linares, J. Chehayeb-Morán, J. Chorda-Ribelles, S. García-Rubio, G. García-de Casasola, A. Gil-Rodrigo, C. Henríquez-Camacho, A. Hernandez-Píriz, C. Hernandez-Quiles, R. Llamas-Fuentes, D. Luordo, R. Marín-Baselga, M. C. Martínez-Díaz, M. Mateos-González, M. Mendez-Bailon, F. Miralles-Aguiar,

R. Nogue, M. Nogué, B. Ortiz de Urbina-Antia, A. Oviedo-García, J. M. Porcel, S. Rodriguez, D. A. Rodríguez-Serrano, T. Sainz, I. M. Sánchez-Barrancos, M. Torres-Arrese, J. Torres-Macho, A. Trueba Vicente, T. Villén-Villegas, J. J. Zafra-Sánchez, and Y. Tung-Chen, "Inter-Rater Variability in the Evaluation of Lung Ultrasound in Videos Acquired from COVID-19 Patients," *Applied Sciences*, vol. 13, p. 1321, Jan. 2023.

[55] J. Gullett, J. P. Donnelly, R. Sinert, B. Hosek, D. Fuller, H. Hill, I. Feldman, G. Galetto, M. Auster, and B. Hoffmann, "Interobserver agreement in the evaluation of B-lines using bedside ultrasound," *Journal of Critical Care*, vol. 30, pp. 1395–1399, Dec. 2015.

[56] T. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare," *Future Healthcare Journal*, vol. 6, pp. 94–98, June 2019.

[57] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," *International Journal of Computer Vision*, vol. 128, pp. 336–359, Feb. 2020.

[58] G. Soldati, M. Demi, A. Smargiassi, R. Inchingolo, and L. Demi, "The role of ultrasound lung artefacts in the diagnosis of respiratory diseases," *Expert Review of Respiratory Medicine*, vol. 13, Jan. 2019.

[59] T. Perrone, G. Soldati, L. Padovini, A. Fiengo, G. Lettieri, U. Sabatini, G. Gori, F. Lepore, M. Garolfi, I. Palumbo, R. Inchingolo, A. Smargiassi, L. Demi, E. E. Mossolani, F. Tursi, C. Klersy, and A. Di Sabatino, "A New Lung Ultrasound Protocol Able to Predict Worsening in Patients Affected by Severe Acute Respiratory Syndrome Coronavirus 2 Pneumonia," *Journal of Ultrasound in Medicine: Official Journal of the American Institute of Ultrasound in Medicine*, vol. 40, pp. 1627–1635, Aug. 2021.

[60] L. Demi, W. van Hoeve, R. J. G. van Sloun, G. Soldati, and M. Demi, "Determination of a potential quantitative measure of the state of the lung using lung ultrasound spectroscopy," *Scientific Reports*, vol. 7, p. 12746, Oct. 2017. Number: 1.

[61] U.S. Food and Drug Administration, "510(k) Premarket Notification: Butterfly iQ Ultrasound System (K202406)," June 2023.

[62] D. Cheng and E. Y. Lam, "Transfer Learning U-Net Deep Learning for Lung Ultrasound Segmentation," Oct. 2021.

[63] F. Yang, G. Zamzmi, S. Angara, S. Rajaraman, A. Aquilina, Z. Xue, S. Jaeger, E. Papagiannakis, and S. K. Antani, "Assessing Inter-Annotator Agreement for Medical Image Segmentation," *IEEE access : practical innovations, open solutions*, vol. 11, pp. 21300–21312, 2023.

[64] K. A. Stewart, S. M. Navarro, S. Kambala, G. Tan, R. Poondla, S. Lederman, K. Barbour, and C. Lavy, "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, pp. 103–120, 2020.

[65] S. Shah, B. A. Bellows, A. A. Adedipe, J. E. Totten, B. H. Backlund, and D. Sajed, "Perceived barriers in the use of ultrasound in developing countries," *Critical Ultrasound Journal*, vol. 7, p. 28, Dec. 2015.

[66] M. M. Leo, I. Y. Potter, M. Zahiri, A. Vaziri, C. F. Jung, and J. A. Feldman, "Using Deep Learning to Detect the Presence and Location of Hemoperitoneum on the Focused Assessment with Sonography in Trauma (FAST) Examination in Adults," *Journal of Digital Imaging*, June 2023.

[67] A. Wijata, J. Andrzejewski, and B. Pyciński, "An Automatic Biopsy Needle Detection and Segmentation on Ultrasound Images Using a Convolutional Neural Network," *Ultrasonic Imaging*, vol. 43, pp. 262–272, Sept. 2021.

[68] E. S. Ebbini and G. Ter Haar, "Ultrasound-guided therapeutic focused ultrasound: Current status and future directions," *International Journal of Hyperthermia*, vol. 31, pp. 77–89, Feb. 2015.

Chapter 3

Curating a UK thyroid ultrasound dataset

As highlighted in Chapter 2, large, high-quality datasets are critical to training and evaluating AI models. Large collections of images can be acquired from digital clinical record systems; however, acquiring high-quality annotations for these datasets is challenging and time-consuming. This is especially true for ultrasound, where dynamic inspection of tissue structures from multiple angles is often critical to clinical interpretation, but only a few images are typically saved per examination, making retrospective analysis challenging without the report's contextual information.

On this theme, a tool was developed to extract clinical information from free-text radiology reports, enabling the curation of high-quality datasets for research and audit. This tool simplified the creation of domain-expert regular expression (regex) models, providing a simple, scalable, and transparent solution for structured data collection in highly regulated research environments.

This tool was applied to extract a comprehensive thyroid ultrasound dataset from National Health Service (NHS) records in Leeds Teaching Hospital Trust (LTHT), extracting information on the presence of thyroid nodules, diffuse thyroid disease, history of thyroid surgery, and British Thyroid Association (BTA) nodule gradings. These provide patient-level annotations for image classification and are useful as contextual information for segmentation. Once released, this will be the largest and most contextually rich dataset available for AI research into thyroid nodules and diffuse thyroid disease.

RADEX: a rule-based clinical and radiology data extraction tool demonstrated on thyroid ultrasound reports

Lewis Howell[1,2], Amir Zarei[3], Tze Min Wah[3,4], James H. Chandler[2], Shishir Karthik[3], Zara Court[3], Helen Hoi-Lam Ng[3], James R. McLaughlan[2,4]

[1]School of Computing, University of Leeds, UK
[2]School of Electronic and Electrical Engineering, University of Leeds, UK
[3]Radiology Department, Leeds Teaching Hospital Trust, Leeds General Infirmary, Leeds, UK
[4]Leeds Institute of Medical Research, University of Leeds, UK

European Radiology 35(8):4506-4517, August 2025.
DOI: 10.1007/s00330-025-11416-4

Abstract

Objectives: Radiology reports contain valuable information for research and audits, but relevant details are often buried within free-text fields. This makes them challenging and time-consuming to extract for secondary analyses, including training Artificial Intelligence (AI) models.

Materials and methods: This study presents a rule-based Radiology Data Extraction Tool (RADEX) to enable biomedical researchers and healthcare professionals to automate information extraction from clinical documents. RADEX simplifies the translation of domain expertise into regular-expression models, enabling context-dependent searching without specialist expertise in Natural Language Processing. Its utility was demonstrated in the multi-label classification of fourteen clinical features in a large retrospective dataset (n=16,246) of thyroid ultrasound reports from five hospitals in the United Kingdom (UK). A tuning subset (n=200) was used to iteratively develop the search strategy, and a hold-out test subset (n=202) was used to evaluate the performance against reference-standard labels.

Results: The dataset cardinality was 3.06, and the label density was 0.34. Cohen's Kappa was 0.94 for rater 1 and 0.95 for rater 2. For RADEX, micro-average sensitivity, specificity, and F1-score were 0.97, 0.96, and 0.94, respectively. The processing time was 12.3 milliseconds per report, enabling fast and reliable information extraction.

Conclusion: RADEX is a versatile tool for bespoke research and audit applications,

where access to labelled data or computing infrastructure is limited, or explainability and reproducibility are priorities. This offers a time-saving and freely available option to accelerate structured data collection, enabling new insights and improved patient care.

Key Points

Question: *Radiology reports contain vital information that is buried in unstructured free-text fields. Can we extract this information effectively for research and audit applications?*

Findings: *A rule-based RAdiology Data Extraction tool (RADEX) is described and used to classify fourteen key findings from thyroid ultrasound reports with sensitivity and specificity >0.95.*

Clinical relevance: *RADEX offers clinicians and researchers a time-saving tool to accelerate structured data collection. This practical approach prioritises transparency, repeatability, and usability, enabling new insights into improved patient care.*

Keywords: Natural Language Processing, Information extraction, Data annotation, Ultrasound, Thyroid

3.1 Introduction

Most information in medical reports is recorded in a narrative 'free-text' format [1], allowing reporters to write without restrictions. While this facilitates effective communication [2], it is often necessary to extract and structure relevant data for research and audit purposes [3]. Manual curation (chart review) is time-consuming, tedious, and expensive. Therefore tools that can extract information automatically are valuable for applications such as diagnostic surveillance, case retrieval, cohort building, quality improvement, and research applications [4–6]. In particular, annotating large datasets for training AI models has been identified as a priority for clinical research [7].

Recently, developments in Natural Language processing (NLP) such as Machine Learning (ML) and Deep Learning (DL), especially Large Language Models (LLMs) such as Bidirectional Encoder Representations from Transformers (BERT) and ChatGPT, have renewed interest in automated information extraction [8, 9]. In particular, domain-specific language models, such as BioBERT [10], ClinicalBERT [11], PubMedBERT [12] and radiology-focused RadBERT [13] have shown promising results for tasks such as text classification, named entity recognition, and summarisation, re-training the base BERT model using clinical text or biomedical literature [14]. For example, in 2022 Wood *et al.* applied BioBERT to classify abnormalities in head Magnetic Resonance Imaging (MRI) reports, achieving an F1-score of 0.962 and 0.930 when evaluated against reference standard report and image labels [15]. Also in 2022, Visser *et al.* extracted actionable findings from radiology reports including X-ray, Ultrasound, Computed Tomography (CT), and MRI, scoring a sensitivity

of 0.841 and specificity of 0.990 [16].

Despite the potential of ML/DL methods, a lack of domain-specific training data and concerns around security, reliability, explainability, and bias have limited their widespread implementation in practice [7, 17]. For these reasons, rule-based methods, which use human knowledge to create computational models, remain popular in medicine [4, 6]. Well-designed rule-based models can achieve similarly high performance to ML/DL techniques [5, 18] while offering improved transparency, replicability and flexibility. The main challenge of rule-based approaches is that developing robust models is difficult and usually requires interdisciplinary collaboration between technical experts who build the model and clinical experts who provide expertise in the domain of the data [17, 19]. Existing databases such as the Unified Medical Language System (UMLS) [20] or Radiology Lexicon (RadLex) [21] can be applied in 'dictionary-based methods' to identify medical terms from large ontologies. However, relying on dictionaries alone often cannot capture the clinical nuances, language variability, and context required for report classification in bespoke research and audit tasks [22]. In these cases, manual knowledge engineering is needed to define custom rules that meet task-specific requirements. These hand-crafted rules are usually defined using the established framework of Regular Expressions (Regex) – sequences of letters and special characters that define patterns for text-matching [23]. Regex offers an efficient and flexible method for pattern-matching, but expressions are challenging to construct and maintain as there is currently no standardised method for their generation and evaluation [24].

This study presents a novel rule-based tool for RADEX which standardises and simplifies the construction of regular expression models using a high-level syntax and iterative refinement protocol. This enables effective translation of clinical domain knowledge into computational models for automated data annotation and report classification. RADEX encourages a systematic approach to form robust, comprehensive, and reproducible search strategies; iteratively refining regular expression models to allow reliable data extraction from unstructured radiology reports (X-Ray, MRI, CT, Ultrasound) and other clinical documents (clinical notes, incident reports, pathology and cytology reports). The tool addresses challenges associated with previous methods, which lack transparent and reproducible reporting, robust evaluation, or rely on access to large labelled datasets [25]. RADEX is aimed at biomedical researchers and healthcare professionals with no specialist prior experience in NLP. It provides a practical, fast, flexible, and free solution which can be integrated with existing clinical systems.

The typical RADEX workflow is demonstrated on a large dataset of thyroid ultrasound reports, a useful example of unstructured reports with variable language [26] and a broad range of diagnostic indications [27]. Thyroid cancer, thyroid nodules, and functional or autoimmune disease have been investigated previously. For example, in 2020 Chen *et al.* used the rule-based tool cTAKES [28] to identify thyroid nodules and classify five

characteristics, detecting nodules with 91.9% accuracy on a dataset of 153 reports [29]. In 2022, Dedhia *et al.* also applied cTAKES, classifying 243 ultrasound reports for nine classes with an F1-score of 0.8122 (precision = 0.74, recall = 0.90) [26]. Recently, Pathak *et al.* applied transformer-based models to extract 16 thyroid nodule characteristics in 490 reports, achieving a maximum F1-Score of 0.9495 when considering partial matches on a relatively small dataset of 98 test reports [30]. In this study, 16,246 thyroid ultrasound reports were classified for 14 features, covering common diffuse and focal pathologies as well as nodule gradings. Results were evaluated against 202 reference-standard labels with consensus from three independent expert reviewers. The anonymised dataset produced could be used for purposes such as service evaluation, improving consistency of reporting, or downstream research such as training of AI computer vision models [29, 31].

The purpose of this study is to develop and evaluate RADEX, a rule-based data extraction tool that leverages domain expertise to efficiently and accurately extract structured information from free-text clinical reports. The tool aims to provide a scalable, user-friendly, and time-saving solution for information extraction and report classification in large datasets, which is demonstrated on thyroid ultrasound reports.

3.2 Methods and materials

3.2.1 Dataset

Formal ethical approval was waived for this retrospective study of collated clinical and imaging reports in accordance with the Institutional Health Research Authority Framework. All radiology reports used in this study were obtained from a retrospective dataset, including all examinations tagged Ultrasound Neck (UNECK), Ultrasound thyroid (UTHYD), or Ultrasound thyroid and parathyroid (UTHPY), extracted from the Computerised Radiology Information System (CRIS) between 01/01/2015 - 31/12/2019 across five different hospitals in the same geographical area in the United Kingdom. Paediatric patients (<18 years) were excluded, and multipart examinations were merged into a single pseudoidentifier. Where multiple reports were present for the same pseudoidentifier, only the most recent report was taken to avoid any possibility of data leakage. This resulted in a dataset comprising 16,246 reports from unique patients. All reports were in an unstructured free-text format and included clinical history, report body, and findings, with embedded headers dividing these sections. The average (mean) total report length was 93 words, and the dataset contained reports written by 222 unique reporters with varying roles and experience, including general or specialist sonographers, radiology trainees, and general or specialist certified radiologists.

To develop the search strategy, a subset of 200 reports was sampled at random as a tuning set. A further 202 reports were sampled as a holdout test dataset, allowing unbiased evaluation of the expected performance and ensuring there was no overfitting. These

402 reports were annotated independently by three reviewers: an experienced radiologist (UK consultant grade; United States (U.S.) attending equivalent) and two foundation doctors with experience reading and interpreting radiology reports. In ambiguous cases, the experienced radiologist served as the adjudicator for reference-standard labels, reviewing the report's associated ultrasound images using the Patient Archive and Communication System (PACS) as necessary to ensure the most accurate annotations possible. Fourteen classes were chosen to cover the features of interest, including common diffuse and focal pathologies observed in ultrasound examination of the thyroid [32–34]. Since reports were from the UK, BTA gradings were used to stratify thyroid nodules based on their malignancy risk [35].

The classes chosen were:

- Thyroid examination - thyroid examination conducted.
- Lymph node examination - lymph node examination conducted.
- Thyroid nodule(s) - one or more thyroid nodules.
- Multiple thyroid nodules - multiple discrete nodules or Multinodular Goitre (MNG).
- Solitary thyroid nodule - finding of thyroid nodule(s) but not multiple nodules or MNG.
- Altered thyroid echotexture - indicative of diffuse thyroid disease such as thyroiditis.
- Goitre - enlarged thyroid including simple and multinodular goitre.
- Previous thyroid surgery - including hemi and total thyroidectomy.
- BTA U1 - British Thyroid Association grading for normal thyroid
- BTA U2 - British Thyroid Association grading for benign thyroid nodule
- BTA U3 - British Thyroid Association grading for indeterminate/equivocal thyroid nodule
- BTA U4 - British Thyroid Association grading for suspicious thyroid nodule
- BTA U5 - British Thyroid Association grading for malignant thyroid nodule
- Normal thyroid - mention of thyroid examination but no finding of thyroid nodule(s), altered echotexture, goitre, or previous thyroid surgery.

3.3 RADEX tool

The RADEX tool is a general-purpose software package aimed at medical professionals and biomedical researchers who require a simple method to extract high-quality structured datasets from free-text clinical reports. RADEX enables rule-based extraction of relevant information, constructing regular-expression models to classify reports and provide interpretable predictions.

Regular expressions are a standard technique implemented in many programming languages (e.g. Java, Python, Perl, Ruby), which offer flexible and efficient searching in large

datasets [24]. These expressions are sequences of characters and special symbols that explicitly define searches, including wildcards, character classes, quantifiers, and boundary matches, which are useful when searching for medical terms in context [23]. For example, in Python the regular expression:

(1) "\bthyr\w*\b",

"\b" represents a word boundary, "\w" matches one or more word characters (letters or numbers), and "*" means "zero or more times". This pattern would match any word that starts with "thyroid", such as "thyroid", "thyroidectomy", or "thyroiditis". In the expression:

(2) "\bthyroid\b(?:\W+\w+){0,3}?\W+\bnodules\b|\bmultinodular\b",

the non-capturing group "(?:\W+\w+){0,2}" matches 0 to 2 words between "thyroid" and "nodules", and the "|" means "OR", meaning the expression matches the word "multinodular" as well as phrases such as "thyroid nodules" and "thyroid contains multiple nodules". While this framework allows highly flexible pattern matching, regex is complex to learn and becomes difficult to interpret as the complexity of patterns increases.

RADEX centres around a user-friendly 'search strategy', which is created with a high-level syntax of keywords, modifiers, and logical connectors. For example, in RADEX, the expressions 'thyr*' and 'thyroid NEAR2 nodules OR multinodular' are equivalent to regular expressions (1) and (2) above. This enables non-expert users to create robust regular-expression-based models that consider context, differentiate between similar terms, and account for variations in spelling, phrasing, and report structure. This approach to developing a search strategy takes inspiration from popular search tools for biomedical literature databases such as PubMed Advanced Search Builder and Ovid Advanced Search, which are familiar to many biomedical researchers and medical professionals, lowering the barrier to entry for using the tool. Similarly to the literature search process of systematic reviews, the RADEX methodology promotes rigour, comprehensivity, and reproducibility in data extraction [36]. A diagram of the typical RADEX workflow is shown in Fig. 3.1, and each step is described in more detail in the sections below.

RADEX was developed in Python 3.9 and is available for Windows, MAC, and Linux OS. The source code for the project is available on GitHub under the BDS 3-Clause Licence, allowing its free use and modification. Documentation, installation instructions, and instructions for developers are available online. Also included is a labelled dataset of 110 synthetic thyroid ultrasound reports, generated using OpenAI's ChatGPT-3.5 [37]. This provides a relevant quick-start example of the RADEX workflow on data that can be readily shared, including a demonstration of a prototype Graphical User Interface (GUI).

3.3.1 Task formulation

The first stage for information extraction is task formulation, where the key requirements and datasets are identified. This includes defining the clinical question, population of inter-

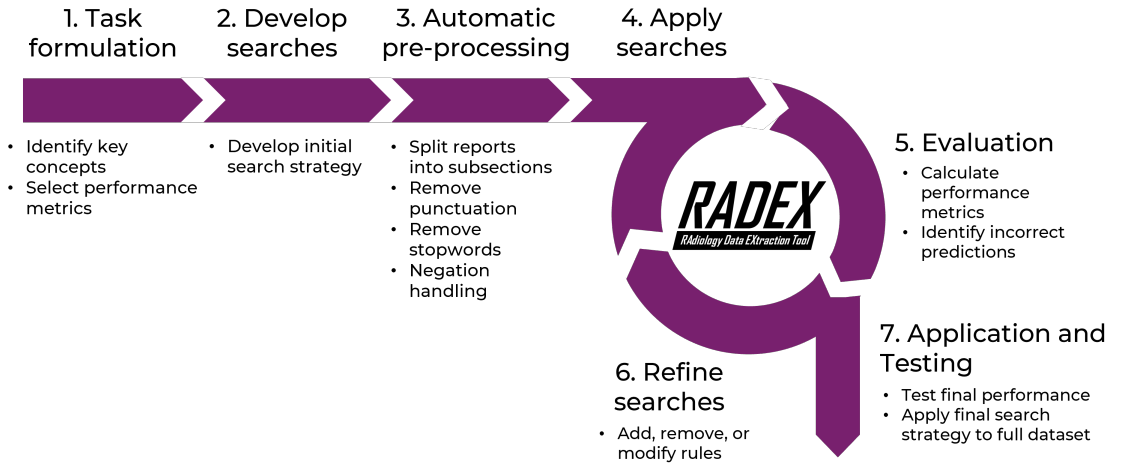


Figure 3.1: Typical workflow for the Radiology Data Extraction Tool (RADEX), including iterative development of searches.

est, and concepts needed to classify reports. For best performance, the task requirements should be sufficiently narrow and well-defined to ensure precise mapping between the tool's outputs and the intended clinical concepts, reducing ambiguity and improving model performance. The user should also consider which classification metrics are most relevant for performance evaluation, depending on the task requirements and characteristics of the dataset [38, 39].

3.3.2 Develop search strategy

The search strategy is then defined using RADEX's simple but flexible syntax, which leverages domain knowledge to extract structured information without requiring the user to have prior expertise in NLP or regular expressions. This RADEX search strategy comprises keywords, modifiers, and logical connectors which define clinical concepts.

Keywords include synonyms, abbreviations, and other terms related to the concept, such as medications, procedures, and diseases. Relevant keywords can be identified using expertise in the domain of the data or using existing biomedical ontologies such as UMLS [20] and RadLex [21]. Modifiers broaden the search scope, accommodating variations in spelling, pleuralised words, and typographical errors. For example, the asterisk ('*') modifier represents any number of characters, which is useful since many related clinical terms share a common root, suffix, or prefix, originating from their Greek or Latin etymology [22]. For example, 'neur*' will match a variety of terms relating to nerves and the nervous system, such as 'neuron', 'neuropathy', and 'neurofibromatosis'. Similarly, '*therap* would match 'radiotherapy' and 'therapeutic'. The question mark ('?') represents any optional single character, most useful for accommodating pluralisation and spelling variation. For example, 'h?emoglobin' would match 'haemoglobin' and 'hemoglobin', and 'wom?n' would match 'woman' or 'women'. The order-independent and order-dependent proximity operators 'NEAR/X' and 'THEN/X' identify terms that are separated by 'X' or fewer words. For

example, 'thyroid NEAR/2 *nodul*' matches phrases such as 'thyroid nodules', 'multinodular thyroid', or 'thyroid contains multiple nodules'. Expressions can be combined using the logical connectors 'AND', 'OR', 'NOT' and 'EXCEPT' (equivalent to 'AND NOT') to build complete definitions for concepts. For example, an expression for identifying thyroid nodules could be: 'thyroid NEAR/2 *nodul* OR (node? EXCEPT lymph node?)'. Here, a search for the word thyroid in proximity to nodules is combined with a search for node/nodes, excluding lymph nodes. Combining these expressions allows the development of search strategies to reliably classify increasingly complex concepts.

3.3.3 Pre-processing

Before running the search strategy, the reports are pre-processed to ensure consistent results. Filtering can be applied to enforce inclusion/exclusion criteria such as date ranges, or patient demographics such as age. Reports are then split into relevant sections based on selected headers, and the text is then standardised: removing stopwords, selecting punctuation, trailing whitespace, and converting the text to lowercase. Sentence tokenisation is applied to break the text into individual sentences. Negated phrases are handled using Chapman's NegEx algorithm [40]. This uses a clinically relevant list of pre-determined modifiers to identify negation modifiers and their scope. For example, the phrases 'there are no thyroid nodules' and 'thyroid nodules were not observed' are negated by the pre- and post- negation terms 'no' and 'not' and are automatically excluded from potential matches.

3.3.4 Apply search strategy

To process the text, the RADEX search strategy is converted to a regular-expression model in the software backend. A grammar function represents the permitted operators and their precedence (NOT > AND > OR), detecting unmatched parentheses, unexpected symbols, or incorrect syntax. Once verified, the search strategy is recursively evaluated, converting each sub-expression to a regular-expression applied to search the text. Matching words and phrases are extracted, and their positional indexes can be used to highlight the report. Depending on the presence or absence of matches and the applied logic, reports are classified for one or more labels. This process, similar to an inference engine, enables traceability at the report level, enhancing explainability and facilitating the iterative development of searches.

3.3.5 Evaluation

The model is then evaluated using an annotated tuning subset of the data. Label-based metrics such as classification accuracy, sensitivity (true positive rate, recall), specificity (true negative rate), precision (positive predictive value), and F1-score (harmonic mean of precision and recall) are calculated to assess the quality of the search strategy. In cases where multiple labels are assigned to each report, the macro- and micro-averaged metrics,

in addition to the example-based Hamming loss (total fraction of wrong labels) and exact match ratio (percentage of reports with all labels predicted correctly) are calculated [41, 42]. In all cases, no single metric captures all important information, and the most useful metrics depend on the study objectives and dataset characteristics [38].

3.3.6 Refine search strategy, application, and testing

The search strategy is then refined based on a review of incorrect predictions in the tuning set, adding, removing, or modifying rules as necessary to build a complete search strategy. RADEX identifies false positive and false negative predictions using reference-standard labels, highlighting matching phrases within the report body to assist this refinement process. The improved searches are then re-applied and re-evaluated on the tuning set, improving the search strategy over several cycles. Once satisfactory performance is achieved, the searches should be assessed on a holdout test dataset. This allows unbiased quantification of the expected performance on the full dataset, assessing the generalisability of the model and ensuring there is no overfitting.

3.4 RADEX for thyroid ultrasound reports

While RADEX is a general tool for clinical information extraction, this study demonstrates its application in the context of thyroid ultrasound reports. Combining multi-label classification ultrasound reports with linked ultrasound images enables the curation of a large labelled dataset of thyroid ultrasound scans, suitable for developing computer vision models in the future. The primary metric for assessing RADEX performance on this dataset was the F1-Score, which was chosen since both high precision and recall were important. A target micro-average F1-score of 0.90 was selected based on the study requirements and the inter-annotator agreement.

Twelve classes were searched directly in the report body, including thyroid examination, lymph node examination, thyroid nodule(s), multiple thyroid nodules, altered thyroid echotexture, goitre, and BTA U1-U5. In addition to the report body, the clinical history was included in searches for previous thyroid surgery. The classes solitary thyroid nodule and normal thyroid were derived from using simple logic, where mention of thyroid nodule(s) and not multiple thyroid nodules implies a single nodule, and normal thyroid is defined as the absence of any other focal or diffuse thyroid abnormality. Simple post-processing rules were applied to ensure consistency in the final dataset, for example, finding a BTA U2 to U5 grading must imply the presence of a thyroid nodule. Once satisfactory scores for the sensitivity and specificity were achieved, the same searches were applied to the test dataset to ensure that the terms used were general enough to work on new data. The final search strategy is shown in Table 3.1.

The initial search strategy was constructed with interdisciplinary input from experienced head and neck radiologists as well as UMLS. These searches were then applied to the

Table 3.1: Search strategy for neck and thyroid ultrasound dataset. The symbols &, |, ¬, and $\sim X$, are shorthand for 'AND', 'OR', 'NOT', and the bidirectional word proximity operator NEAR/X, with adjacency X as an integer. The modifiers '*' and '?' are wildcard modifiers.

Class	**Search strategy**
Thyroid examination	thyroid*
Lymph node examination	lymph* | (node & ¬thyroid* node),
Thyroid nodules(s)	(nodul* & ¬(nodul*∼2cartilage | nodul*∼2parotid | nodule*∼2salivary | nodul*∼2submandibular)) | (cyst* & ¬(thyroglossal cyst | cyst*∼2parotid | cyst*∼2salivary | cyst*∼2submandibular | cyst*∼1sebaceous | cyst*∼1inclusion | cyst*∼1lymph* | cyst*∼1node)) | thyroid∼4lesion* | thyroid∼1node* | adenoma
Multiple Thyroid Nodules	(nodules & ¬(nodules∼2cartilage | nodules∼2parotid | nodules∼2salivary | nodules∼2submandibular)) | (cysts & ¬(thyroglossal cysts | cysts∼2parotid | cysts∼2salivary | cysts∼2submandibular | cysts∼2sebaceous | cysts∼2inclusion | cysts∼2lymph* | cysts∼2node)) | thyroid∼4lesions | thyroid∼1nodes | multi?nodul* | nodul*∼2both | other∼2nodul* | second*∼2nodul* | (superior∼3nodul* & inferior∼3nodul*) | (nodule*∼3left & nodule*∼3right) | (nodule*∼3left & nodule*∼3isthmus) | (nodule*∼3right & nodule*∼3isthmus) | mng
Altered thyroid echotexture	thyroiditis | grav?s | heterogen*∼2echo* | heterogen*∼2thyroid | heterogen*∼1texture | inflamed*
Goitre	goit?r? | mng | multi?nodul* thyroid | enlarged∼2thyroid | (enlarged∼2gland & ¬enlarged∼2parathy*) | enlarged∼2lobe | hyperplasia
Previous thyroid surgery	thyroidectomy | lobectomy | isthmusectomy | (surgery & ¬refer*∼2surgery) | resect* | incomplete∼2thyroid | post?op | bed | hemiagenesis
BTA U1	u1
BTA U2	u2 & ¬(u2?u3 | u2 ∼1 ?3)
BTA U3	u3 & ¬(u3?u4 | u3 ∼1 ?4 | u2?u3 | u2 ∼1 ?3)
BTA U4	u4 & ¬(u4?u5 | u4 ∼1 ?5 | u3?u4 | u3 ∼1 ?4)
BTA U5	u5 & ¬(u4?u5 | u4 ∼1 ?5)
Solitary thyroid nodule	"Thyroid nodules(s)" & ¬"Multiple Thyroid Nodules"
Normal thyroid	"Thyroid examination" & ¬("Thyroid nodule(s)" | "Altered thyroid echotexture" | "Goitre" | "Previous thyroid surgery")

clinical dataset, and entries where there was disagreement between the reference standard and predicted labels were flagged for manual review. The search criteria were iteratively refined in several cycles, adding or removing phrases from the search strategy to reduce the number of incorrect predictions. All computation was conducted on a research laptop with average specifications (AMD 4600H Central Processing Unit (CPU), 16GB Random Access Memory (RAM)).

3.5 Results

In the reference-standard labels, the average number of labels per report (cardinality) was 3.06, and the percentage of possible labels assigned on average per report (label density), was 34%. The inter-annotator reliability was assessed using Cohen's Kappa, which was 0.94 for rater 1 and 0.95 for rater 2, indicating a strong agreement with the adjudicator. The time taken to annotate 200 reports was approximately 2.5 h per annotator, which if extrapolated to the full dataset, suggests approximately 203 h of manual labelling would be needed per annotator.

Applying the RADEX search strategy in Table 1 enabled classification and highlighting of the reports for the 14 different classes (Fig. 3.2). Applying the iterative tuning approach to the development of the search strategy increased the F1-score across all classes (Fig. 3.2), improving the micro-average by 0.222. Generally, false positive predictions arose from semantic context errors, negated, and hypothetical sentences, while false negatives were due to missing keywords and negation errors.

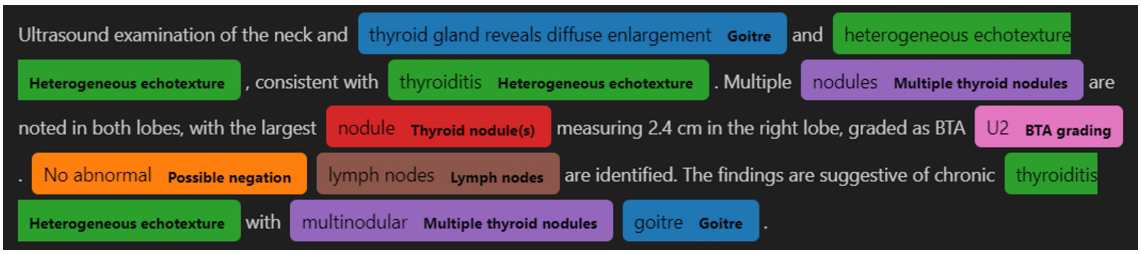


Figure 3.2: A synthetic neck and thyroid ultrasound report, highlighted using a rule-based radiology data extraction tool. This example was correctly classified as positive for goitre (blue), altered (heterogeneous) thyroid echotexture (green), multiple thyroid nodules (purple), and lymph node examination (brown). The phrase 'no abnormal' was matched as a possible negation modifier to the term 'lymph nodes'

Using RADEX, the Hamming loss for the nine main classes (excluding the BTA gradings) was 0.0418, suggesting, on average, less than 5% of the total labels were incorrectly predicted. The harshest metric, the exact match ratio, was 0.733, showing the percentage of reports with all nine classes correct. Label-based metrics and their micro- and macro-averages are given in Table 3.2, scoring each class individually, with all classes except solitary thyroid nodule and goitre having F1 scores > 0.8. Semantically, solitary thyroid nodule was the most complex concept to predict, since it was difficult to distinguish cases

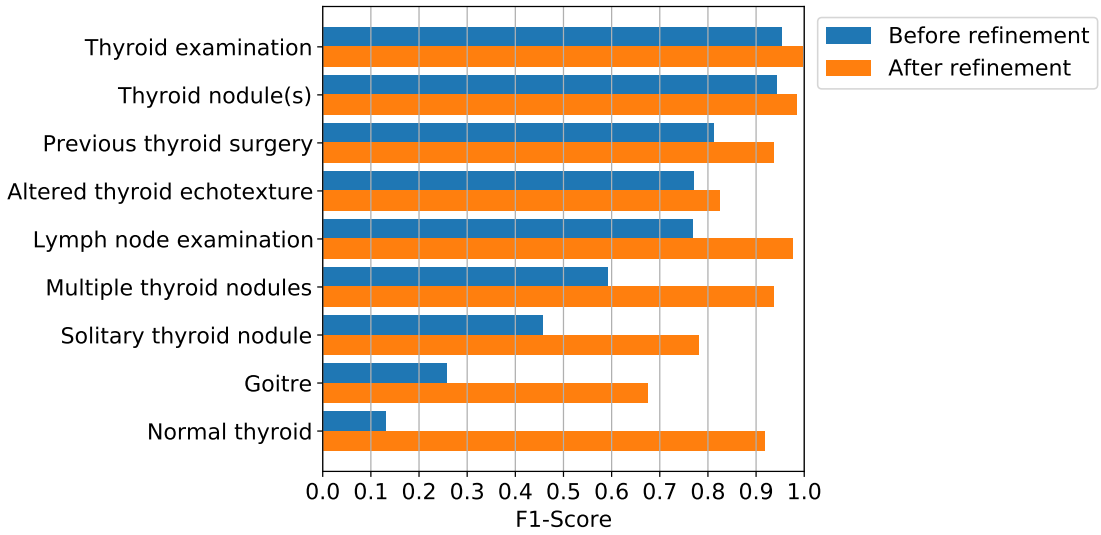


Figure 3.3: F1-score for rule-based predictions on the holdout test clinical dataset, before and after refining the searches.

where a single nodule was mentioned multiple times, from cases with multiple nodules present in the thyroid. Notably, goitre had a high sensitivity with no false positives in the test set but lower precision. This is because of the conservative approach taken to labelling this class and the imprecise language used in its reporting. For example, phrases such as 'thyroid gland appears slightly enlarged' were commonly predicted as false positives, requiring additional context outside the report body to predict these cases accurately. Overall, the micro-average sensitivity and specificity across the nine classes were 0.971 and 0.957, respectively, exceeding the targets set at the start of the study.

Table 3.2: Scores for multilabel classification of neck and thyroid ultrasound reports with regular expressions.

	Precision (PPV)	Sensitivity (Recall/TPR)	Specificity (TNR)	F1-Score
Thyroid examination	0.994	1.000	0.978	0.997
Normal thyroid	0.846	1.000	0.990	0.917
Previous thyroid surgery	0.902	0.974	0.976	0.937
Thyroid nodule(s)	0.966	1.000	0.954	0.983
Multiple thyroid nodules	0.943	0.932	0.956	0.937
Solitary thyroid nodule	0.719	0.852	0.949	0.780
Altered thyroid echotexture	0.824	0.824	0.984	0.824
Goitre	0.509	1.000	0.851	0.675
Lymph node examination	0.982	0.971	0.900	0.977
Micro average	0.917	0.971	0.957	0.943
Macro average	0.854	0.950	0.949	0.892

Importantly, the recall was generally high, meaning the rate of false negatives was low. This is significant since a manual review of positive predictions in the predictions can eliminate false positives, but finding missed reports among the full dataset (false negatives) requires a review of more total reports. The true and false positives and negatives are visualised

with confusion matrices in Fig. 3.4. In addition to these nine classes, the BTA gradings for thyroid nodules were predicted well, with micro-average F1-Score of 0.967.

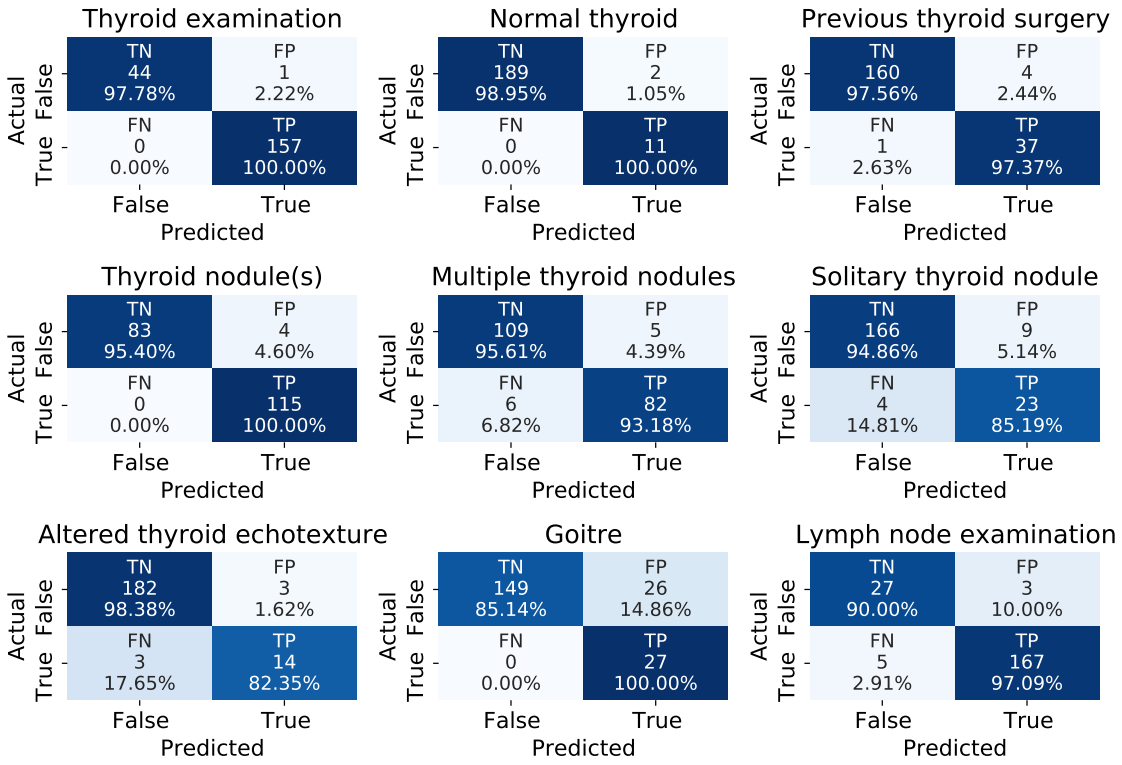


Figure 3.4: Class-wise confusion matrices for multilabel data. True positive (TP), True negative (TN), False positive (FP) and False Negative (FN) counts and normalised scores for each class.

This method is also relatively computationally inexpensive, requiring no specialist hardware or cloud infrastructure. The total processing time for 16,246 reports was 200s, or 12.3 ms per report. As expected, processing time increased linearly with the number and average length of the reports, as well as increasing with the number and complexity of searches.

3.6 Discussion

This study presented RADEX, a rule-based data extraction tool for classification and information extraction in clinical free-text reports. This was demonstrated on a large dataset of 16,246 thyroid ultrasound reports, classifying 14 clinical findings and scoring a micro-average sensitivity of 0.971 and specificity of 0.957, which exceeded the targets set for the study. RADEX processed the full dataset in just 200 s, which represented a large time-saving benefit when compared to the 203 h estimated for manual annotation. Thyroid ultrasound reports presented an interesting case study for classification due to their variable language [26] and a broad range of clinical concepts [27]. Compared with

similar work applying NLP to thyroid ultrasound reports, the methods used improved on previous rule-based approaches and performed approximately as well as DL methods. Using dictionary-based approaches to classify thyroid nodules in reports, Dedhia *et al.* reported an F1-score of 0.812 [26], and Chen *et al.* scored 0.919 in accuracy [29]. This study achieved an F1-Score of 0.943 and an accuracy of 0.980. The F1-Score was also within one percentile of the DL transformer-based approach reported by Pathak *et al.* which was 0.945 [30]. However, it should be noted that since these studies use different datasets and class definitions, it is difficult to compare these results directly. In this study, the micro-average F1-score of 0.943 illustrates a low false positive and false negative rate, which is important since previous approaches using handcrafted regular expressions often perform poorly in recall [24]. This improvement was attributed to the flexible, iterative refinement process used, enabling the creation of custom rules that were sufficiently general to capture language variations among the 222 reporters and five different healthcare organisations in the dataset.

The main advantages of RADEX are that accurate results can be obtained quickly, results are transparent and reproducible, and the tool can be used for free with no special hardware requirements. The tool prioritises usability, leveraging the power of regular expressions but without requiring the user to understand their complex syntax. This simplifies the process of creating robust regular-expression models, allowing users to annotate large datasets with no specialist NLP expertise. The iterative refinement process ensures the search strategy is comprehensive, and the independent evaluation ensures the final model is generalisable within the domain of the dataset. The final search strategy can also be easily exported and shared without compromising patient confidentiality, supporting reproducibility, which is critical in clinical NLP applications [25].

Compared with existing methods, RADEX provides greater flexibility than dictionary-based approaches and better explainability than DL methods at the cost of requiring the search strategy to be manually defined. Dictionary-based methods can extract recognised concepts such as disease names, symptoms, and treatments from clinical ontologies. However, they struggle with capturing specific, context-dependent, or nuanced information, as dictionaries alone often cannot capture all required patterns to define a concept [22]. ML/DL methods automatically learn patterns and relationships by analysing text datasets, allowing for more accurate extraction of nuanced information. However, they require large labelled datasets, expensive computational resources, and often lack transparency, making it difficult to explain or reproduce their outputs in clinical settings. RADEX bridges these gaps by simplifying the process of defining and refining custom rulesets. This makes it suitable for standalone use or in conjunction with existing dictionary-based or ML/DL methods. For example, RADEX could be used to supplement dictionary-based methods in cases of difficult-to-extract concepts, or to provide an explainable baseline for DL/ML models to be evaluated against.

This explainability is important to support decision-making [6] and to satisfy ethical considerations and data governance standards related to accountability [43]. While regular-expression models alone can be difficult to decipher; the high-level syntax used to develop the RADEX search strategy is easy to understand, aiding interpretability of model predictions. This is a major advantage compared with ML/DL methods, where the non-linear operations in 'black box' DL models make explainability difficult [44]. In particular, since LLMs rely on statistical patterns in their training data, they lack a true understanding of underlying medical concepts, which can lead to inaccurate predictions [45]. The human-created search strategy of RADEX can be modified to correct specific errors, meet changing requirements, or adapt to terminology changes in different regions or over time [2]. This is not the case with DL models, which require re-training using additional labelled training. This is resource-intensive and does not guarantee performance improvement. Importantly, RADEX is also scalable to large datasets and has no special hardware requirements. This helps to alleviate concerns regarding cost and data security since all processing is performed locally, rather than relying on secure cloud computing infrastructure, which restricts the use of AI in clinical settings [5].

The main limitation of the RADEX method is its limited applicability to complex semantic tasks, where understanding the wider context and the subtlety of human language is important. In these cases, DL or hybrid techniques may be more appropriate [19], but rule-based approaches are still useful as a baseline to justify the associated implementation challenges and explainability tradeoff [6]. Inter-annotator agreement also presents a challenge, where different interpretations of the same data can introduce uncertainty [23]. This is reflected by the inter-annotator agreement (Cohen's Kappa = 0.94), showing a minority of reports containing ambiguous findings that required cross-referencing with clinical images in the reference-standard dataset. While this approach minimises uncertainty, it does limit the theoretical maximum prediction accuracy as this information is not available to the model, requiring a multimodal approach. A better strategy may have been to quantify uncertainty using a method such as NegBio [46], allowing a secondary review of unclear cases. Furthermore, while test set metrics give an unbiased estimate of the expected performance on the larger dataset, this may not represent the wider generalisability to external datasets, since language use can vary significantly between institutions and over time. Since data was collected in the UK, all reports were in British English, and thyroid nodule malignancy risk was given by the British Thyroid Association (BTA) grading system. However, similar principles could be applied to reports in other languages or different grading systems, such as the European Thyroid Association (ETA) and American College of Radiologists (ACR) Thyroid Imaging Reporting and Data System (TI-RADS) [47, 48]. This would require further refinement of searches, which can be achieved using the iterative RADEX method.

Overall, RADEX provides a practical alternative to dictionary-based and ML/DL methods for free-text information extraction, effectively leveraging clinical expertise to create custom rule-based models for bespoke research and audit tasks. This was demonstrated in the

classification of thyroid ultrasound reports but could be similarly applied to other radiology reports and clinical documents such as clinical notes, incident reports, pathology, and cytology reports in the future. Micro-average sensitivity, specificity, and F1-Score were 0.971, 0.957, and 0.943, respectively, and the Hamming loss was 0.0418. No specialist hardware was required, and the technique was shown to be scalable to large datasets, taking 12.3 ms/report. This method offers a practical alternative or precursor to other approaches, such as ML and DL, especially where access to labelled training data, computing infrastructure, explainability, or reproducibility are concerns. This provides a time-saving and freely available tool for future studies in clinical decision support, quality improvements, unbiased identification of disease cohorts for research studies, query-based case retrieval, and research tasks such as annotating medical images for AI analysis.

3.7 Acknowledgements

Thanks to Adam Hastings for his contributions to the graphic user interface branch of the software repository. This work used data provided by patients and collected by the UK National Health Service as part of their care and support.

3.8 Funding

This study has received funding by the Engineering and Physical Sciences Research Council (EPSRC) from the following awards, EP/S024336/1 and EP/V04799X/1.

Bibliography

[1] D. L. Rubin and C. E. Kahn, "Common Data Elements in Radiology," *Radiology*, vol. 283, pp. 837–844, June 2017.

[2] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, and H. Liu, "Clinical information extraction applications: A literature review," *Journal of Biomedical Informatics*, vol. 77, pp. 34–49, Jan. 2018.

[3] M. Chen, R. Ball, L. Yang, N. Moradzadeh, B. Chapman, D. Larson, C. Langlotz, T. Amrhein, and M. Lungren, "Deep learning to classify radiology free-text reports," *Radiology*, vol. 286, no. 3, pp. 845–852, 2018.

[4] N. Linna and C. E. Kahn, "Applications of natural language processing in radiology: A systematic review," *International Journal of Medical Informatics*, vol. 163, p. 104779, July 2022.

[5] E. Pons, L. M. M. Braun, M. G. M. Hunink, and J. A. Kors, "Natural Language Processing in Radiology: A Systematic Review," *Radiology*, vol. 279, pp. 329–343, May 2016.

[6] A. Casey, E. Davidson, M. Poon, H. Dong, D. Duma, A. Grivas, C. Grover, V. Suárez-Paniagua, R. Tobin, W. Whiteley, H. Wu, and B. Alex, "A systematic review of natural language processing applied to radiology reports," *BMC Medical Informatics and Decision Making*, vol. 21, p. 179, June 2021.

[7] H. Wu, M. Wang, J. Wu, F. Francis, Y.-H. Chang, A. Shavick, H. Dong, M. T. C. Poon, N. Fitzpatrick, A. P. Levine, L. T. Slater, A. Handy, A. Karwath, G. V. Gkoutos, C. Chelala, A. D. Shah, R. Stewart, N. Collier, B. Alex, W. Whiteley, C. Sudlow, A. Roberts, and R. J. B. Dobson, "A survey on clinical natural language processing in the United Kingdom from 2007 to 2022," *npj Digital Medicine*, vol. 5, p. 186, Dec. 2022.

[8] N. Schopow, G. Osterhoff, and D. Baur, "Applications of the Natural Language Processing Tool ChatGPT in Clinical Practice: Comparative Study and Augmented Systematic Review," *JMIR Medical Informatics*, vol. 11, p. e48933, Nov. 2023.

[9] M. A. Fink, "From data to insights: how natural language processing and structured reporting advance data-driven radiology," *European Radiology*, vol. 33, pp. 7494–7495, Nov. 2023.

[10] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, pp. 1234–1240, Feb. 2020.

[11] K. Huang, J. Altosaar, and R. Ranganath, "ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission," Nov. 2020.

[12] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing," *ACM Transactions on Computing for Healthcare*, vol. 3, pp. 2:1–2:23, Oct. 2021.

[13] A. Yan, J. McAuley, X. Lu, J. Du, E. Y. Chang, A. Gentili, and C.-N. Hsu, "RadBERT: Adapting Transformer-based Language Models to Radiology," *Radiology: Artificial Intelligence*, vol. 4, p. e210258, June 2022.

[14] A. Turchin, S. Masharsky, and M. Zitnik, "Comparison of BERT implementations for natural language processing of narrative medical documents," *Informatics in Medicine Unlocked*, vol. 36, p. 101139, Jan. 2023.

[15] D. A. Wood, S. Kafiabadi, A. Al Busaidi, E. L. Guilhem, J. Lynch, M. K. Townend, A. Montvila, M. Kiik, J. Siddiqui, N. Gadapa, M. D. Benger, A. Mazumder, G. Barker, S. Ourselin, J. H. Cole, and T. C. Booth, "Deep learning to automate the labelling of head MRI datasets for computer vision applications," *European Radiology*, vol. 32, pp. 725–736, Jan. 2022.

[16] J. J. Visser, M. de Vries, and J. A. Kors, "Automatic detection of actionable findings

and communication mentions in radiology reports using natural language processing," *European Radiology*, vol. 32, pp. 3996–4002, June 2022.

[17] E. Ford, M. Oswald, L. Hassan, K. Bozentko, G. Nenadic, and J. Cassell, "Should free-text data in electronic medical records be shared for research? A citizens' jury study in the UK," *Journal of Medical Ethics*, vol. 46, pp. 367–377, June 2020.

[18] I. Solti, C. R. Cooke, F. Xia, and M. M. Wurfel, "Automated Classification of Radiology Reports for Acute Lung Injury: Comparison of Keyword and Machine Learning Based Natural Language Processing Approaches," *Proceedings. IEEE International Conference on Bioinformatics and Biomedicine*, vol. 2009, pp. 314–319, Nov. 2009.

[19] S. Hassanpour and C. P. Langlotz, "Information extraction from multi-institutional radiology reports," *Artificial Intelligence in Medicine*, vol. 66, pp. 29–39, Jan. 2016.

[20] O. Bodenreider, "The Unified Medical Language System (UMLS): integrating biomedical terminology," *Nucleic Acids Research*, vol. 32, pp. D267–D270, Jan. 2004.

[21] C. P. Langlotz, "RadLex: A New Method for Indexing Online Educational Materials," *RadioGraphics*, vol. 26, pp. 1595–1597, Nov. 2006.

[22] S. Fu, D. Chen, H. He, S. Liu, S. Moon, K. J. Peterson, F. Shen, L. Wang, Y. Wang, A. Wen, Y. Zhao, S. Sohn, and H. Liu, "Clinical concept extraction: A methodology review," *Journal of Biomedical Informatics*, vol. 109, p. 103526, Sept. 2020.

[23] T. Cai, A. A. Giannopoulos, S. Yu, T. Kelil, B. Ripley, K. K. Kumamaru, F. J. Rybicki, and D. Mitsouras, "Natural Language Processing Technologies in Radiology Research and Clinical Applications," *Radiographics*, vol. 36, pp. 176–191, Jan. 2016.

[24] D. D. A. Bui and Q. Zeng-Treitler, "Learning regular expressions for clinical text classification," *Journal of the American Medical Informatics Association*, vol. 21, pp. 850–857, Sept. 2014.

[25] S. Velupillai, H. Suominen, M. Liakata, A. Roberts, A. D. Shah, K. Morley, D. Osborn, J. Hayes, R. Stewart, J. Downs, W. Chapman, and R. Dutta, "Using clinical Natural Language Processing for health outcomes research: Overview and actionable suggestions for future advances," *Journal of Biomedical Informatics*, vol. 88, pp. 11–19, Dec. 2018.

[26] P. H. Dedhia, K. Chen, Y. Song, E. LaRose, J. R. Imbus, P. L. Peissig, E. A. Mendonca, and D. F. Schneider, "Ambiguous and Incomplete: Natural Language Processing Reveals Problematic Reporting Styles in Thyroid Ultrasound Reports," *Methods of Information in Medicine*, vol. 61, pp. 011–018, May 2022.

[27] R. Loor-Torres, M. Duran, D. Toro-Tobon, M. M. Chavez, O. Ponce, C. S. Jacome, D. S. Torres, S. A. Perneth, V. Montori, E. Golembiewski, M. B. Osorio, J. W. Fan, N. S. Ospina, Y. Wu, and J. P. Brito, "A Systematic Review of Natural Language Pro-

cessing Methods and Applications in Thyroidology," *Mayo Clinic proceedings. Digital health*, vol. 2, pp. 270–279, June 2024.

[28] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 507–513, 2010.

[29] K. J. Chen, P. H. Dedhia, J. R. Imbus, and D. F. Schneider, "Thyroid Ultrasound Reports: Will the Thyroid Imaging, Reporting, and Data System Improve Natural Language Processing Capture of Critical Thyroid Nodule Features?," *Journal of Surgical Research*, vol. 256, pp. 557–563, Dec. 2020.

[30] A. Pathak, Z. Yu, D. Paredes, E. P. Monsour, A. O. Rocha, J. P. Brito, N. S. Ospina, and Y. Wu, "Extracting Thyroid Nodules Characteristics from Ultrasound Reports Using Transformer-based Natural Language Processing Methods," *American Medical Informatics Association. Annual Symposium proceedings*, vol. 2023, pp. 1193–1200, 2023.

[31] C. S. Jacome, D. S. Torres, J. W. Fan, R. Loor-Torres, M. Duran, M. A. Zahidy, E. Cabezas, M. Borras-Osorio, D. Toro-Tobon, Y. Wu, Y. Wu, N. S. Ospina, and J. P. Brito, "Thyroid Ultrasound Appropriateness Identification Through Natural Language Processing of Electronic Health Records," *Mayo Clinic Proceedings. Digital Health*, vol. 2, pp. 67–74, Mar. 2024.

[32] BMUS, "Head and Neck Ultrasound Practice Standards," tech. rep., British Medical Ultrasound Society, Sept. 2021.

[33] M. Knappe, M. Louw, and R. T. Gregor, "Ultrasonography-guided fine-needle aspiration for the assessment of cervical metastases," *Archives of Otolaryngology–Head & Neck Surgery*, vol. 126, pp. 1091–1096, Sept. 2000.

[34] V. Chaudhary and S. Bano, "Thyroid ultrasound," *Indian Journal of Endocrinology and Metabolism*, vol. 17, no. 2, pp. 219–227, 2013.

[35] P. Perros, K. Boelaert, S. Colley, C. Evans, R. M. Evans, G. Gerrard BA, J. Gilbert, B. Harrison, S. J. Johnson, T. E. Giles, L. Moss, V. Lewington, K. Newbold, J. Taylor, R. V. Thakker, J. Watkinson, and G. R. Williams, "Guidelines for the management of thyroid cancer," *Clinical Endocrinology*, vol. 81, no. s1, pp. 1–122, 2014.

[36] W. M. Bramer, G. B. de Jonge, M. L. Rethlefsen, F. Mast, and J. Kleijnen, "A systematic approach to searching: an efficient and complete method to develop literature searches," *Journal of the Medical Library Association*, vol. 106, pp. 531–541, Oct. 2018.

[37] OpenAI, "ChatGPT 3.5," 2023.

[38] S. A. Hicks, I. Strümke, V. Thambawita, M. Hammou, M. A. Riegler, P. Halvorsen, and S. Parasa, "On evaluation metrics for medical applications of artificial intelligence," *Scientific Reports*, vol. 12, p. 5979, Apr. 2022.

[39] L. Maier-Hein, A. Reinke, P. Godau, M. D. Tizabi, F. Buettner, E. Christodoulou, B. Glocker, F. Isensee, J. Kleesiek, M. Kozubek, M. Reyes, M. A. Riegler, M. Wiesenfarth, A. E. Kavur, C. H. Sudre, M. Baumgartner, M. Eisenmann, D. Heckmann-Nötzel, T. Rädsch, L. Acion, M. Antonelli, T. Arbel, S. Bakas, A. Benis, M. B. Blaschko, M. J. Cardoso, V. Cheplygina, B. A. Cimini, G. S. Collins, K. Farahani, L. Ferrer, A. Galdran, B. van Ginneken, R. Haase, D. A. Hashimoto, M. M. Hoffman, M. Huisman, P. Jannin, C. E. Kahn, D. Kainmueller, B. Kainz, A. Karargyris, A. Karthikesalingam, F. Kofler, A. Kopp-Schneider, A. Kreshuk, T. Kurc, B. A. Landman, G. Litjens, A. Madani, K. Maier-Hein, A. L. Martel, P. Mattson, E. Meijering, B. Menze, K. G. M. Moons, H. Müller, B. Nichyporuk, F. Nickel, J. Petersen, N. Rajpoot, N. Rieke, J. Saez-Rodriguez, C. I. Sánchez, S. Shetty, M. van Smeden, R. M. Summers, A. A. Taha, A. Tiulpin, S. A. Tsaftaris, B. Van Calster, G. Varoquaux, and P. F. Jäger, "Metrics reloaded: recommendations for image analysis validation," *Nature Methods*, vol. 21, pp. 195–212, Feb. 2024.

[40] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan, "A simple algorithm for identifying negated findings and diseases in discharge summaries," *Journal of Biomedical Informatics*, vol. 34, pp. 301–310, Oct. 2001.

[41] A. N. Tarekegn, M. Giacobini, and K. Michalak, "A review of methods for imbalanced multi-label classification," *Pattern Recognition*, vol. 118, p. 107965, Oct. 2021.

[42] M. S. Sorower, "A Literature Survey on Algorithms for Multi-label Learning," (Oregon State University, Corvallis), 2010.

[43] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Machine Intelligence*, vol. 1, pp. 389–399, Sept. 2019.

[44] L. F. Donnelly, R. Grzeszczuk, C. V. Guimaraes, W. Zhang, and G. S. Bisset III, "Using a Natural Language Processing and Machine Learning Algorithm Program to Analyze Inter-Radiologist Report Style Variation and Compare Variation Between Radiologists When Using Highly Structured Versus More Free Text Reporting," *Current Problems in Diagnostic Radiology*, vol. 48, pp. 524–530, Nov. 2019.

[45] E. Ullah, A. Parwani, M. M. Baig, and R. Singh, "Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology - a recent scoping review," *Diagnostic Pathology*, vol. 19, p. 43, Feb. 2024.

[46] Y. Peng, X. Wang, L. Lu, M. Bagheri, R. Summers, and Z. Lu, "NegBio: a high-performance tool for negation and uncertainty detection in radiology reports," *AMIA*

Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science, vol. 2017, pp. 188–196, 2018.

[47] G. Russ, S. J. Bonnema, M. F. Erdogan, C. Durante, R. Ngu, and L. Leenhardt, "European Thyroid Association Guidelines for Ultrasound Malignancy Risk Stratification of Thyroid Nodules in Adults: The EU-TIRADS," *European Thyroid Journal*, vol. 6, pp. 225–237, Sept. 2017.

[48] F. N. Tessler, W. D. Middleton, E. G. Grant, J. K. Hoang, L. L. Berland, S. A. Teefey, J. J. Cronan, M. D. Beland, T. S. Desser, M. C. Frates, L. W. Hammers, U. M. Hamper, J. E. Langer, C. C. Reading, L. M. Scoutt, and A. T. Stavros, "ACR Thyroid Imaging, Reporting and Data System (TI-RADS): White Paper of the ACR TI-RADS Committee," *Journal of the American College of Radiology*, vol. 14, pp. 587–595, May 2017.

Chapter 4

Ultrasound 3D reconstruction and quality assurance

As discussed in Chapter 1, more precise and streamlined treatment planning is critical for improving High-Intensity Focused Ultrasound (HIFU) delivery. Since preoperative CT/MRI is generally not used to assess thyroid nodules, and 2D ultrasound provides limited options for volumetric planning, 3D ultrasound offers a promising alternative. This allows a detailed understanding of the size and relative positioning of the treatment target and surrounding anatomy, which is crucial for effective HIFU planning.

In the context of the HIFU procedure, where robotics and ultrasound are already used for treatment navigation and real-time guidance, robotic-tracked 3D ultrasound reconstruction offers a practical approach. It does not rely on preoperative imaging and simplifies the procedure by using a single imaging modality. Using the robot arm for tracking also allows a single coordinate system to be used for both planning and treatment navigation, eliminating uncertainty associated with co-registration with external tracking systems.

This chapter describes the in-house system developed for 3D ultrasound based on tracked B-mode reconstruction and a custom phantom designed to evaluate the geometric fidelity of the output volumes. The system was built on open-source frameworks and supports optical tracking, Electromagnetic (EM) tracking, and robotic kinematic tracking, enabling flexible use across a range of diagnostic and interventional applications.

The methods developed in this chapter, including calibration, synchronisation, data coordination, segmentation, and 3D reconstruction, were critical to the thyroid reconstruction demonstrated in Chapter 5. Detailed Quality Assurance (QA) testing showed the system was sufficiently accurate for clinical 3D reconstruction, with further improvements limited by the imaging resolution. It also clarified the roles of EM and optical tracking in the context of robotic US, and demonstrated that the system was robust to changes in speed and insonation angle, informing subsequent work.

A Unified Platform and Quality Assurance Framework for 3D Ultrasound Reconstruction with Robotic, Optical, and Electromagnetic Tracking

Lewis Howell[1,2], Manisha Waterston[1,2], Tze Min Wah[3,4], James H. Chandler[2], James R. McLaughlan[2,4]

[1]School of Computing, University of Leeds, UK
[2]School of Electronic and Electrical Engineering, University of Leeds, UK
[3]Radiology Department, Leeds Teaching Hospital Trust, Leeds General Infirmary, Leeds, UK
[4]Leeds Institute of Medical Research, University of Leeds, UK

Submitted to IEEE Transactions on Medical Robotics and Bionics

---

Abstract

Three-dimensional (3D) Ultrasound (US) can facilitate diagnosis, treatment planning, and image-guided therapy. However, current studies rarely provide a comprehensive evaluation of volumetric accuracy and reproducibility, highlighting the need for robust QA frameworks, particularly for tracked 3D US reconstruction using freehand or robotic acquisition.

This study presents a QA framework for 3D US reconstruction and a flexible open source platform for tracked US research. A custom phantom containing geometric inclusions with varying symmetry properties enables straightforward evaluation of optical, electromagnetic, and robotic kinematic tracking for 3D US at different scanning speeds and insonation angles.

A standardised pipeline performs real-time segmentation and 3D reconstruction of geometric targets (DSC = 0.97, FPS = 46) without GPU acceleration, followed by automated registration and comparison with ground-truth geometries. Applying this framework showed that our robotic 3D US achieves state-of-the-art reconstruction performance (DSC-3D = 0.94 ± 0.01, HD95 = 1.17 ± 0.12), approaching the spatial resolution limit imposed by the transducer.

This work establishes a flexible experimental platform and a reproducible validation methodology for 3D US reconstruction. The proposed framework enables robust cross-platform comparisons and improved reporting practices, supporting the safe and effective clinical translation of 3D ultrasound in diagnostic and image-guided therapy applications.

Keywords: Ultrasound, 3D reconstruction, Quality assurance, Robotic ultrasound, Imag-

ing guidance.

4.1 Introduction

Ultrasound (US) is widely used to image soft tissues, providing a non-ionising, low-cost, portable, and real-time imaging modality [1]. Beyond diagnostic applications, US also supports intra-operative guidance [2], image-guided biopsy [3], microwave and radiofrequency ablation [4, 5], and emerging non-invasive modalities such as HIFU and histotripsy [6–8]. In these applications, spatial understanding of the relative shapes and positions of anatomical structures is critical to ensure diagnostic accuracy, safety, and treatment efficacy. However, traditional Two-dimensional (2D) US provides a limited cross-sectional field of view, which reduces options for volumetric analysis and makes US-guided applications dependent on the operator's scanning technique and skill in geometric reasoning [9, 10].

Tracked Three-dimensional (3D) US addresses this limitation by combining conventional 2D imaging with position and orientation tracking to reconstruct 3D anatomy [11, 12]. Compared with dedicated 3D transducers, this approach enables reconstruction over larger areas, uses widely available and more cost-effective 2D transducers and machines, and can be readily integrated with imaging-guidance systems [11, 13, 14]. To acquire position and orientation tracking data, optical and EM tracking have been used in freehand approaches [12]. For example, Chung *et al.* demonstrated freehand optical-tracked US for carotid artery 3D imaging [14], while Krönke *et al.* applied EM-tracked US for thyroid volumetry, demonstrating improved repeatability relative to standard 2D approaches [15]. More recently, robotic US has improved scanning stability, trajectory control, and pose estimation [16]. Faoro *et al.* used robotic US for 3D vessel reconstruction as a pre-treatment planning step for transcarotid revascularisation [17], Sun *et al.* demonstrated 3D musculoskeletal imaging [18], and Zhou *et al.* proposed a robotic US system for kidney and lesion imaging for HIFU imaging-guidance. Robotic kinematic tracking can be used independently for pose estimation or combined with optical or EM tracking to simplify co-registration with patient or tool coordinate frames, supporting multimodal image guidance and interventional workflows [19].

Despite widespread investigation of tracked US for 3D reconstruction, there is currently no standardised Quality Assurance (QA) framework specifically designed to validate tracked 3D US reconstruction systems. Existing commercial QA phantoms are expensive and intended primarily to assess 2D or dedicated 3D transducers [13]. Consequently, most existing tracked US studies report task-specific outcomes, such as 3D measurement accuracy for a specific organ, but lack a comprehensive evaluation of their systems' capabilities. Task-specific metrics are often also dependent on segmentation of the test object, meaning that errors associated with calibration, tracking, and 3D reconstruction cannot be isolated from those of anatomical segmentation models (often black-box Artificial Intelligence (AI)-based) used to process the US images. The robustness of these systems is

also rarely assessed, particularly how scanning speed and insonation angle affect the quality and reliability of 3D reconstruction. This makes cross-system comparisons difficult, thereby limiting evidence-based development.

To address this, we present two complementary contributions. First, we introduce a flexible, open source tracked US platform for 3D reconstruction that supports interchangeable pose estimation using kinematic, optical, or EM tracking within a unified coordinate system. Second, we present an open source QA framework for evaluating tracked 3D US reconstruction, including a purpose-designed agar phantom with geometric inclusions (sphere, ellipsoid, cylinder, and triangular prism) to enable direct comparison with ground-truth volumes. Using controlled scanning and simultaneous multimodal tracking, we compare kinematic, optical, and EM tracking for 3D reconstruction, providing insights for future research.

4.2 Methods

This study presents a flexible platform for 3D US that integrates conventional brightness-mode (B-mode) US imaging with robotic, optical, or EM tracking. The workflow for 3D reconstruction comprised four stages: (A) tracked image acquisition, (B) calibration and spatial registration, (C) segmentation and 3D reconstruction, (D) quality assurance.

4.2.1 Tracked Image Acquisition

The system uses open source tools to support integration with multiple imaging and tracking systems. Modules for data acquisition, robotic navigation, data streaming, segmentation, and visualisation ran on a distributed system using OpenIGTLink for real-time communication [20] (Fig. 4.1). ROS2 (Foxy) interfaced with the robot and optical tracking systems, PLUS (v2.9.0) handled coordination synchronisation [21], and 3DSlicer (v5.2.2) with SlicerIGT enabled visualisation, reconstruction, and analysis [22]. Segmentation and 3D reconstruction were implemented in Python 3.9 on an Ubuntu 20.04 workstation (Intel i7, 32 GB RAM). Full instructions for the system setup, as well as relevant scripts, configuration files, and example code is included in the GitHub repository (https://github.com/SMRTUS/US-QA-3D).

US images were acquired using a point-of-care US system (GE, Venue Fit) and a L4-20t-RS transducer (3–20 MHz, 38.4 mm Field of View (FOV)) and frames were captured with an HDMI-to-USB 3.0 capture card (Magewell, USB Capture HDMI Gen 2) and processed at the scanner's native resolution (1920x1080) and frame rate (15-30 FPS) using OpenCV. This acquisition approach is broadly compatible with modern US systems providing HDMI output, though device-specific APIs may reduce latency.

High-resolution harmonic B-mode imaging used 10/20 MHz frequency, 78 dB dynamic range, speckle reduction, and virtual convex mode to extend the FOV. Using these imag-

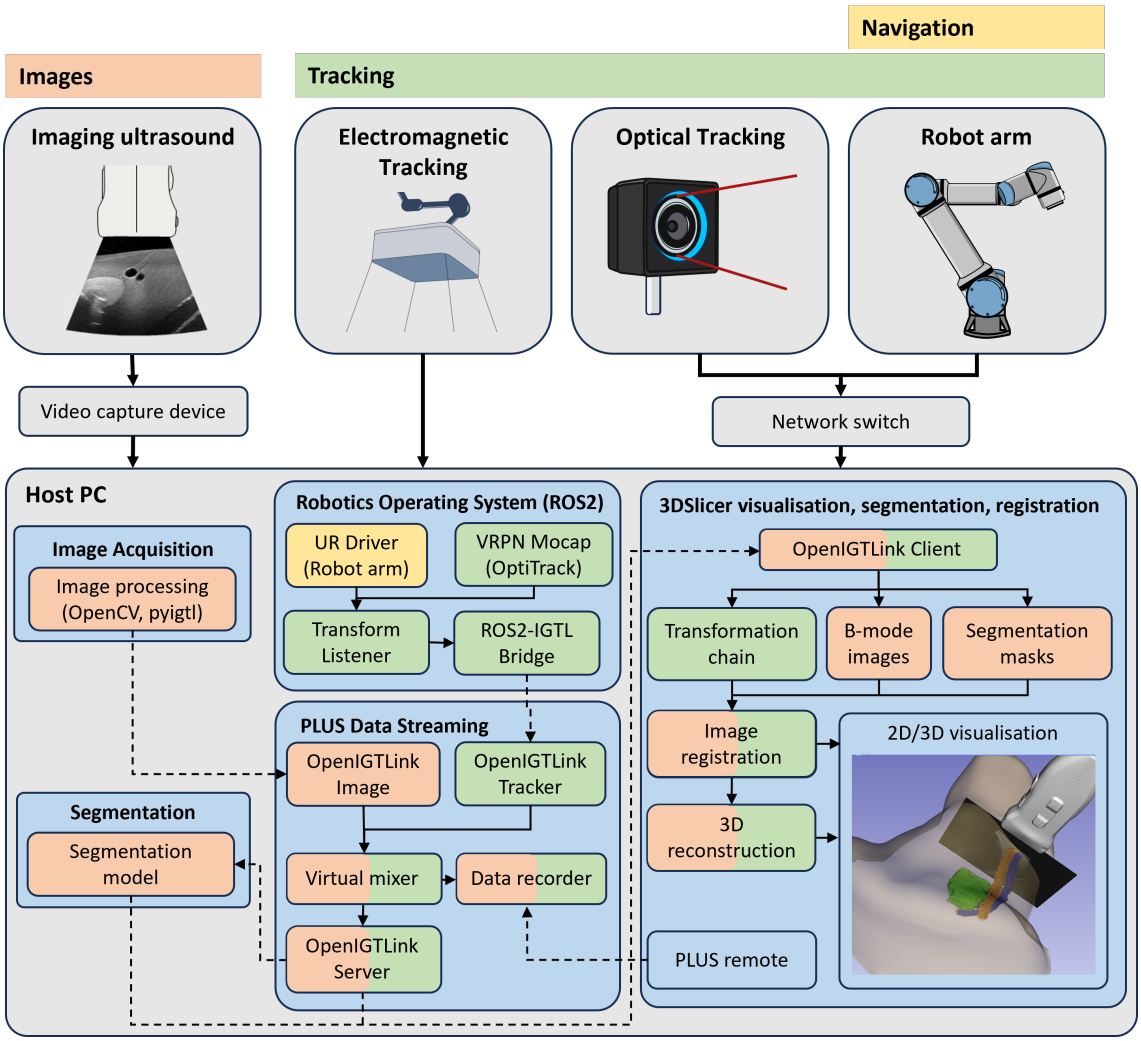


Figure 4.1: Technical diagram showing hardware devices (grey) and software modules (blue) included as part of the experimental system used for real-time 3D US reconstruction. The OpenIGTLink protocol was used to exchange data between modules (dashed arrows). The data flow within modules (thin solid arrows) and connections between hardware devices are also shown (thick solid arrows).

ing parameters, the axial, lateral, and elevational resolution (slice thickness) were 0.25-0.5 mm, 0.5-1.0 mm, and 0.7-2.0 mm respectively at 3 cm focal depth, measured using a multipurpose QA phantom (CIRS, 040GSE) [23].

A 6 Degrees of Freedom (DoF) collaborative robotic arm (Universal Robots, UR3e) was used to move and track the transducer pose. The transducer was secured in a custom two-part 3D-printed holder mounted onto the robot flange. The robot was interfaced via ROS2 and the Universal Robots drivers [24], which provided real-time pose estimates based on the robot's internal kinematics. A custom ROS2 transform listener converted axis-angle-representation poses to transformation matrices and streamed them to PLUS via the ROS2-IGTL-bridge [25]. The module MoveIt for ROS2 was used for motion planning [26] to generate consistent scanning trajectories. Baseline QA scans used linear translation at 5.0 mm/s, unless otherwise stated.

The transducer's position and orientation were also captured using optical and EM tracking. The optical tracking system used four infrared cameras (OptiTrack, Prime 13), positioned to maximise coverage of the working area, and calibrated using a standard wand-based dynamic calibration (OptiTrack Motive v2.2.0) [27]. A rigid-body frame with four 6.4 mm retroreflective markers was mounted to the transducer holder for passive tracking. A ROS2 VRPN client [28] broadcast timestamped poses at 60 FPS, and a custom ROS2 listener forwarded the timestamped transformations to PLUS via ROS2-IGTL-bridge [29].

EM tracking used a Planar 20-20 Field Generator, control unit, sensor interface unit, and 6-DoF sensor (NDI, Aurora), with manufacturer-reported positioning accuracy of 0.48 mm and orientation accuracy of 0.3° [30]. The field generator covered a 500 mm^3 workspace and was positioned to minimise Electromagnetic Interference (EMI). The control unit was connected via USB 3.1, and timestamped poses were acquired directly through PLUS's native Aurora interface (Fig. 4.1) [31].

4.2.2 Calibration and registration

Static transforms defining the system geometry included the flange-to-transducer offset from Computer-Aided Design (CAD) measurements, robot linkage transforms from Denavit-Hartenberg (DH) parameters [32], image-to-transducer transform from spatial calibration, and coordinate registration transforms. These were organised in 3DSlicer using the Medical Reality Modelling Language (MRML) transform hierarchy, enabling consistent propagation of dynamic motion and tracing of US image coordinates relative to the robot base.

Temporal and spatial calibration were performed using the PLUS freehand calibration tool (fCal) [21]. Temporal calibration estimated the latency between the image and tracking streams by imaging the bottom of a water tank while moving the transducer in a periodic oscillatory motion. Spatial calibration determined the static transform between the US image and the transducer (ImageToTransducer) using a 3-layer N-wire phantom [33], giving a sub-millimetre registration error.

All tracking modalities (robotic, optical, and EM) were aligned to a common world coordinate system defined by the robot base frame. Rigid transforms between coordinate-tracking systems were computed via fiducial registration using at least 8 reference points across the workspace. The 3DSlicer volume reslice driver rendered the US image within the 3D scene during scanning.

4.2.3 Segmentation and 3D reconstruction

Real-time 3D reconstruction combined frame-by-frame segmentation with voxel-based volume mapping, providing live visual feedback. Segmentation ran in parallel with the imaging pipeline, so the latency from model inference did not bottleneck visualisation. Tracked frames were processed to produce segmentation masks which retained the original pose and timestamp for synchronised 3D reconstruction.

A threshold-based model segmented anechoic targets in B-mode images of the phantom for QA (D). Processing steps included Region of Interest (ROI) cropping, median filtering (kernel size 7) to reduce speckle noise while preserving edges, and Otsu's thresholding to create binary masks [34, 35]. Masks were refined using morphological closing to fill gaps and morphological opening to remove small islands. A trimming function removed dark vertical regions from poor coupling. The resulting binary masks were then used as the input for 3D reconstruction.

Real-time 3D reconstruction used the PLUS Volume Reconstruction module in 3DSlicer (SlicerIGT) [22], with voxel-based forward mapping to transform each 2D pixel into a 3D voxel coordinate, and nearest neighbour interpolation to assign voxel values within a predefined grid [11]. When multiple pixels mapped to the same voxel, the final value was determined by maximum-value compounding.

Volumes were updated continuously and rendered using 3DSlicer's Visualisation Toolkit (VTK) GPU Ray Casting, enabling real-time 2D and 3D visualisation [36]. Segmentation overlays were displayed on synchronised B-mode images, and multiplanar formatting allowed for axial, sagittal, coronal, or oblique views of the resulting volumes. This allowed for clear visualisation of anatomical details, aligned with patient anatomy or procedural targets, regardless of the original scan orientation [11].

4.2.4 Evaluation and Quality Assurance

To evaluate reconstruction accuracy and repeatability, a purpose-built phantom for Ultrasound Quality Assurance in 3-Dimensions (US-QA-3D) was developed. The phantom contained four inclusions with known geometry: a sphere, an ellipsoid, a cylinder, and an equilateral triangular prism, selected for their varied geometric properties and relevance to anatomical structures. The inclusions were cast from an agar mixture without acoustic scatterers, producing anechoic targets whose dimensions serve as ground truth for 3D reconstruction. Inclusions were embedded in a tissue-mimicking agar block contain-

ing polydisperse glass beads to introduce speckle noise on B-mode, creating clear contrast against the anechoic targets. Phantom preparation details are provided in Appendix B.1.1.

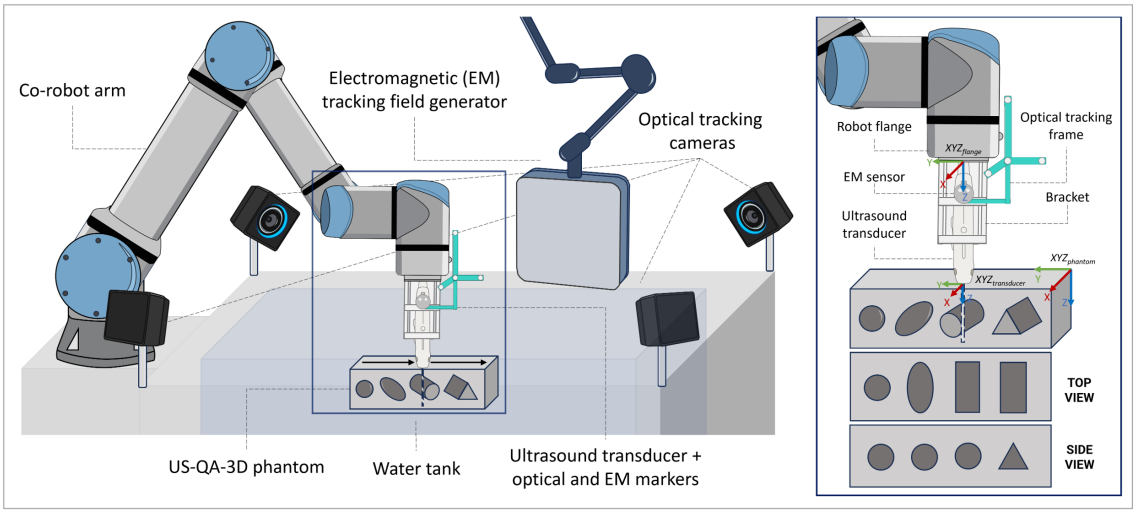


Figure 4.2: Experimental setup for experiments with US-QA-3D phantom, showing the robotic arm, ultrasound transducer, tracking assembly, water tank with geometric phantom, optical cameras, and electromagnetic field generator. The co-robotic arm was used for navigation. An EM sensor and an optical tracking frame enable simultaneous multimodal tracking alongside the robot's kinematics. The reference frames for the robot flange (XYZ_{flange}), transducer ($XYZ_{trasnducer}$), and phantom ($XYZ_{phantom}$) are shown. The US-QA-3D phantom with geometric inclusions is shown in oblique, top, and side views.

QA experiments were conducted with the phantom suspended in a large water tank filled with filtered, degassed water. This ensured consistent acoustic coupling and unrestricted robot motion. The phantom was scanned using robotic navigation with simultaneous EM, optical, and robotic kinematic tracking (Fig. 4.2). Optical cameras were arranged to maximise marker visibility, and the EM field generator was positioned to minimise EMI, without restricting probe motion. A $\sim$1 cm water layer separated the transducer from the phantom to prevent target deformation.

Baseline scans used automated linear trajectories at 5.0 mm/s, with the imaging plane orthogonal to the phantom's long axis. Additional experiments varied scan speed (2.5–17.5 mm/s) and insonation angles (30°, 45°, and 90° about the axial axis and $\pm$ 20° about the lateral axis). At large axial angles, multi-sweep scans were required due to the limited B-mode FOV, and overlapping regions were combined using maximum-voxel compounding. Image and tracking data were recorded through PLUS and triggered via the PLUS remote for 3DSlicer.

Reconstructed volumes were analysed using a Python pipeline with the 3DSlicer Application Programming Interface (API). Intensity thresholding and connected-component analysis separated the four geometric targets, after which geometric descriptors were calculated, including the volume, surface area, centroid location, Feret diameter, roundness, flatness, elongation, and principal axes of variation [35].

Reference models derived from physical measurements of the agar objects were reconstructed in 3DSlicer VTK and co-registered with the corresponding segmented shapes using Iterative Closest Point (ICP) registration. For quantitative evaluation, 3DSlicer's Segment Comparison and Segment Statistics modules were used. Agreement between predicted (X) and reference (Y) geometries was measured using the Dice Similarity Coefficient (DSC):

$$\mathrm{DSC} = \frac{2\,|X \cap Y|}{|X| + |Y|} \tag{4.1}$$

In 2D, DSC measured the pixel-wise agreement between predicted and manually annotated segmentation masks, whereas in 3D (DSC-3D), it quantified voxel-wise agreement between reconstructed and reference volumes.

Boundary accuracy was assessed using the Hausdorff Distance (HD), which measures the difference between the surfaces of the predicted shape (X_s) and reference shape (Y_s). For each surface point ($x \in X_s$) and ($y \in Y_s$), the set of shortest distances to the corresponding shape ($d_{X_s \to Y_s}$ and $d_{Y_s \to X_s}$) was calculated:

$$d_{X_s \to Y_s} = \left\{ \min_{y \in Y_s} \|x - y\| \;:\; x \in X_s \right\} \tag{4.2}$$

$$d_{Y_s \to X_s} = \left\{ \min_{x \in X_s} \|y - x\| \;:\; y \in Y_s \right\} \tag{4.3}$$

This was used to find the (bidirectional) maximum Hausdorff Distance ($\mathrm{HD}_{\max}$), defined by the maximum of both sets of shortest distances:

$$\mathrm{HD}(X_s, Y_s) = \max\left(\max(d_{X_s \to Y_s}),\ \max(d_{Y_s \to X_s})\right) \tag{4.4}$$

and the 95th-percentile HD (HD_{95}):

$$\mathrm{HD}_{95}(X, Y) = \max\left(P_{95}(d_{X \to Y}),\ P_{95}(d_{Y \to X})\right) \tag{4.5}$$

Both the HD and 95th Percentile Hausdorff Distance (HD95) are positive values expressed in millimetres, with 0 mm indicating a perfect prediction.

To complement voxel-based metrics, shape-fitting provided geometry-specific information. Surface meshes were converted to point clouds and fitted using SciPy optimisation to estimate the sphere centre and radius; ellipsoid centre and three semi-axes; cylinder centre, radius, height, and axis; and triangular prism height, axis orientation, and base vertices.

4.3 Results

4.3.1 System Performance

Our system for tracked US reconstruction demonstrated reliable, real-time performance in multimodal tracking QA experiments (Table 4.1). Tight synchronisation and low system latency were essential for accurate real-time 3D reconstruction and visualisation. Frame-by-frame synchronisation was handled using PLUS's internal buffering, with a local time offset uncertainty of < 0.5 ms, as measured by the FCal temporal calibration algorithm. End-to-end latency from US capture to visualisation was approximately 90.2 ± 7.4 ms, consistent with expected delays for image transfer and display using PLUS/OpenIGTLink and 3DSlicer [21]. Volume reconstruction and VTK rendering allowed live updates every 0.2s.

Accurate tracking and registration were also critical to achieve sub-millimetre precision in the output volumes. Image-to-transducer calibration with the N-wire technique gave a Root mean square (RMS) error of 0.81 ± 0.15 mm, consistent with the reported performance of this technique [33]. Robotic kinematic tracking using the robot arm's internal joint transformations yielded highly accurate end-effector tracking, with reported pose positioning errors of < 0.03 mm. Optical tracking calibration achieved a mean reprojection 3D error of < 0.1 mm across all experiments, but tracking was affected by marker visibility, leading to gaps in the reconstructed volumes when the line of sight was obstructed. EM tracking accuracy was limited by field distortion, despite careful placement of the EM field generator to minimise it. The RMS error was 0.1-0.2 mm in controlled conditions but increased depending on the tracker's position relative to the robot, sometimes exceeding several millimetres. Spatial alignment across tracking modalities had a mean RMS registration error of 0.19 ± 0.07 mm.

Table 4.1: System performance, including timing and synchronisation, calibration and registration, and tracking accuracy.

	Metric	Value
System Timing	Synchronisation uncertainty	< 0.5 ms
	End-to-end latency	90.2 ± 7.4 ms
	Volume update rate	0.2 s
Calibration and registration	Image-to-transducer RMS error	0.81 ± 0.15 mm
	Cross-modality RMS registration error	0.19 ± 0.07 mm
Tracking Accuracy	Robot kinematic tracking accuracy	< 0.03 mm
	Optical tracking reprojection error	< 0.1 mm
	EM tracking RMS error	0.1 – 0.2 mm

4.3.2 Geometric phantom

Evaluation using the US-QA-3D phantom enabled controlled assessment of reconstruction accuracy, confirming the system's ability to reconstruct known 3D shapes with low error. Combined, voxel-, surface-, shape-, and fit-based metrics provided a comprehensive eval-

uation of system performance under controlled conditions and highlighted the importance of accurate tracking and consistent scanning for high-quality 3D reconstruction.

Construction of the US-QA-3D phantom was straightforward, and B-mode imaging showed good contrast between the anechoic inclusions and speckle pattern of the outer media (Fig. 4.3a). The segmentation algorithm was tuned to define the ROI, segment the inclusions, and optimise kernels for post-processing. Small bubbles introduced to the agar inclusions during manufacturing did not affect segmentation, as closed-contour filling reliably removed them. Automatic thresholding-based segmentation delineated the inclusions accurately (Fig. 4.3b). Comparison with hand-annotated ground truth images (Fig. 4.3c) showed excellent agreement, with a mean DSC of 0.97 on a test dataset of 50 images. The thresholding approach was also computationally efficient, with a processing time of 21.8 ± 6.4 ms on a mid-range processor.

The segmentation model was applied for real-time 3D-tracked US reconstruction with robotic navigation and simultaneous optical, EM, and kinematic tracking. During scanning, reconstructed volumes were rendered and updated in real time (Supplementary Video), allowing the operator to confirm complete coverage of the ROI (Fig. 4.4a). A second 3D view displayed the robot, transducer, and tracked US frames (Fig. 4.4b). The live US machine feed and B-mode images were shown in a dedicated 2D view (Fig. 4.4c), while a second 2D view displayed the segmentation overlay (Fig. 4.4d).

Baseline reconstruction quality was evaluated using segmentation-, shape-, and fitting-based metrics, averaged across the four geometric targets (Table 4.2). Robotic kinematic and optical tracking performed well, with mean DSC-3D of 0.94 ± 0.01 and 0.93 ± 0.01, respectively, while EM-based reconstruction was poorer, with DSC-3D of 0.86 ± 0.02, likely due to the influence of EMI. This is also supported by the HD and HD95, which measure the surface fidelity of the reconstructed shapes. For robotic kinematic and optical tracking, the maximum HD was < 2 mm and HD95 remained < 1.2 mm, but EM-tracked reconstruction had a HD95 > 2 mm. Fig. 4.5a shows segmented surfaces from the same imaging sequence using each tracking method, coloured by surface error relative to the co-registered ground truth.

Shape-based metrics provided additional validation on the geometric reconstruction accuracy. The roundness of the segmented sphere was 0.98 ± 0.01 for both robotic and optical tracking, close to the expected value of 1.0. The ellipsoid elongation was 2.07 ± 0.01 for robotic tracking and 2.06 ± 0.01 for optical tracking, both close to the expected value of 2.0. EM-tracked reconstructions were visually poorer, with a sphere roundness of 0.95 ± 0.01 and ellipsoid elongation of 1.80 ± 0.02 indicating shape distortion.

Fit-based shape analysis also demonstrated low parameter errors for sphere and ellipsoid models under robotic and optical tracking. Fitted radii deviated by less than 0.5 mm for the sphere Radius R, and for the ellipsoid semi-axis radii R_y and R_z, but were slightly higher for R_x, potentially due to variation in lateral image resolution across the wide virtual

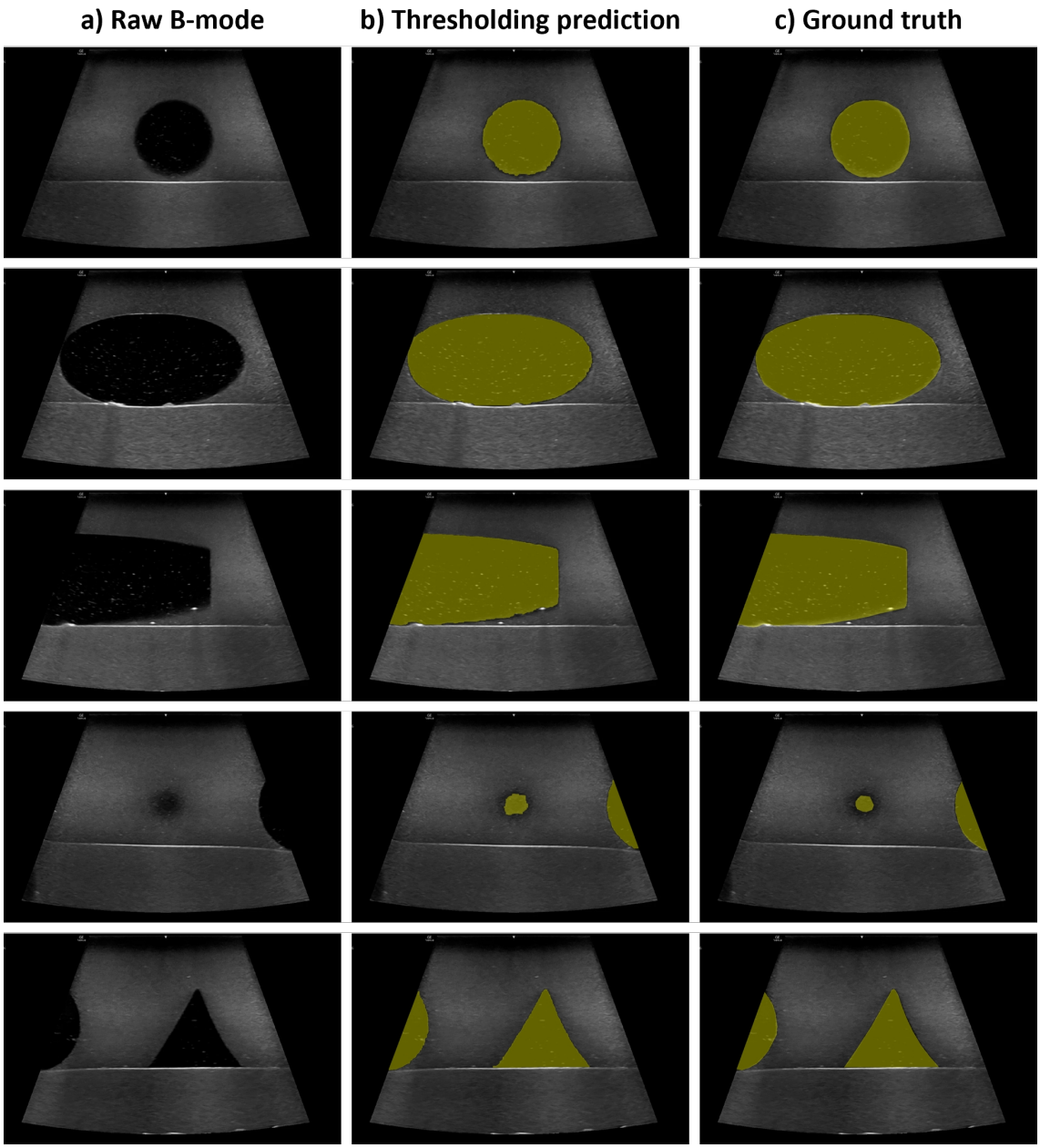


Figure 4.3: Thresholding-based segmentation of objects in the US-QA-3D phantom. Comparison shows a) B-mode images, b) B-mode images with overlaid prediction mask, and c) B-mode images with overlaid ground-truth reference mask.

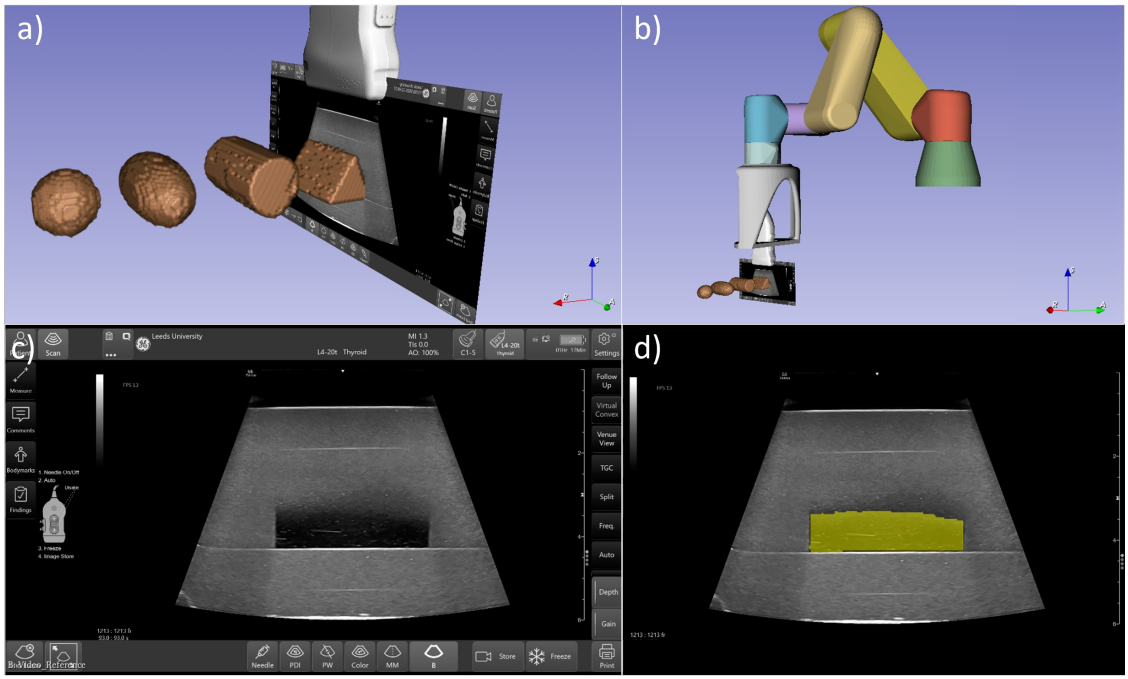


Figure 4.4: Dual 3D view of live real-time volumetric reconstruction of robotic-tracked US in the US-QA-3D phantom. a) Close-up of voxel-based volume reconstruction showing the tracked US frame and reconstructed geometry. b) Robot arm and transducer assembly during the scan. c) B-mode image from the US machine showing a cross-section of the triangular prism inclusion, which appears anechoic compared to the background. d) B-mode image with segmentation overlay from the thresholding algorithm. For a video of the live 3D reconstruction, see Supplementary Video.

Table 4.2: Baseline scans of the US-QA-3D phantom, using simultaneous tracking with robotic kinematic, optical and electromagnetic (EM) tracking. Fitting metrics provide the following shape measurements: sphere radius (R); ellipsoid semi-major axis (R1) and semi-minor axis (R2) radii; cylinder radius (R) and height (H); and triangular prism length (L) and height (H). All metrics are given as mean ± standard deviation. The error relative to the ground truth is shown in brackets for the fitting metrics.

		Robotic	Optical	EM
Segmentation Metrics	DSC-3D	0.94 ± 0.01	0.93 ± 0.01	0.86 ± 0.02
	HD (mm)	1.79 ± 0.37	1.93 ± 0.73	3.87 ± 0.87
	HD95 (mm)	1.17 ± 0.12	1.19 ± 0.12	2.43 ± 0.44
Shape metrics	Volume error (%)	11.80 ± 1.73	11.90 ± 1.47	20.05 ± 4.77
	Surface area error (%)	8.90 ± 2.33	10.41 ± 6.76	18.10 ± 8.54
	Sphere roundness	0.98 ± 0.01	0.98 ± 0.01	0.95 ± 0.01
	Ellipsoid elongation	2.08 ± 0.01	2.07 ± 0.04	1.80 ± 0.02
Fitting metrics	Sphere R (mm)	12.05 ± 0.03 (+0.48)	12.07 ± 0.01 (+0.50)	12.13 ± 0.09 (+0.56)
	Ellipsoid R1 (mm)	12.73 ± 0.09 (+0.40)	12.83 ± 0.37 (+0.50)	14.75 ± 0.11 (+2.42)
	Ellipsoid R2 (mm)	26.56 ± 0.11 (+1.91)	26.59 ± 0.07 (+1.94)	26.74 ± 0.13 (+2.09)
	Cylinder R (mm)	11.23 ± 0.08 (-0.73)	11.14 ± 0.08 (-0.82)	11.93 ± 0.26 (-0.03)
	Cylinder H (mm)	44.94 ± 0.40 (+2.37)	45.14 ± 0.13 (+2.57)	45.00 ± 0.46 (+2.43)
	Tri. prism L (mm)	22.33 ± 0.47 (-1.18)	22.42 ± 0.26 (-1.09)	24.87 ± 0.96 (+1.36)
	Tri. prism H (mm)	38.54 ± 0.07 (+1.56)	38.71 ± 0.35 (+1.73)	40.69 ± 0.85 (+3.71)

convex imaging window. The cylinder and triangular prism also exhibited slightly higher parameter errors for dimensions along the lateral axis (x-axis), with fitted dimensions up to 2.57 mm larger than the measured values (for the cylinder reconstructed with optical tracking). Despite this, all shape-fitting functions converged successfully with low residual error.

The largest errors were observed when segmentation quality was poor, particularly in cases of surface dropout, which was occasionally seen with EM tracking. Depending on the scan orientation, shapes were also slightly elongated in the elevational direction, resulting in errors up to 1 mm in the fitted shape parameters. This was likely due to the transducer's elevation thickness (0.7-2.0 mm), which causes signal spill-out from and out-of-plane elongation due to contributions above or below the imaging plane. In most cases, the shape-fitting parameters slightly exceeded the reference values, indicating a small bias toward overestimating the size of the reference objects.

Comparison with resolution-limited theoretical maximums indicated that reconstruction accuracy was largely constrained by US imaging rather than calibration, tracking, synchronisation, or segmentation. For example, given a reference sphere of radius r, the volume V is

$$V = 4/3 \cdot \pi r^3, \tag{4.6}$$

and a reconstructed sphere with radius error r', the volume V' is

$$V' = 4/3 \cdot \pi (r + r')^3. \tag{4.7}$$

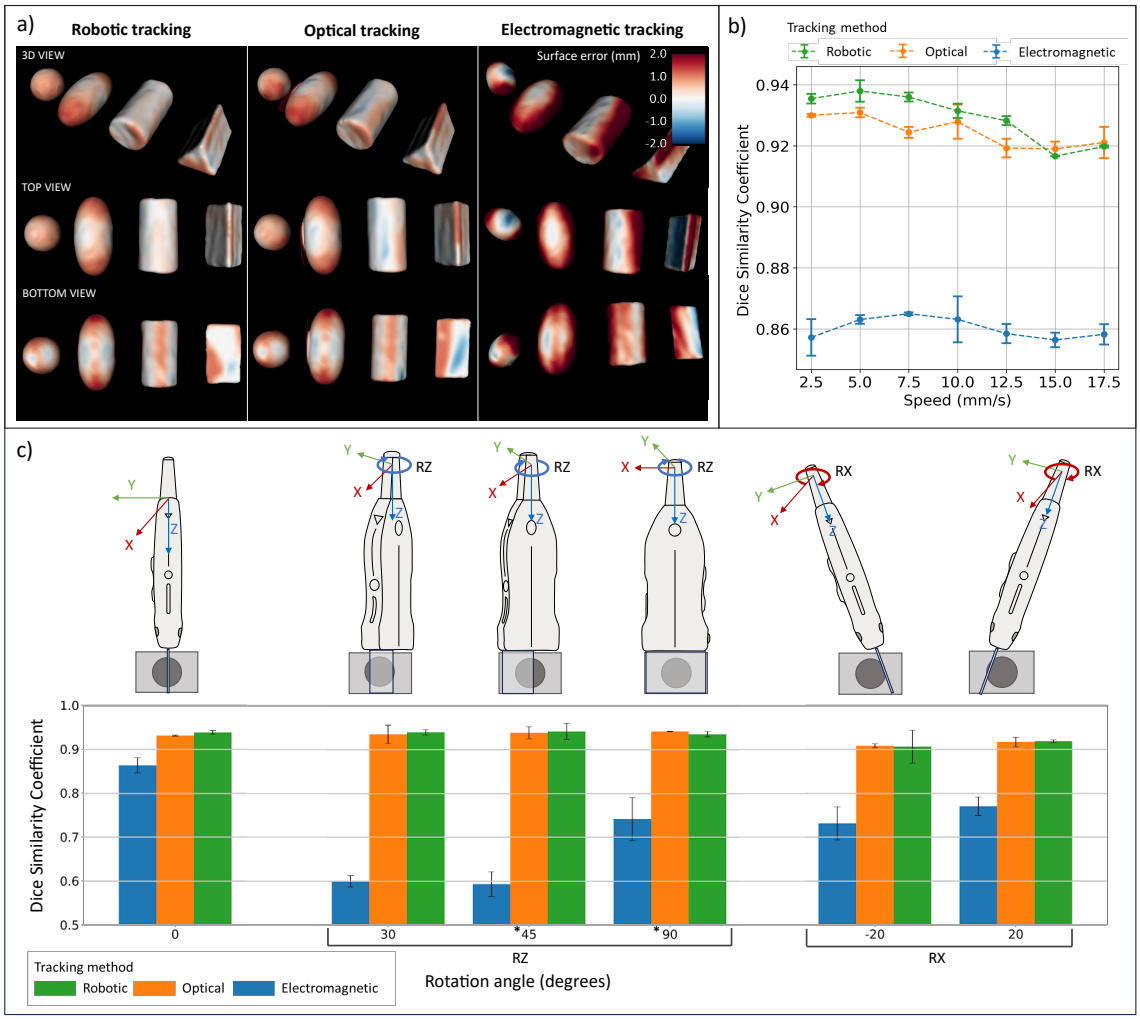


Figure 4.5: 3D reconstruction of objects in US-QA-3D phantom for different tracking modalities (robotic, optical, and electromagnetic). a) Heat maps show the surface deviation error compared with co-registered reference objects. b) Mean DSC across shapes at different scanning speeds. c) Mean DSC at varying insonation angles. Error bars show the standard deviation across three repeats. *These experiments required multi-sweep (compound) scanning.

If the spheres are perfectly aligned and concentric, the theoretical resolution-limited DSC-3D is

$$\frac{2V}{V+V'} = \frac{2r^3}{r^3+(r+r')^3} \tag{4.8}$$

In the US-QA-3D phantom, the sphere had radius $r = 11.5$ mm and a resolution-limited uncertainty $r' = 0.5$ mm, which gives an expected maximum DSC-3D of 0.94, closely matching the experimental values of 0.94 ± 0.01 for robotic, and 0.93 ± 0.01 for optical-tracked 3D reconstruction. Similarly, the theoretical resolution-limited sphere volume error is approximately 12%, close to the experimental values of 11.80 ± 1.73 and 11.90 ± 1.47.

Best results were obtained using slow, constant scanning speeds, with slight degradation in reconstruction quality when the scanning speed exceeded 7.5 mm/s (Fig. 4.5b). At speeds ≤ 7.5 mm/s, robotic tracking performed as well as or slightly better than optical tracking and was much better than EM tracking. Under these conditions, reconstruction was both accurate and repeatable, with a mean DSC-3D > 0.92 and HD95 < 1.2 mm. At higher navigation speeds, reconstruction quality remained acceptable but decreased slightly for both robotic and optical tracking, likely due to minor synchronisation errors producing surface unevenness and small gaps. Overall, robotic tracking produced the most consistent reconstructions, followed closely by optical tracking, while EM tracking performed worse when used alongside the robot arm.

For robotic and optical tracking, insonation angle had little or no effect on the mean DSC-3D, even during multi-sweep scanning (Fig. 4.5c). Axial rotations (z-axis) produced accurate reconstructions despite the non-zero angle between the normal vector of the imaging plane and the scanning direction. For rotations $> 30°$, maximum-value compounding enabled complete reconstruction, with no clear demarcation between the sections from multi-sweep scanning. Rotations about the lateral axis (x-axis) introduced an oblique angle between the transducer face and the phantom surface, increasing artefacts from reflection and refraction and slightly reducing the DSC-3D compared with perpendicular scanning. Across all experiments, the standard deviation between repeats was small, demonstrating good repeatability. In EM scanning experiments, the DSC-3D of the reconstructed shape was more angle-dependent since field distortion effects varied with the relative angle between the sensor and the scanner, as well as the robot position.

4.4 Discussion and Conclusion

This study presents a flexible, open source tracked US platform for 3D reconstruction and a standardised QA framework for quantitative system validation. We demonstrate that accurate volumetric reconstruction can be achieved with a conventional 2D B-mode transducer combined with interchangeable pose tracking using robotic kinematic, optical, or EM tracking. The platform also supports simultaneous multimodal tracking, which enables co-registered applications with other tracked tools or external systems.

Compared with existing phantom-based validation approaches designed for dedicated 3D transducers, the US-QA-3D framework evaluates volumetric shape reconstruction to expose clinically relevant failure modes of tracked US, including shape distortion, boundary noise, and scaling errors [37]. The US-QA-3D phantom incorporates shapes with different symmetry properties, helping to identify errors from calibration, temporal misalignment of data streams, anisotropic distortion in tracking, and reconstruction artefacts that may not be visible in point-based tests. The phantom was inexpensive and simple to construct, and the geometry of the inclusions could be adapted to different applications. Using threshold-based segmentation and automatic registration enabled simple QA without GPUs or specialised hardware, further simplifying implementation. Segmentation of B-mode images achieved high accuracy (DSC = 0.97) and real-time performance (45 FPS), demonstrating its suitability for assessing errors associated with tracking and 3D reconstruction. This effectively decouples tracking and reconstruction errors from segmentation errors, a distinction often overlooked in prior studies. By separating these components, the QA framework enables meaningful comparisons for different systems, transducers, and tracking modalities and provides a practical tool to accelerate future development.

Applying this framework showed that both robotic kinematic tracking and optical tracking can achieve shape reconstruction close to the physics-based resolution limits of the US transducer (DSC-3D = 0.94, HD95 = 1.2 mm). This is significant because optical tracking is highly accurate, with widely reported sub-millimetre RMS positional accuracy [38]. Demonstrating performance comparable to optical tracking confirms that modern 6-DOF robots can provide sufficiently accurate pose tracking for high-quality 3D US, while eliminating the line-of-sight issues and workflow complexity associated with optical cameras. The QA also clarified the role of EM tracking in tracked US, which performed well in freehand calibration, but poorly when used alongside the robot arm. This is consistent with the well-documented susceptibility of EM tracking to magnetic field distortion [39] and highlights the importance of volumetric QA when using EM tracking in clinical settings, where ferromagnetic objects and EMI sources are often unavoidable [39].

Nevertheless, there are important limitations in this study that should form the basis for future research. This study used a single high-frequency linear probe, and further investigation is needed to evaluate performance with other transducers, systems, and imaging settings. Furthermore, while the QA framework assessed geometric reconstruction under controlled conditions, future work should investigate robustness in non-idealised settings to quantify the additional effect of tissue deformation, acoustic shadowing, and patient motion on 3D reconstruction. Finally, although optical, EM, and robotic kinematic tracking have been the dominant techniques for US tracking, recent trackerless methods have also been developed that use AI models to estimate transducer trajectories from video clips [12, 40–42]. These methods, while promising, were not assessed here as they remain difficult to validate quantitatively and are not yet suitable for safety-critical image-guided interventions.

By standardising volumetric QA for tracked US, this work supports transparent performance reporting and facilitates cross-system comparison and evidence-based development of 3D US. The proposed system showed accurate 3D reconstruction with robotic, optical, and EM tracking which makes it extensible to a range of previously investigated clinical applications, including 3D reconstruction for organ volumetry [43], vascular assessment to determine the severity of stenosis [14, 44, 45], and HIFU treatment planning for tumour ablation [46]. Ultimately, the framework will improve confidence in the reliability of 3D US reconstruction, which is essential for safe and effective translation in clinical practice.

4.5 Acknowledgments

This study has received funding from the Engineering and Physical Sciences Research Council (EPSRC) through the following awards: EP/S024336/1 and EP/V04799X/1. J. Chandler was supported by the Leverhulme Trust and the Royal Academy of Engineering under a RAEng/Leverhulme Trust Research Fellowship (LTRF-2425-21-154).

Bibliography

[1] Z. Chenhao, Z. Ruqi, S. Xiaozhuo, W. Jiabo, L. Jiao, and M. Yajuan, "Ultrasound in medicine from 2014 to 2024: A bibliometric review," *Medicine*, vol. 105, p. e46890, Jan. 2026.

[2] N. M. Bekedam, J. N. Smit, P. K. de Koekkoek Doll, M. J. A. van Alphen, R. L. P. van Veen, L. H. E. Karssemakers, M. B. Karakullukcu, and L. E. Smeele, "Intra-operative resection margin model of tongue carcinoma using 3D reconstructed ultrasound," *Advances in Oral and Maxillofacial Surgery*, vol. 4, p. 100154, Oct. 2021.

[3] P. Chatelain, A. Krupa, and N. Navab, "3D ultrasound-guided robotic steering of a flexible needle via visual servoing," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2250–2255, May 2015.

[4] D. Li, Z. Cheng, G. Chen, F. Liu, W. Wu, J. Yu, Y. Gu, F. Liu, C. Ren, and P. Liang, "A multimodality imaging-compatible insertion robot with a respiratory motion calibration module designed for ablation of liver tumors: a preclinical study," *International Journal of Hyperthermia*, vol. 34, pp. 1194–1201, Nov. 2018.

[5] T. Boers, S. J. Braak, W. M. Brink, M. Versluis, and S. Manohar, "3D ultrasound guidance for radiofrequency ablation in an anthropomorphic thyroid nodule phantom," *European Radiology Experimental*, vol. 8, p. 115, Oct. 2024.

[6] L. J. Howell, V. W.-S. Chan, J. Zhong, H. H.-L. Ng, J. H. Chandler, J. R. McLaughlan, and T. M. Wah, "Histotripsy: New Frontiers for Noninvasive Ablation," *The Arab Journal of Interventional Radiology*, vol. 9, p. 61, Dec. 2025.

[7] C. Y. An, J. H. Syu, C. S. Tseng, and C.-J. Chang, "An Ultrasound Imaging-Guided Robotic HIFU Ablation Experimental System and Accuracy Evaluations," *Applied Bionics and Biomechanics*, vol. 2017, p. e5868695, Apr. 2017.

[8] H. H. L. Ng, V. W.-S. Chan, L. Howell, T. Shiwani, J. Zhong, J. Brandon, A. Samson, J. Chandler, J. Mclaughlan, and T. M. Wah, "Early Adoption of Image-Guided Histotripsy Therapy in Interventional Oncology: Challenges and Opportunities in the UK," *British Journal of Radiology*, p. tqag047, Mar. 2026.

[9] Q. Huang and Z. Zeng, "A Review on Real-Time 3D Ultrasound Imaging Technology," *BioMed Research International*, vol. 2017, p. 6027029, 2017.

[10] Y. Wang, Q. Yang, L. Drukker, A. Papageorghiou, Y. Hu, and J. A. Noble, "Task model-specific operator skill assessment in routine fetal ultrasound scanning," *International Journal of Computer Assisted Radiology and Surgery*, vol. 17, no. 8, pp. 1437–1444, 2022.

[11] F. Mohamed, C. V. Siang, F. Mohamed, and C. V. Siang, "A Survey on 3D Ultrasound Reconstruction Techniques," in *Artificial Intelligence - Applications in Medicine and Biology*, IntechOpen, Apr. 2019.

[12] C. Peng, Q. Cai, M. Chen, and X. Jiang, "Recent Advances in Tracking Devices for Biomedical Ultrasound Imaging Applications," *Micromachines*, vol. 13, p. 1855, Oct. 2022.

[13] M. H. Mozaffari and W.-S. Lee, "Freehand 3-D Ultrasound Imaging: A Systematic Review," *Ultrasound in Medicine & Biology*, vol. 43, pp. 2099–2124, Oct. 2017.

[14] S.-W. Chung, C.-C. Shih, and C.-C. Huang, "Freehand three-dimensional ultrasound imaging of carotid artery using motion tracking technology," *Ultrasonics*, vol. 74, pp. 11–20, Feb. 2017.

[15] M. Krönke, C. Eilers, D. Dimova, M. Köhler, G. Buschner, L. Schweiger, L. Konstantinidou, M. Makowski, J. Nagarajah, N. Navab, W. Weber, and T. Wendler, "Tracked 3D ultrasound and deep neural network-based thyroid segmentation reduce interobserver variability in thyroid volumetry," *PLOS ONE*, vol. 17, p. e0268550, July 2022.

[16] A. M. Priester, S. Natarajan, and M. O. Culjat, "Robotic ultrasound systems in medicine," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 60, pp. 507–523, Mar. 2013.

[17] G. Faoro, S. Maglio, S. Pane, V. Iacovacci, and A. Menciassi, "An Artificial Intelligence-Aided Robotic Platform for Ultrasound-Guided Transcarotid Revascularization," *IEEE Robotics and Automation Letters*, vol. 8, pp. 2349–2356, Apr. 2023.

[18] D. Sun, A. Cappellari, B. Lan, M. Abayazid, S. Stramigioli, and K. Niu, "Automatic

Robotic Ultrasound for 3D Musculoskeletal Reconstruction: A Comprehensive Framework," *Technologies*, vol. 13, p. 70, Feb. 2025.

[19] A. Schmidt, O. Mohareri, S. DiMaio, M. C. Yip, and S. E. Salcudean, "Tracking and mapping in medical computer vision: A review," *Medical Image Analysis*, vol. 94, p. 103131, May 2024.

[20] J. Tokuda, G. S. Fischer, X. Papademetris, Z. Yaniv, L. Ibanez, P. Cheng, H. Liu, J. Blevins, J. Arata, A. J. Golby, T. Kapur, S. Pieper, E. C. Burdette, G. Fichtinger, C. M. Tempany, and N. Hata, "OpenIGTLink: an open network protocol for image-guided therapy environment," *The international journal of medical robotics and computer assisted surgery*, vol. 5, pp. 423–434, Dec. 2009.

[21] A. Lasso, T. Heffter, A. Rankin, C. Pinter, T. Ungi, and G. Fichtinger, "PLUS: Open-source toolkit for ultrasound-guided intervention systems," *IEEE Transactions on Biomedical Engineering*, vol. 61, pp. 2527–2537, 2014.

[22] T. Ungi, A. Lasso, and G. Fichtinger, "Open-source platforms for navigated image-guided interventions," *Medical Image Analysis*, vol. 33, pp. 181–186, Oct. 2016.

[23] Computerized Imaging Reference Systems Inc., "Multi-Purpose, Multi-Tissue Ultrasound Phantom," 2013.

[24] Universal Robots, "Universal Robots ROS2 Driver," Oct. 2025.

[25] L. Connolly, A. Deguet, S. Leonard, J. Tokuda, T. Ungi, A. Krieger, P. Kazanzides, P. Mousavi, G. Fichtinger, and R. H. Taylor, "Bridging 3D Slicer and ROS2 for Image-Guided Robotic Interventions," *Sensors (Basel, Switzerland)*, vol. 22, p. 5336, July 2022.

[26] "moveit/moveit2," Oct. 2025.

[27] OptiTrack (NaturalPoint Inc.), "Calibration v2.3 OptiTrack Documentation," Oct. 2022.

[28] A. Sun, "alvinsunyixiao/vrpn_mocap," Sept. 2025.

[29] Tobias Frank and Junichi Tokuda, "openigtlink/ros2_igtl_bridge," Dec. 2024.

[30] Northern Digital Inc, "Minimally Invasive Procedures with Electromagnetic Tracking Technology," 2024.

[31] Perk Lab, "PlusLib: NDI Vega, Polaris and Aurora pose trackers," Aug. 2025.

[32] Universal Robots, "DH Parameters for calculations of kinematics and dynamics," Nov. 2026.

[33] G. Carbajal, A. Lasso, A. Gomez, and G. Fichtinger, "Improving N-wire phantom-based freehand ultrasound calibration," *International journal of computer assisted radiology and surgery*, vol. 8, July 2013.

[34] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, pp. 62–66, Jan. 1979.

[35] J. C. Russ and F. B. Neal, *The Image Processing Handbook*. Boca Raton: CRC Press, 7 ed., Sept. 2018.

[36] M. Bozorgi and F. Lindseth, "GPU-based multi-volume ray casting within VTK for medical applications," *International Journal of Computer Assisted Radiology and Surgery*, vol. 10, pp. 293–300, Mar. 2015.

[37] D. Iommi, J. Hummel, and M. L. Figl, "Evaluation of 3D ultrasound for image guidance," *PLoS ONE*, vol. 15, p. e0229441, Mar. 2020.

[38] A. M. Aurand, J. S. Dufour, and W. S. Marras, "Accuracy map of an optical motion capture system with 42 or 21 cameras in a large measurement volume," *Journal of Biomechanics*, vol. 58, pp. 237–240, June 2017.

[39] F. Poulin and L. P. Amiot, "Interference during the use of an electromagnetic tracking system under OR conditions," *Journal of Biomechanics*, vol. 35, pp. 733–737, June 2002.

[40] Y. Zhou, C. Chen, J. Yao, J. Yu, B. Feng, L. Sui, Y. Yan, X. Chen, Y. Liu, X. Zhang, H. Wang, Q. Pan, W. Zou, Q. Zhang, L. Lin, C. Xu, S. Yuan, Q. He, X. Ding, P. Liang, V. Y. Wang, and D. Xu, "A deep learning based ultrasound diagnostic tool driven by 3D visualization of thyroid nodules," *npj Digital Medicine*, vol. 8, p. 126, Feb. 2025.

[41] H. Yandong, Z. Shiqi, J. Lanting, H. Wenxin, C. Leyao, and H. Hejing, "Establishment and preliminary application of personalized three-dimensional reconstruction of thyroid gland with automatic detection of thyroid nodules based on ultrasound videos," *Journal of Applied Clinical Medical Physics*, vol. 25, p. e14332, Mar. 2024.

[42] W. Wein, M. Lupetti, O. Zettinig, S. Jagoda, M. Salehi, V. Markova, D. Zonoobi, and R. Prevost, "Three-Dimensional Thyroid Assessment from Untracked 2D Ultrasound Clips," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020* (A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, eds.), Lecture Notes in Computer Science, (Cham), pp. 514–523, Springer International Publishing, 2020.

[43] J. Zielke, C. Eilers, B. Busam, W. Weber, N. Navab, and T. Wendler, "RSV: Robotic Sonography for Thyroid Volumetry," *IEEE Robotics and Automation Letters*, vol. 7, pp. 3342–3348, Apr. 2022.

[44] S. Merouche, L. Allard, E. Montagnon, G. Soulez, P. Bigras, and G. Cloutier, "A Robotic Ultrasound Scanner for Automatic Vessel Tracking and Three-Dimensional Reconstruction of B-Mode Images," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 63, pp. 35–46, Jan. 2016.

[45] K. AlMuhanna, M. M. Hossain, L. Zhao, J. Fischell, G. Kowalewski, M. Dux, S. Sikdar, and B. K. Lal, "Carotid plaque morphometric assessment with three-dimensional ultrasound imaging," *Journal of Vascular Surgery*, vol. 61, pp. 690–697, Mar. 2015.

[46] L. Daunizeau, A. Nguyen, M. Le Garrec, J. Y. Chapelon, and W. A. N'Djin, "Robot-assisted ultrasound navigation platform for 3D HIFU treatment planning: Initial evaluation for conformal interstitial ablation," *Computers in Biology and Medicine*, vol. 124, p. 103941, Sept. 2020.

Chapter 5

Co-robotic 3D reconstruction for thyroid ultrasound

3D reconstruction of tracked ultrasound has the potential to improve diagnostic accuracy, spatial understanding for treatment planning, and clinician-patient communication. Combining developments in lightweight multi-organ segmentation from Chapter 2 and robotic-tracked 3D reconstruction techniques from Chapter 4 facilitated live multi-organ reconstruction with high accuracy, demonstrated in the context of thyroid ultrasound. For tracking and navigation, collaborative robotics (co-robotics) was investigated to improve acquisition stability and also keep the clinician in the loop, simplifying the procedure and addressing challenges associated with fully automated scanning.

This work demonstrates real-time 3D US reconstruction, with live volume reconstruction and rendering of the thyroid and vessels. Different scanning techniques were compared, including 'passive' co-robotic backdrive-enabled manual guidance, 'active' co-robotic admittance-control scanning, and freehand scanning with optical and Electromagnetic (EM) tracking. This was used for thyroid volumetry, demonstrating that the 3D US reconstruction is more accurate than the standard 2D sonographic approach. This has ramifications for diagnostic applications in diffuse thyroid disease and for radioiodine therapy planning, where dosimetry is calculated based on the thyroid volume.

In the context of HIFU, this segmentation-based multi-organ 3D US reconstruction could be readily integrated into the pre-treatment planning workflow. Existing systems, such as the Histosonics Edison histotripsy platform, already use a 6-DoF co-robotic arm, meaning co-robotic admittance-control scanning could be feasible with minimal changes to the core hardware. The phantom 3D reconstruction model segments the thyroid and major vessels, which are important for HIFU planning. However, the nodule volume is needed for therapy planning (a feature not present in the phantom), and the trachea, recurrent laryngeal nerve, and oesophagus should be considered in future work aimed at translating the findings using clinical data.

AI-assisted co-robotic ultrasound for three-dimensional thyroid imaging and volumetry

Lewis Howell[1,2], Manisha Waterston[1,2], Amir Zarei[3], Tze Min Wah[3,4], James H. Chandler[2], James R. McLaughlan[2,4]

[1]School of Computing, University of Leeds, UK
[2]School of Electronic and Electrical Engineering, University of Leeds, UK
[3]Radiology Department, Leeds Teaching Hospital Trust, Leeds General Infirmary, Leeds, UK
[4]Leeds Institute of Medical Research, University of Leeds, UK

Submitted to IEEE Transactions on Medical Robotics and Bionics

---

Abstract

Accurate thyroid volumetry is essential for diagnosis, monitoring, and treatment planning. However, conventional approaches rely on two-dimensional (2D) Ultrasound (US), which is operator-dependent and can be inaccurate for patients with enlarged or multinodular thyroids. Three-dimensional (3D) US can overcome these challenges, but has been limited by workflow and segmentation challenges. This study presents an AI-assisted co-robotic US system for real-time 3D reconstruction and automated assessment of the thyroid and adjacent vasculature. The system uses collaborative robotics for hand-guided scanning, with frame-by-frame deep learning segmentation and voxel-based reconstruction. Evaluation with a tissue-mimicking anatomical thyroid phantom showed that the co-robotic acquired 3D US thyroid volume (37.7 ± 1.3 mL) closely matched the computed tomography (CT) ground-truth value (37.9 ± 1.8 mL) and outperformed the standard 2D ellipsoid formula B-mode US approach (30.9 ± 2.0 mL). Co-robotic acquisition improved scanning stability compared with freehand optical and electromagnetic tracked approaches, increasing the accuracy of measurements of thyroid lobe dimensions, vessel diameters, and total thyroid volume. AI-assisted co-robotic 3D US enables accurate, reproducible assessment of the thyroid and vessels while retaining the accessibility and safety of 2D US imaging. This has the potential to improve planning for thyroid procedures, including radioiodine therapy, US-guided biopsy, radiofrequency ablation, and High-Intensity Focused Ultrasound (HIFU).

Keywords: Ultrasound, 3D Reconstruction, Artificial Intelligence, Thyroid, Robotic Ultrasound.

5.1 Introduction

US is the primary imaging modality for clinical evaluation of the thyroid gland. It is non-ionising, portable, and cost-effective, and provides high-resolution imaging of the thyroid and nodules [1, 2]. Beyond diagnosis, US is also used for imaging guidance in biopsies, intra-operative navigation, and emerging non-invasive therapies such as High-Intensity Focused Ultrasound (HIFU) and histotripsy [3–5]. Accurate thyroid volumetry is also important in radioiodine therapy planning for Graves' disease, where dosimetry is calculated based on the thyroid mass, derived from US volumetry [6, 7].

Despite its widespread use, most clinical US applications rely on Two-dimensional (2D) imaging. Thyroid volume is typically determined by measuring the gland in approximately orthogonal planes and modelling each lobe as an ellipsoid (often with an empirical correction factor) [8]. This approach is operator-dependent, and previous studies have shown that the ellipsoid approach is inaccurate for irregular or enlarged thyroid glands. Inter-operator variability in thyroid volume measurement ranges from 11–28% [9], which can influence clinical decision-making, particularly in treatment planning where accurate volumetry is required.

Three-dimensional (3D) US can reduce reliance on geometric assumptions by directly reconstructing anatomy from acquired image data [10]. Dedicated 3D transducers exist but are limited by cost, restricted Field of View (FOV), lower frame rates, and incompatibility with standard clinical systems [11]. An alternative approach is tracked 3D US, in which a conventional 2D transducer is swept across the anatomy, and its pose is recorded to reconstruct a volumetric image [12]. This enables 3D imaging over a large area without the need for specialist US systems [9, 13], but introduces challenges with tracking and computation. Previous systems for neck and thyroid US have included freehand optical-tracked US to image the carotid artery [14] and EM tracked US for thyroid volumetry [9]. However, optical tracking requires line-of-sight between infrared cameras and retroreflective markers on the transducer [15], and EM tracking is susceptible to Electromagnetic Interference (EMI) [16, 17]. Freehand scanning also depends on operator skill and hand stability, limiting its reproducibility [13].

Robotic ultrasound (RUS) systems have emerged as a promising solution to improve the reproducibility of tracked US acquisition [18–20]. RUS enables teleoperation and automated scanning [21–23]; however, a large skill gap remains between human experts and autonomous US scanning systems [24]. Collaborative robotic (co-robotic) approaches offer a middle ground between freehand and automated RUS, allowing the operator to control the scanning while still benefiting from robotic stability, speed regulation, and kinematic tracking [21, 25, 26].

Neck and thyroid imaging are well-suited to RUS because the anatomy is superficial, respiratory motion is negligible, target structures are small and well-defined, and diagnostic

workflows already rely on US assessment [9, 10]. Several recent studies have investigated RUS for either thyroid [7, 23, 27–31] or carotid artery scanning [14, 32–37]; however, none have yet demonstrated a unified method for diagnostics and treatment planning in the neck. Previous work by Huang *et al.* demonstrated RUS 3D reconstruction in a thyroid phantom using B-mode intensity mapping [27, 28], and Kaminski *et al.* applied force feedback to improve 3D reconstruction in the same phantom [30]. Kojcev *et al.* later evaluated RUS and conventional US thyroid volumetry in four healthy volunteers, suggesting that their robotic approach was more repeatable than conventional sonography [31]. However, these studies lacked the organ-specific segmentation required for automated volumetry and treatment planning.

More recently, Artificial Intelligence (AI) and Deep Learning (DL) methods have been applied to real-time segmentation and adaptive robotic control [29, 33, 38, 39]. Zielke *et al.* introduced an AI-guided robotic system for thyroid scanning using live segmentation for motion control [7], recording a mean thyroid volume of 32.48 ± 0.93 mL, compared to 23.75 ± 4.83 mL for experts, and 30.01 mL for the Computed Tomography (CT) baseline. Su *et al.* combined automated RUS with a DL classifier for thyroid nodule classification in nineteen patients [29], but without 3D reconstruction. Lin *et al.* demonstrated fully automated robotic scanning of a thyroid phantom with accurate volumetry (34.3 ± 0.3 mL vs. 34.5 mL on CT) and good force reproducibility (2.1 ± 0.5 N) [23], but did not evaluate the effect of scanning speed on 3D reconstruction, or include 3D renderings of the reconstructed anatomy. Jiang *et al.* applied RUS and DL for carotid artery imaging in 122 volunteers with expert-level biometric accuracy (± 0.42 mm), but did not extend to 3D reconstruction [35].

Therefore, important clinical gaps remain. Many RUS thyroid studies focus on robotic control strategies for automation, but few investigate co-robotic strategies. Others report volumetry based on post-processing, without 3D reconstruction, or without quantitative comparison to CT reference imaging. Real-time multi-organ 3D reconstruction of the thyroid and surrounding anatomy has not yet been demonstrated, despite its relevance to biopsy planning, ablation, and non-invasive therapies, where understanding the spatial relationships between structures is important for safety and efficacy.

In this study, we present an AI-assisted co-robotic US platform for real-time 3D reconstruction of the thyroid and adjacent vasculature using a conventional clinical US system (Fig. 5.1). A collaborative robotic approach maintains direct clinician control while improving acquisition stability and tracking accuracy. DL–based segmentation enables live reconstruction and automated extraction of thyroid volume, lobe dimensions, and vessel diameters. We demonstrate that co-robotic 3D US improves volumetric accuracy compared to conventional 2D ellipsoid-based estimation, supporting its potential role in thyroid diagnostic applications, disease monitoring, and interventional planning.

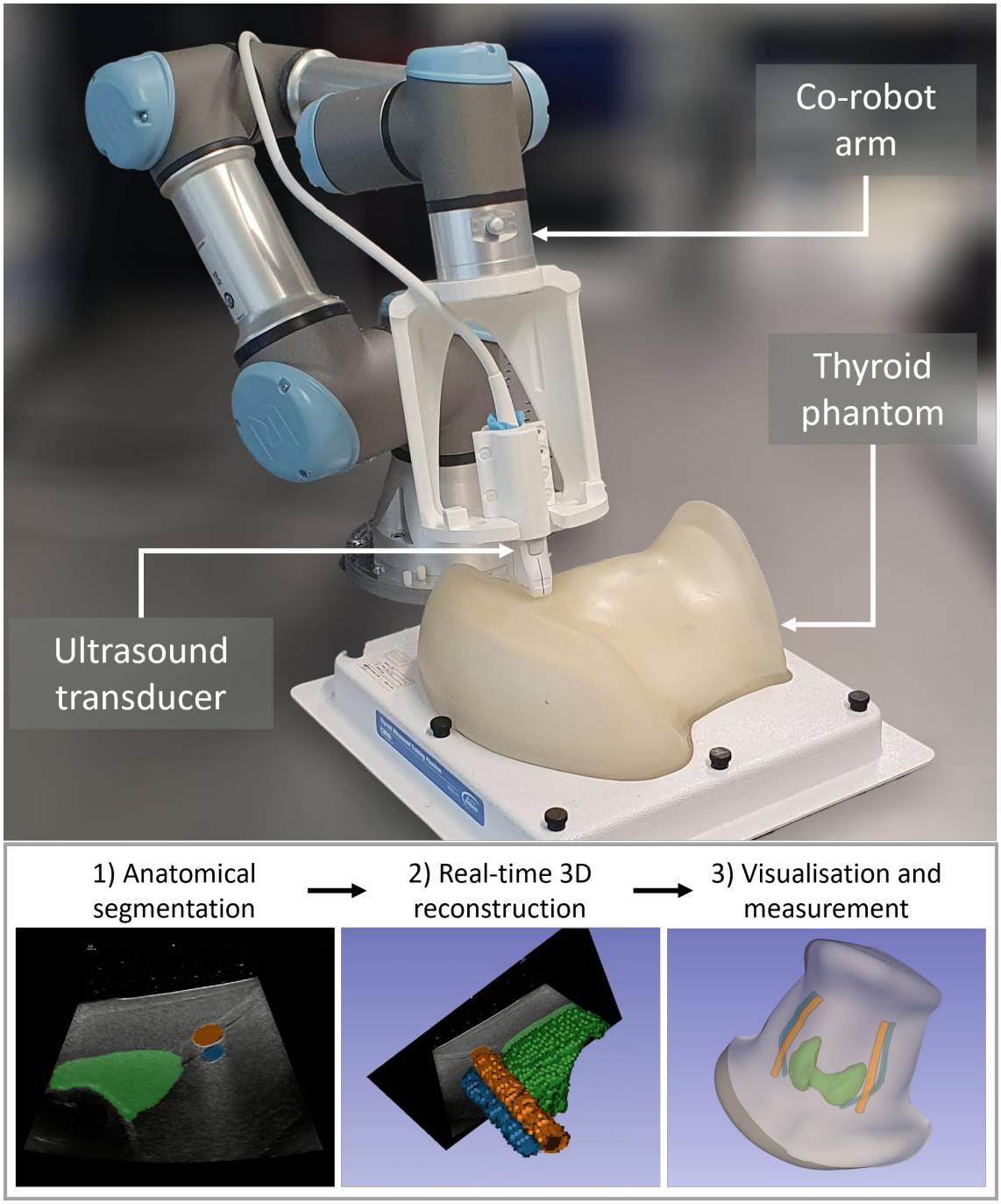


Figure 5.1: System for co-robotic 3D reconstruction in a thyroid anatomical phantom (CIRS, Model 074) using a linear US transducer (GE Healthcare, L4-20T-RS) mounted on a 6-DoF collaborative robot arm (Universal Robots, UR3e). A deep learning model segments the thyroid and vessels in each frame, and a real-time 3D reconstruction algorithm spatially compounds the segmented images from tracked transducer poses, enabling visualisation and automatic volumetric assessment of the thyroid and major vessels.

5.2 Methods

This study evaluated an AI-assisted co-robotic US system [40] for 3D reconstruction and volumetric assessment of the thyroid gland and adjacent vasculature (Fig. 5.1). Co-robotic scanning was compared with freehand scanning using optical and EM tracking. Experiments were conducted using a thyroid training phantom, and the US-derived thyroid volume was compared with the CT gold-standard reference volume.

5.2.1 System overview

Experiments used a commercially available tissue-mimicking thyroid phantom (Model 074, CIRS), which includes a slightly enlarged thyroid gland, trachea, common carotid arteries, and internal jugular veins. This phantom has been used as a baseline in previous RUS studies [7, 23, 27, 28, 30].

A high-frequency linear transducer (GE L4-20t-RS) and clinical US system (GE Venue Fit) captured Brightness Mode (B-mode) images for 3D reconstruction. Imaging parameters were typical of those used for thyroid US [41]: harmonic imaging at 10/20 MHz, 78 dB dynamic range, speckle reduction, and virtual convex mode to extend the FOV. Gain and Time Gain Compensation (TGC) were adjusted for optimal contrast. US spatial and temporal calibration used PLUS fCal with a 3-layer N-wire phantom [42, 43].

In co-robotic experiments, the transducer was mounted on a 6-DoF collaborative robot (UR3e, Universal Robots). Two control modes were compared: 'passive' backdrive-enabled manual guidance (Universal Robots Freedrive) and 'active' admittance control with force sensing (Robotiq Force Copilot, ActiveDrive) to assist movement, regulate speed, and maintain contact force. Freehand experiments used optical tracking with four cameras (OptiTrack, Prime 13) and EM tracking with a 6-DoF sensor (NDI, Aurora) to capture the transducer position and orientation.

All scans were performed by a single operator following standard clinical thyroid US technique [2, 41, 44]. Three sweeps were acquired: two transverse superior–inferior sweeps of the left and right thyroid lobes and vessels, and a midline sweep of the isthmus and trachea. Transducer motion was slow (2.8-3.5 mm/s), imaging perpendicular to the phantom surface with light, consistent pressure to ensure consistent coupling and minimise deformation. Imaging was acquired with freehand (optical and EM) and co-robotic (active and passive) navigation, each with three repeats.

5.2.2 Segmentation and 3D reconstruction

Real-time reconstruction used frame-by-frame segmentation and voxel-based mapping to provide immediate visual feedback during scanning. A lightweight U-Net model was trained for multiclass anatomical segmentation of the thyroid, carotid artery, and jugular vein in phantom B-mode images (Fig. 5.2). This used an encoder-decoder architecture to extract

hierarchical features from the B-mode image, which were then progressively upsampled and fused with skip connections to produce pixel-wise class prediction masks. Details of the architecture are given in Table 5.1.

The model was trained on 539 hand-annotated images, with a 20% validation split to regulate a dynamic training regime using learning-rate decay and early stopping to prevent overfitting. Training used the Adam optimiser [45] to minimise a Dice-cross-entropy combo loss (weighted 2:1) [46], and US-specific data augmentations (emulating varying gain, depth, and TGC) to improve generalisation [47]. Performance was evaluated on 100 test images, sampled from separate cine-loops to minimise the risk of data leakage. In 3D reconstruction, the trained model was used in inference mode and ran in parallel with the imaging pipeline to minimise latency. The resulting segmented frames were used to inform 3D reconstruction.

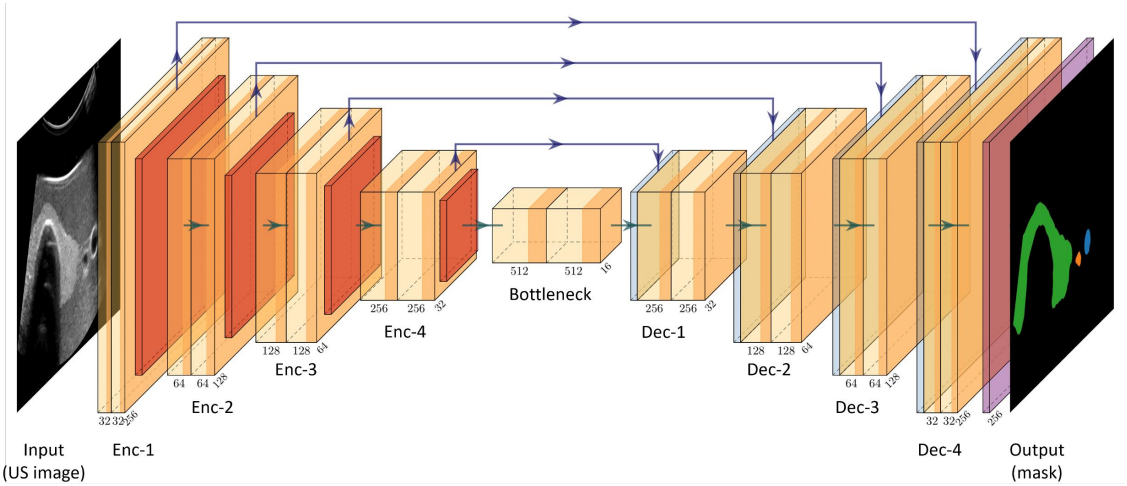


Figure 5.2: Lightweight U-Net architecture for multiclass anatomical segmentation in the thyroid phantom. The input image is mapped to the output segmentation map via a convolutional encoder-decoder network. Encoder (Enc) blocks extract hierarchical features from the input image and compress spatial information. Decoder (Dec) reconstruct the segmentation map and recovers spatial details. Skip connections are shown as blue arrows, and the direct flow of information between layers as green arrows.

Segmented US frames inherited the original pose and timestamp for 3D reconstruction, ensuring tight synchronisation. Voxel-based forward mapping transformed each pixel to 3D space, using nearest-neighbour interpolation and maximum-value compounding for real-time 3D reconstruction [11]. Volumes were updated and rendered using 3DSlicer's Visualisation Toolkit (VTK) GPU Ray Casting [48]. Post-processing removed small islands (< 10 mm^3), followed by 3 mm (3x3) morphological opening and closing operations to reduce small extrusions and fill small cavities. A 3 mm median filter smoothed the reconstructed surface while preserving the overall morphology. Finally, multiplanar formatting re-sliced the volumes to enable interactive inspection in axial, sagittal, coronal, or oblique planes, regardless of the original scan orientation.

Table 5.1: CNN structure for real-time multiclass thyroid segmentation. The CNN uses encoder (Enc) blocks, with an increasing number of convolutional filters per layer (32, 64, 128, 256). Each Enc block consists of 2D convolutional layers with 3x3 kernels (Conv3x3), Batch Normalisation (BN), LeakyReLU activation (α=0.1), and Dropout (rate = 0.2), followed by 2x2 Max-pooling. The bottleneck layer has 512 filters. Decoder (Dec) blocks mirror the encoder, using bilinear upsampling, feature concatenation (Concat) with the corresponding Enc block output, and convolutional blocks with a decreasing number of filters (256, 128, 64, 32).

Stage	**Resolution (H×W)**	**Channels**	**Details**
Input	256×256	1	Input (B-mode ultrasound image)
Enc-1	128×128	32	2×(Conv3×3 + BN + LeakyReLU + Dropout), MaxPool 2×2, stride 2
Enc-2	64×64	64	2×(Conv3×3 + BN + LeakyReLU + Dropout), MaxPool 2×2, stride 2
Enc-3	32×32	128	2×(Conv3×3 + BN + LeakyReLU + Dropout), MaxPool 2×2, stride 2
Enc-4	16×16	256	2×(Conv3×3 + BN + LeakyReLU + Dropout), MaxPool 2×2, stride 2
Bottleneck	16×16	512	2×(Conv3×3 + BN + LeakyReLU + Dropout)
Dec-4	32×32	256	UpSample ×2, Concat Enc-4, 2×(Conv3×3 + BN + LeakyReLU + Dropout)
Dec-3	64×64	128	UpSample ×2, Concat Enc-3, 2×(Conv3×3 + BN + LeakyReLU + Dropout)
Dec-2	128×128	64	UpSample ×2, Concat Enc-2, 2×(Conv3×3 + BN + LeakyReLU + Dropout)
Dec-1	256×256	32	UpSample ×2, Concat Enc-1, 2×(Conv3×3 + BN + LeakyReLU + Dropout)
Output	256×256	4	1×1 Conv + Softmax (pixel-wise class probabilities)

The thyroid left and right lobes and isthmus were then segmented along longitudinal planes lateral to the trachea walls, and the maximum length, width, and anterior-posterior diameter of each lobe were extracted. Vessels were separated laterally, and centrelines were extracted to find the average diameter with SlicerVMTK. The 3D thyroid volume ($V_{\text{3D-TV}}$) was calculated directly from the voxel count and grid spacing.

5.2.3 Thyroid volumetry and quantitative analysis

In line with previous studies, manual segmentation of a CT scan was used to determine the gold-standard reference thyroid volume [7, 23], allowing direct comparison with the $V_{\text{3D-TV}}$ from 3D US reconstruction. The shape of the co-registered thyroid model from 3D US was also compared with the CT reference using the Dice Similarity Coefficient (DSC), which quantified the voxel-wise volumetric agreement (Equation 5.1). This was also applied in 2D to find the pixel-wise agreement between AI-segmented and manually-annotated test data.

$$\text{DSC} = \frac{2\,|X \cap Y|}{|X| + |Y|} \tag{5.1}$$

The 95th Percentile Hausdorff Distance (HD95) was also calculated (Equation 5.2) to quantify the surface distance error between the thyroid models from CT and 3D US, measured as the 95th percentile of closest-point distances in both directions.

$$\text{HD}_{95}(X, Y) = \max\left(P_{95}(d_{X \to Y}),\ P_{95}(d_{Y \to X})\right) \tag{5.2}$$

Additionally, reference thyroid and vessel measurements were obtained using 2D B-mode, which is the clinical standard measurement technique [1, 2]. Thyroid lobe length (L) and anterior-posterior diameter (AP) were measured in sagittal views, and width (W) was measured in transverse views [1]. The thyroid volume was calculated using the clinical standard ellipsoid formula ($V_{\text{US-TV}}$, Equation 5.3) [8] and a lower correction factor formulation ($V_{\text{US-LCF}}$, Equation 5.4) [8, 49, 50].

$$V_{\text{US-TV}} = L \times W \times AP \times \pi/6 \quad (\approx 0.524) \tag{5.3}$$

$$V_{\text{US-LCF}} = L \times W \times AP \times 0.479 \tag{5.4}$$

5.2.4 Measuring movement smoothness

To assess the smoothness of freehand and co-robotic US scanning, tracking data were acquired for scans of the left lobe of a thyroid phantom with a single slow sweep (2.8-3.5 mm/s). For each experiment, the position, orientation, and force-torque values were recorded in ROS2 at 30 Hz. Noise in the tracking data was suppressed using a Savitzky-Golay filter [51] (order 3, window size 11) and the Log Dimensionless Jerk (LDLJ) was

calculated across four non-overlapping 5 s windows, quantifying the smoothness of movement by measuring deviations from an idealised perfectly smooth motion. The LDLJ (Equation 5.5) is a normalised metric derived from the second derivative of speed with respect to time, where lower (less negative) values indicate a smoother scan:

$$\mathrm{LDLJ} = -\ln\left((t_2 - t_1)^3\, v_{\mathrm{peak}}^2 \int_{t_1}^{t_2} \left(\frac{d^2 v(t)}{dt^2}\right)^2 dt\right) \tag{5.5}$$

where t_1 and t_2 are the start and stop times of the movement, v_{peak} is the maximum speed, and $\frac{d^2v(t)}{dt^2}$ is the jerk. We also report the Spectral Arc Length (SPARC) (Equation 5.6), as proposed by Balasubramanian *et al.*, which is less sensitive to noise than the LDLJ [52]. SPARC is given by

$$\mathrm{SPARC} = -\int_{\omega_l}^{\omega_h} \sqrt{\left(\frac{1}{\omega_c}\right)^2 + \left(\frac{d\hat{V}(\omega)}{d\omega}\right)^2}\, d\omega \tag{5.6}$$

Where $\hat{V}(\omega)$ is the normalised Fourier magnitude spectrum of the velocity and ω_c is adaptively selected based on the normalised Fourier magnitude spectrum to avoid noise.

5.3 Results

The system's baseline 3D reconstruction performance was first evaluated using a quality assurance phantom with anechoic targets of known geometry [40]. The base volume reconstruction accuracy was measured as DSC = 0.94 ± 0.01, and the surface reconstruction error was HD95 = 1.17 ± 0.12 mm, indicating that tracking and 3D reconstruction errors were small relative to the transducer's spatial resolution (∼0.5-1 mm). Robotic kinematic tracking had positioning errors of < 0.03 mm [53], optical tracking had a mean 3D reprojection error of < 0.1 mm, and EM tracking had an Root mean square (RMS) error of around 0.1-0.2 mm. Spatial calibration of the B-mode images had an RMS error of 0.81 ± 0.15 mm.

The system was then applied for thyroid 3D US. Training the lightweight segmentation U-Net took 9 min 43 s (76 epochs) and achieved DSC values of 0.88, 0.90, and 0.74 for the thyroid, carotid artery, and jugular vein, respectively, as assessed on the independent test set. Most images were segmented accurately (Fig. 5.3), but the jugular vein was occasionally misclassified as the carotid artery, which has a similar morphology. Some false-positive regions were present for the thyroid class, mostly in the far-field region around the trachea, where the thyroid boundary was unclear.

The model inference time was 31.3 ± 3.8 ms with 3.2 ± 0.9 ms for pre-processing, allowing segmentation at 27 FPS. This was sufficient for live 3D reconstruction in thyroid scanning experiments (Fig. 5.4), enabling the user to monitor the segmentation overlay and generate 3D anatomical structures during the scan (Supplementary Video 1). Time offset

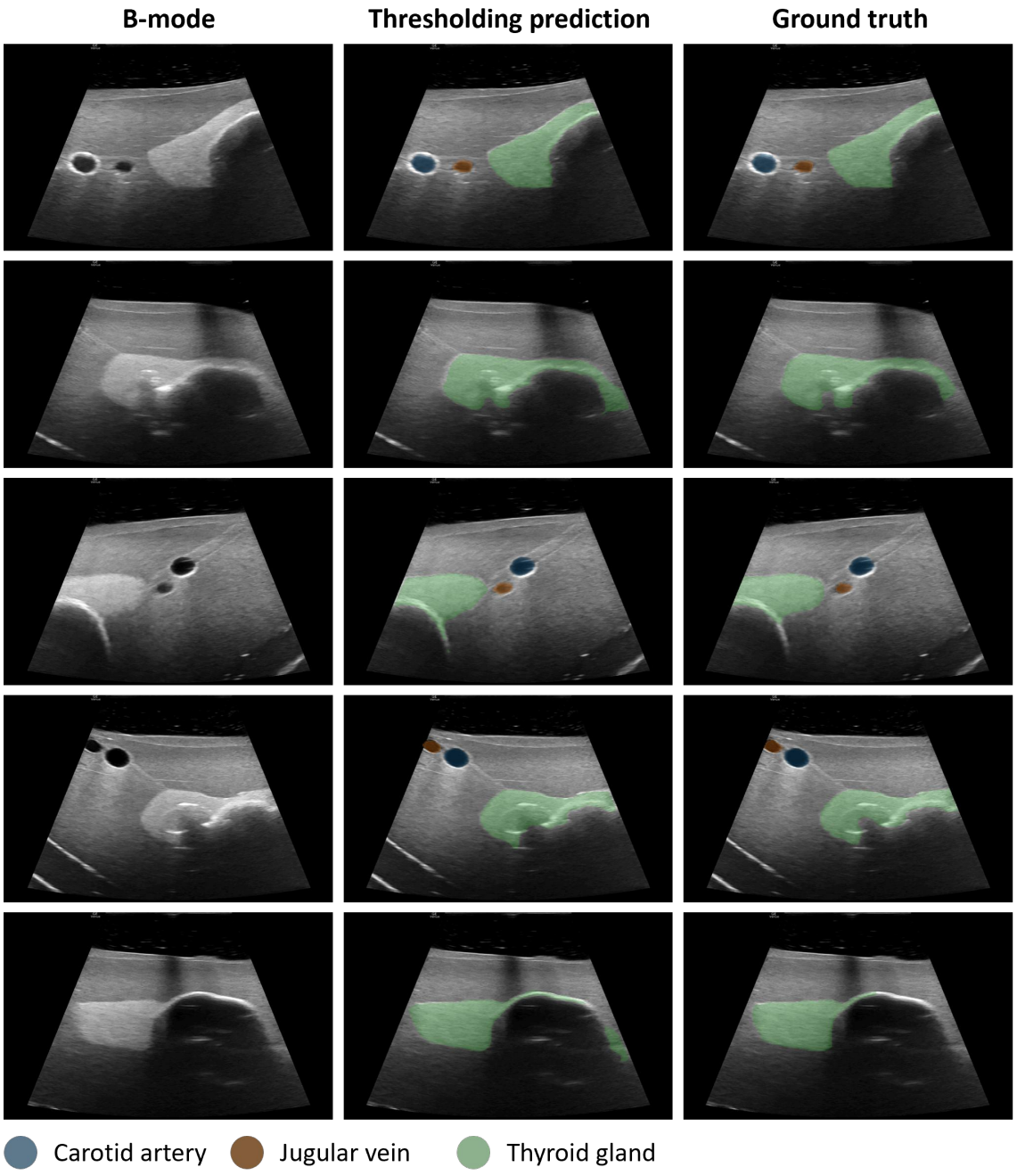


Figure 5.3: AI-based segmentation of anatomy in the thyroid phantom. The comparison shows the B-mode US image, overlaid prediction mask, and overlaid ground-truth reference mask.

uncertainty between imaging and tracking was < 0.5 ms, and end-to-end latency from US capture to visualisation was < 100 ms. Live volume reconstruction and rendering allowed volumes to be updated every 0.5 s using GPU acceleration.

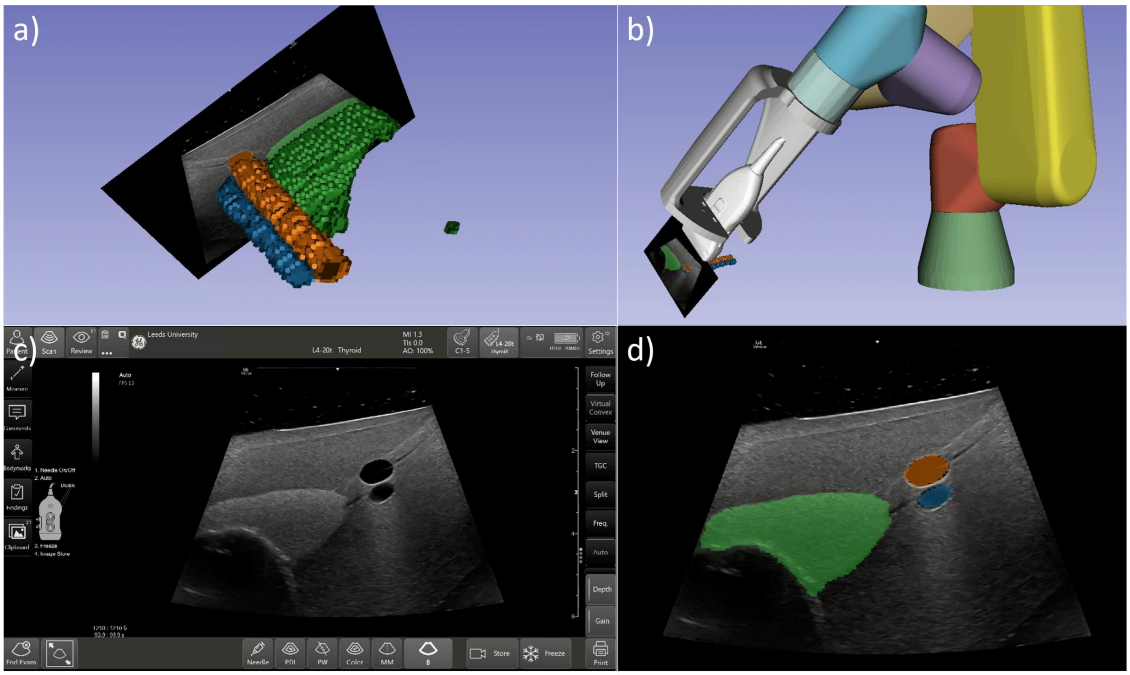


Figure 5.4: Dual 3D view of real-time volumetric reconstruction of co-robotic US in a thyroid training phantom: a) close-up of voxel-based volume reconstruction showing the tracked US frame and reconstructed carotid artery (orange), jugular vein (blue), and thyroid (green), b) co-robot arm and transducer assembly. c) B-mode image from the US machine, d) B-mode image with segmentation overlay. For a video of the live 3D reconstruction, see Supplementary Video 1.

The three-sweep co-robotic scanning approach captured the complete thyroid volume (Fig. 5.5a). Real-time 3D reconstruction of the thyroid and major vessels produced volumes with some surface noise (Fig. 5.5b), which was smoothed and suppressed by post-processing (Fig. 5.5c). The thyroid was split into its left and right lobes and isthmus, and thyroid and vessel measurements were extracted (Fig. 5.5d). All scans captured the general structure of the thyroid lobes and adjacent vessels, but the level of detail and fidelity varied with the tracking method and acquisition approach. Consistent segmentation of anatomical structures across repeats supported the repeatability of the AI-based approach.

After scanning and 3D reconstruction, the volumes were automatically post-processed and re-sliced using Multiplanar Reconstruction (MPR) to generate axial, sagittal, and coronal (or oblique) views of the anatomy, similar to CT or Magnetic Resonance Imaging (MRI) (Fig. 5.6). The rendered volumes could also be manipulated (rotated, zoomed, and panned) to inspect the anatomy from all angles in 3D.

Table 5.2 summarises the 3D reconstruction results. Average diameters ($\varnothing$) for the left and right carotid artery and jugular vein were within the range of uncertainty of the measured

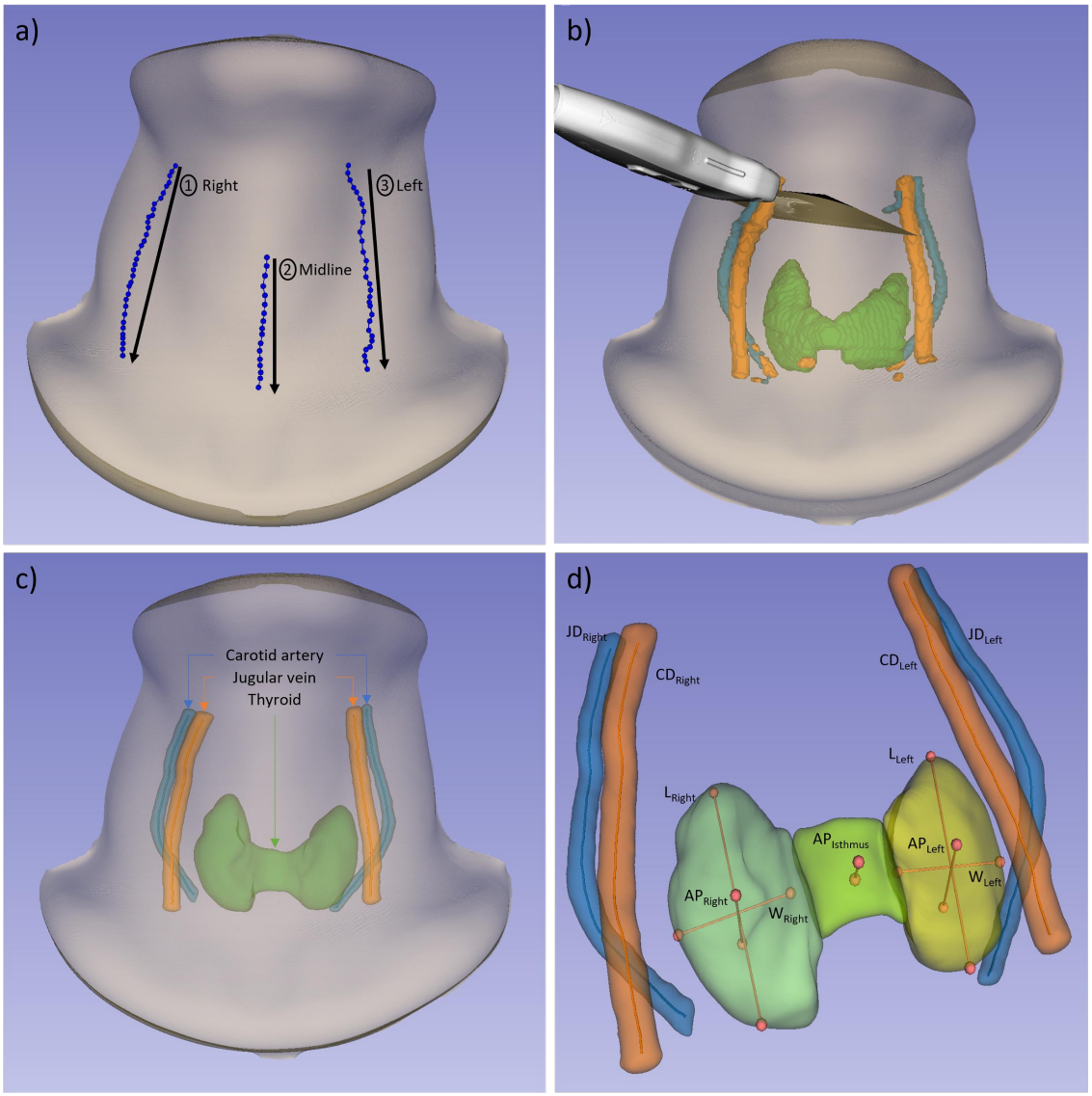


Figure 5.5: AI-assisted co-robotic tracked US 3D reconstruction and visualisation in thyroid phantom. a) Scanning methodology, showing the three sweeps needed to fully reconstruct the neck anatomy. Blue points show the transducer's trajectory during the scan. b) Voxel-based reconstruction of the carotid artery (orange), jugular vein (blue), and thyroid (green) (1 mm voxel size). c) Post-processed segmented volumes. d) Measurements of the thyroid, which is split into the left lobe, right lobe, and isthmus. L = length, W = width, AP = anterior-posterior diameter, JD = jugular diameter, CD = carotid diameter.

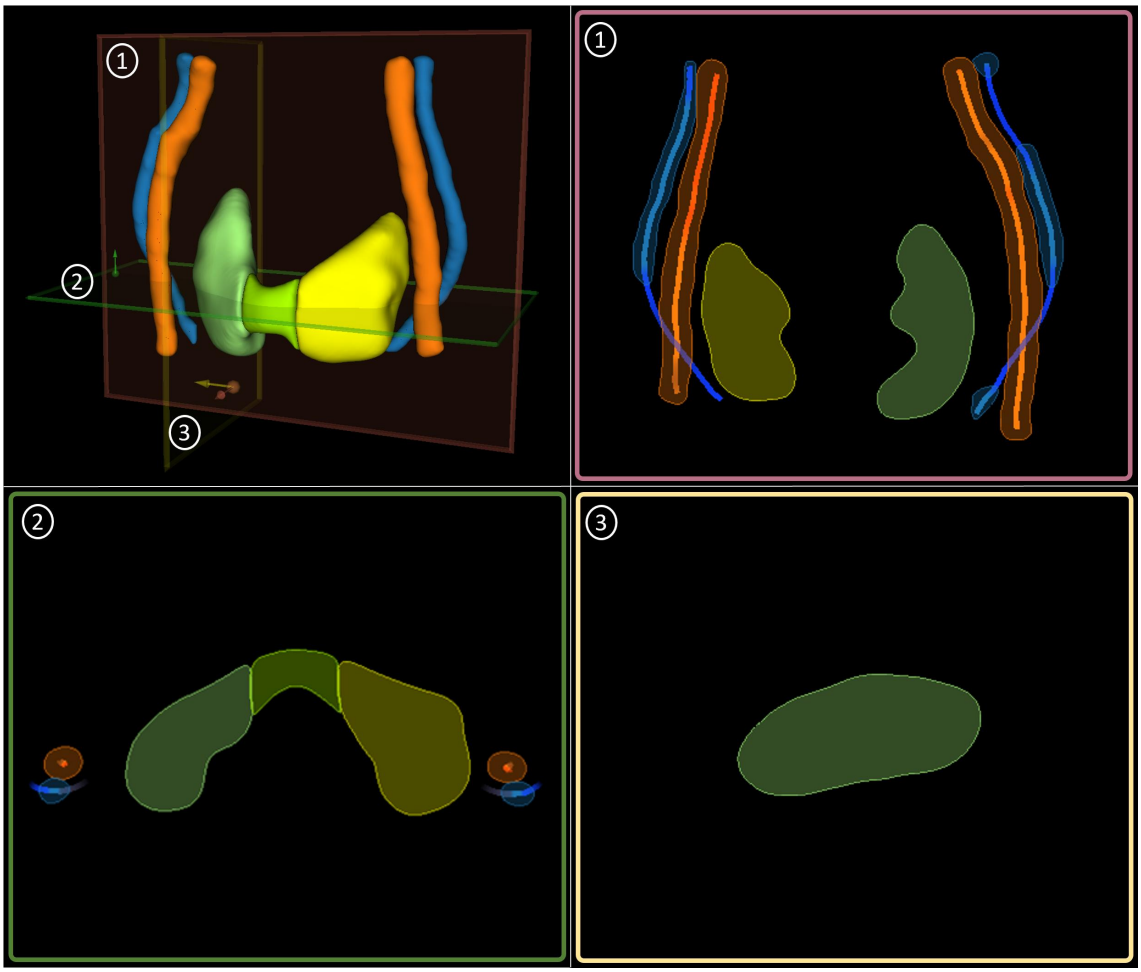


Figure 5.6: 3D render and cross-sectional images of the thyroid and major vessels in an anatomical phantom from MPR. Volumes are generated using a co-robotic tracked US system with active guidance and real-time AI anatomical segmentation for 3D reconstruction. The visualisation shows closed-surface representations of the reconstructed thyroid, split into its left and right lobes and isthmus. Also shown are the left and right carotid arteries, the jugular vein, and projections of their centrelines. Cross-sectional views show a slice through the segmented volume: 1) coronal plane, 2) axial plane, 3) sagittal plane (right thyroid lobe).

Table 5.2: Measurements of the phantom anatomy, comparing baseline values (B-mode and CT) with AI-based 3D US reconstruction under co-robotic (active and passive guidance) and freehand (optical and EM tracked) guidance. Values are mean ± std.

			Baseline		Co-robotic		Freehand	
			B-mode	CT	Active	Passive	Optical	EM
Vessels	Carotid	⌀(left) / mm	6.8 ± 0.7	N/A	7.0 ± 0.5	6.5 ± 0.2	7.2 ± 0.1	6.9 ± 0.1
		⌀(right) / mm	6.8 ± 1.1	N/A	7.1 ± 0.2	6.4 ± 0.5	7.6 ± 0.1	7.5 ± 0.1
	Jugular	⌀(left) / mm	4.4 ± 0.7	N/A	5.2 ± 0.2	4.4 ± 0.3	5.0 ± 0.1	5.1 ± 0.1
		⌀(right) / mm	4.6 ± 0.6	N/A	5.1 ± 0.2	4.4 ± 0.1	5.3 ± 0.1	5.4 ± 0.2
Thyroid	Left lobe	L / mm	48.4 ± 5.6	49.8 ± 0.8	48.9 ± 1.3	47.2 ± 0.2	54.0 ± 3.7	51.4 ± 1.3
		W / mm	24.6 ± 0.9	23.4 ± 0.4	23.1 ± 1.8	21.9 ± 2.2	21.8 ± 1.6	23.2 ± 1.1
		AP / mm	23.0 ± 0.8	24.1 ± 0.6	24.0 ± 0.8	22.4 ± 0.5	24.6 ± 1.3	20.6 ± 1.1
	Right lobe	L / mm	55.2 ± 1.1	55.6 ± 0.3	55.3 ± 1.0	54.6 ± 0.2	52.2 ± 2.2	56.5 ± 0.9
		W / mm	24.2 ± 0.1	23.5 ± 0.6	25.3 ± 1.2	23.4 ± 0.4	26.3 ± 2.3	23.4 ± 1.1
		AP / mm	23.2 ± 0.9	22.9 ± 0.4	22.3 ± 0.9	23.9 ± 0.9	22.1 ± 0.4	22.1 ± 0.4
	Total volume	$V_{\text{US-TV}}$ / mL	30.9 ± 2.0	30.6 ± 1.2	30.7 ± 2.8	28.2 ± 2.0	31.3 ± 2.0	28.3 ± 1.0
		$V_{\text{LCF-TV}}$ / mL	28.0 ± 1.8	27.8 ± 1.1	28.0 ± 2.6	25.7 ± 1.9	28.4 ± 1.8	25.8 ± 0.9
		$V_{\text{3D-TV}}$ / mL	N/A	37.9 ± 1.8	37.7 ± 1.3	33.4 ± 0.8	40.9 ± 1.1	37.8 ± 1.7

baseline (B-mode). For active co-robotic scanning, the right carotid artery and jugular vein were 7.1 ± 0.2 mm and 5.1 ± 0.2 mm; for freehand optical scanning, 7.6 ± 0.1 mm and 5.3 ± 0.1 mm, compared with 6.8 ± 1.1 mm and 4.6 ± 0.6 mm measured directly on 2D B-mode. Vessel measurement uncertainty for 3D reconstruction was also smaller than the B-mode baseline measurements. Thyroid lobe L, W, and AP diameters were within 1.5 mm of Baseline B-mode values using co-robotic active scanning.

3D US thyroid volumes were compared with the clinical standard technique using the conventional ellipsoid formula ($V_{\text{US-TV}}$, Equation 5.3) and lower-correction-factor ellipsoid formula ($V_{\text{LCF-TV}}$, Equation 5.4). Co-robotic active and freehand optical $V_{\text{US-TV}}$ values were 30.7 ± 2.8 mL and 31.3 ± 2.0 mL respectively, differing < 0.5 mL from the B-mode baseline value of 30.9 ± 2.0 mL. Co-robotic passive and freehand EM scans showed larger deviations (up to 2.7 mL), but remained within standard error limits of the reference measurements.

The CT-derived gold-standard thyroid volume was 37.9 ± 1.8 mL (Fig. 5.7), 7.4 mL higher than the estimate obtained using the ellipsoid formula. This was because, unlike the $V_{\text{3D-TV}}$ from CT, the $V_{\text{US-TV}}$ did not capture the irregular geometry of the enlarged thyroid and excluded the isthmus. The $V_{\text{3D-TV}}$ measurements from 3D US showed strong agreement with the CT reference, supporting this reasoning. Co-robotic active guidance yielded a volume of 37.7 ± 1.3 mL, while freehand EM scanning produced 37.8 ± 1.7 mL, corresponding to small biases of +0.2 mL (0.5%) and +0.1 mL (0.3%), respectively, relative to the CT baseline. Geometric agreement between US- and CT-derived reconstructions was also high, with a mean surface distance of 1.06 mm, a DSC of 0.87, and an HD95 of 3.07 mm. Vascular structures were not visible in the reference CT images and could therefore not be evaluated this way.

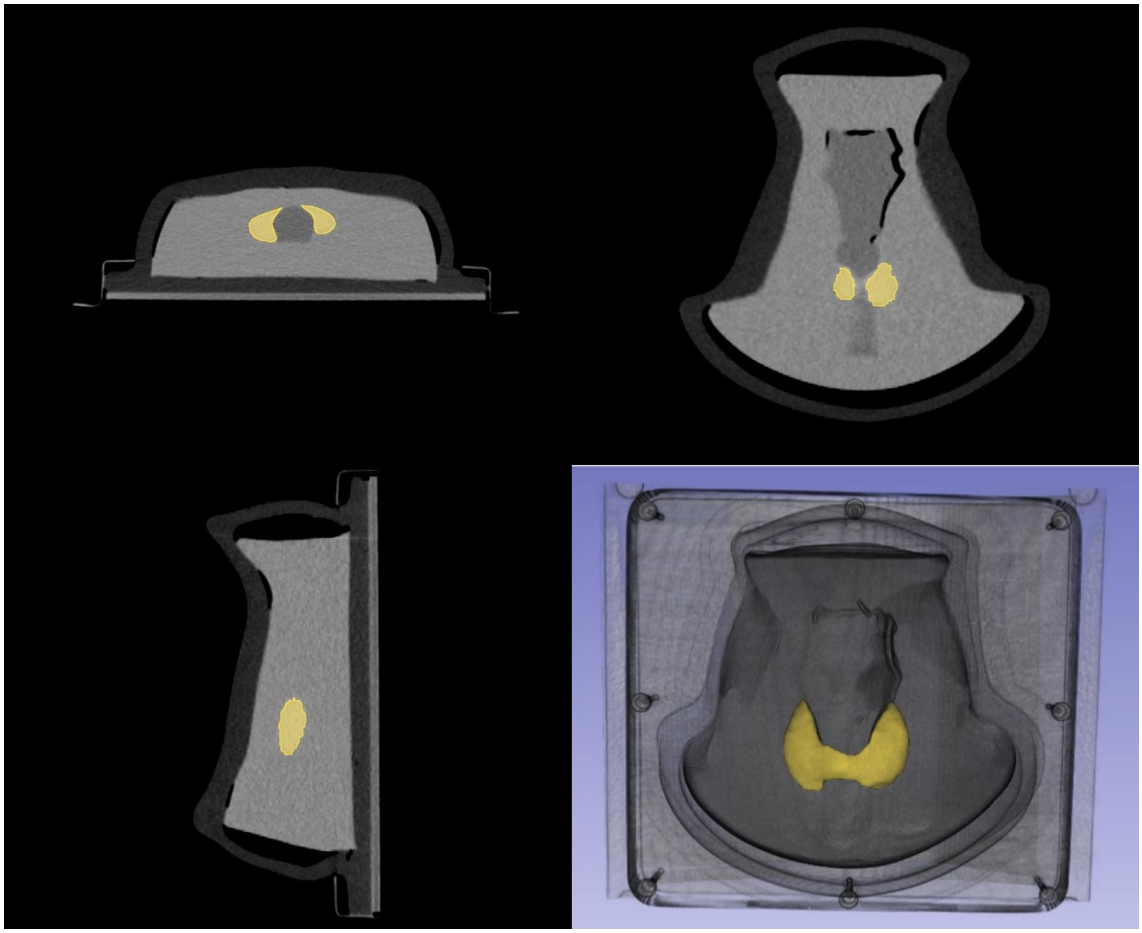


Figure 5.7: Reference CT imaging of thyroid phantom (CIRS Model 074), showing axial, coronal, and sagittal views. The thyroid was manually segmented in each slice, and is shown in yellow. The volume of the thyroid was 37.9 ± 1.8 mL.

Comparing freehand and co-robotic scanning approaches showed that both approaches enable accurate 3D US reconstruction, with the best method depending on the clinical application. Freehand scanning allowed unrestricted motion but was more prone to dropout, causing the thyroid and vessels to appear irregular in some cases, with visible discontinuities, increased surface noise, or holes in the 3D volumes, likely due to tracking jitter. Co-robotic scanning provided greater stability, improving the consistency of tracked US acquisition, but the scanning area was limited to the robot's workspace. Co-robotic trajectories were smoother than freehand scans, indicated by lower LDLJ and SPARC values, despite similar median scanning speeds (3.5 ± 1.0 and 2.8 ± 0.6 mm/s) (Figure 5.8). Of the two co-robotic control modes, active compliance with speed regulation enabled more precise control of the transducer's speed and orientation and helped the operator maintain a safe probe contact force, with a mean of 2.2 N, compared to 5.6 N under passive manual guidance, as measured by the robot's force-torque sensor.

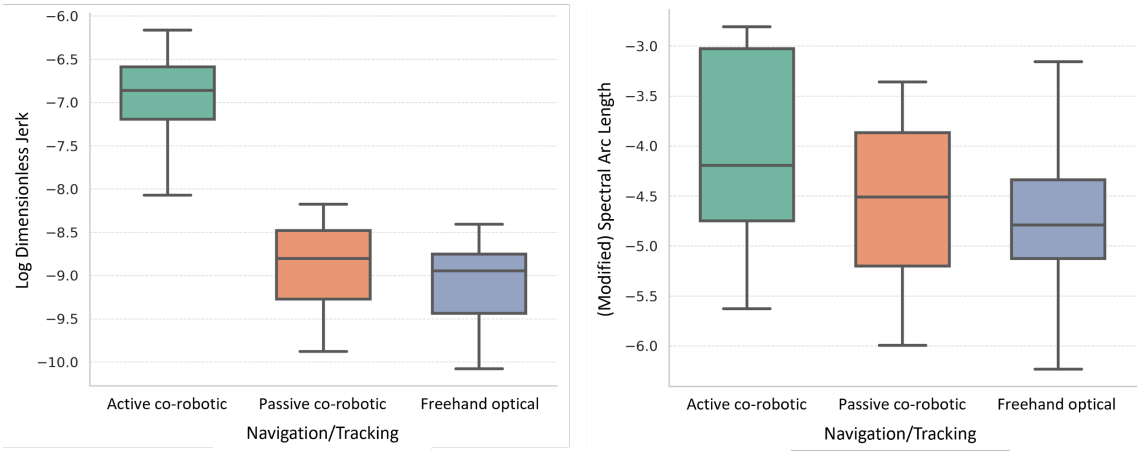


Figure 5.8: Log dimensionless jerk and spectral arm length for single-sweep scans of thyroid phantom using co-robotic navigation (active and passive) and freehand navigation (optical tracking). Boxplots show the median and quartiles for LDLJ and SPARC across five repeated sweeps.

5.4 Discussion and Conclusion

This study demonstrates that accurate 3D US imaging of the thyroid and major vessels can be achieved using a conventional 2D B-mode transducer, a collaborative robot arm, and real-time AI-based segmentation. This system enables low-latency acquisition and volume reconstruction, allowing anatomical structures to be visualised and measured immediately. These capabilities support the use of tracked US for thyroid volumetry, treatment planning, and image-guided interventions.

A key feature of the proposed framework is the use of co-robotic US rather than a fully autonomous RUS. In this approach, the clinician remains in direct control of the scan while the robot provides mechanical stability, force regulation, and trajectory smoothing.

At present, fully automatic RUSs still fall short of human scanning capabilities, lacking the ability to fully comprehend anatomy, diseases, physiology, and the physics of US imaging, all of which are important for ensuring optimal imaging and patient safety [25, 26]. Co-robotic systems may therefore represent a practical intermediate stage of automation, assisting the clinician while addressing the safety, regulatory, and ethical challenges associated with fully autonomous RUS [54], aiding in the translation of these devices from research settings to clinical practice [24].

This co-robotic approach also aligns well with current clinical workflows, particularly in surgical and minimally invasive interventional environments where robots are already used to position needles or therapy transducers. Because the clinician hand-guides the robot, the learning curve is shorter than with other robotic approaches, facilitating clinical translation. Using the same robotic platform for imaging and treatment also ensures that all spatial data is within a single coordinate system. This removes the necessity for external optical or EM tracking [55], simplifying the workflow and overcoming common issues such as line-of-sight occlusion and EMI in procedures such as image-guided biopsy and HIFU. In procedures requiring multiple tracked objects, registration may be more complex, and multimodal tracking with co-registered robotic and optical or EM tracking may be beneficial.

The experimental results demonstrate that co-robotic scanning can provide stable acquisition and high-quality 3D reconstructions of thyroid anatomy. Compared with freehand scanning, co-robotic trajectories were smoother and less susceptible to probe instability or undersampling, resulting in more consistent volumetric reconstructions. Force-regulated compliance also helped the operator maintain stable probe contact with the tissue surface, which is important for image quality and patient comfort. Despite these differences, both freehand and co-robotic scanning approaches produced accurate thyroid reconstructions, suggesting that the tracked US framework is robust to different acquisition strategies.

The lightweight U-Net enabled real-time segmentation of the thyroid and major vessels at 27 FPS, with a thyroid DSC of 0.88, comparable to previous thyroid US segmentation methods with DSC in the range 0.83–0.88 [9, 56]. This allowed anatomical structures to be reconstructed and anatomical measurements to be extracted automatically, potentially reducing operator-dependence. Thyroid length and carotid diameter measurements were consistent with reference B-mode US values, supporting the reliability of the reconstructed volumes for quantitative assessment. Carotid diameters were within 3-5% error of reference US measurements, compared to 7-9% errors recorded in a previous freehand optical-tracked 3D US study [14].

Direct volumetric measurement of the thyroid from the reconstructed 3D US data showed strong agreement with the CT-derived reference volume, with errors below 1%, compared with previous 1-8% errors using autonomous RUS [7, 23]. In contrast, thyroid volumes estimated using the clinical-standard ellipsoid formula were substantially lower, underestimating the true volume of the phantom's slightly enlarged thyroid by approximately 22%.

This finding is consistent with previous studies showing that the ellipsoid approximation does not accurately represent the irregular shape seen in patients with an enlarged or multinodular thyroid [1, 7, 9]. The ability to measure thyroid volume directly from reconstructed US data may therefore provide a more accurate and reproducible alternative to conventional 2D sonographic volumetry.

Nevertheless, there are important limitations. This study used a single high-frequency linear transducer, and further investigation is needed to evaluate the approach across other US systems and imaging settings. In addition, validation was performed using an anatomical phantom rather than *in vivo*, where patient motion, deformation, and acoustic shadowing would introduce additional challenges [31, 57]. To translate the device for clinical use, the AI segmentation model would require additional training and validation using large annotated clinical datasets, which can be expensive and challenging to acquire in practice [58, 59]. This model would have to be sufficiently generalisable across anatomical variants, such as the thyroid pyramidal lobe, which is present in around 20% of individuals [60], as well as pathologies that affect thyroid shape, such as exophytic thyroid nodules and multinodular goitre. Furthermore, only the thyroid and major vessels were segmented in this study. Extending the segmentation to include additional structures, such as the trachea, oesophagus, muscles, and lymph nodes, may further improve the clinical utility of reconstructed volumes for surgical planning [61]. Inter-operator variability was not evaluated, and future studies should investigate how different clinicians interact with the co-robotic system and how these interactions affect scanning performance [62]. Understanding operator workload and ergonomics is also important to assess whether co-robotic approaches hinder natural scanning techniques or increase musculoskeletal strain, which is a recognised issue in diagnostic sonography [63] but remains largely unexplored in RUS.

The tracked US framework presented here could support a range of clinical applications beyond thyroid volumetry. 3D US can improve spatial understanding of anatomy and assist in treatment planning, particularly when determining lesion location and surgical margins in thyroid disease [64]. Similarly, 3D vessel reconstruction enables measurement of the lumen cross-sectional area, vessel curvature, and tortuosity, which have clinical value for vascular interventions and plaque assessment in the carotid artery [14, 37, 65], as well as cannulation of the internal jugular vein for robotic cardiac surgery [66]. More broadly, co-registered 3D US imaging could facilitate minimally- and non-invasive interventions such as microwave or radiofrequency ablation [4, 67], biopsy [68], and HIFU procedures [5, 69, 70], providing a spatial map of the anatomy and treatment pathway. Future systems may also integrate 3D US with Augmented Reality (AR) to further improve spatial understanding and clinical communication in next-generation imaging [11, 71].

5.5 Acknowledgments

This study has received funding from the Engineering and Physical Sciences Research Council (EPSRC) through the following awards: EP/S024336/1 and EP/V04799X/1. J. Chandler was supported by the Leverhulme Trust and the Royal Academy of Engineering under a RAEng/Leverhulme Trust Research Fellowship (LTRF-2425-21-154).

Bibliography

[1] A. Viduetsky and C. L. Herrejon, "Sonographic Evaluation of Thyroid Size: A Review of Important Measurement Parameters," *Journal of Diagnostic Medical Sonography*, vol. 35, pp. 206–210, May 2019.

[2] M. Dighe, R. Barr, J. Bojunga, V. Cantisani, M. C. Chammas, D. Cosgrove, X. W. Cui, Y. Dong, F. Fenner, M. Radzina, S. Vinayak, J. M. Xu, and C. F. Dietrich, "Thyroid Ultrasound: State of the Art Part 1 – Thyroid Ultrasound reporting and Diffuse Thyroid Diseases," *Medical Ultrasonography*, vol. 19, pp. 79–93, Jan. 2017.

[3] N. M. Bekedam, J. N. Smit, P. K. de Koekkoek Doll, M. J. A. van Alphen, R. L. P. van Veen, L. H. E. Karssemakers, M. B. Karakullukcu, and L. E. Smeele, "Intra-operative resection margin model of tongue carcinoma using 3D reconstructed ultrasound," *Advances in Oral and Maxillofacial Surgery*, vol. 4, p. 100154, Oct. 2021.

[4] D. Li, Z. Cheng, G. Chen, F. Liu, W. Wu, J. Yu, Y. Gu, F. Liu, C. Ren, and P. Liang, "A multimodality imaging-compatible insertion robot with a respiratory motion calibration module designed for ablation of liver tumors: a preclinical study," *International Journal of Hyperthermia*, vol. 34, pp. 1194–1201, Nov. 2018.

[5] L. J. Howell, V. W.-S. Chan, J. Zhong, H. H.-L. Ng, J. H. Chandler, J. R. McLaughlan, and T. M. Wah, "Histotripsy: New Frontiers for Noninvasive Ablation," *The Arab Journal of Interventional Radiology*, vol. 9, p. 61, Dec. 2025.

[6] P. Szumowski, M. Mojsak, S. Abdelrazek, M. Sykała, A. Amelian-Fiłonowicz, D. Jurgilewicz, and J. Myśliwiec, "Calculation of therapeutic activity of radioiodine in Graves' disease by means of Marinelli's formula, using technetium (99mTc) scintigraphy," *Endocrine*, vol. 54, pp. 751–756, Dec. 2016.

[7] J. Zielke, C. Eilers, B. Busam, W. Weber, N. Navab, and T. Wendler, "RSV: Robotic Sonography for Thyroid Volumetry," *IEEE Robotics and Automation Letters*, vol. 7, pp. 3342–3348, Apr. 2022.

[8] C. Konca and A. H. Elhan, "Unveiling the Accuracy of Ultrasonographic Assessment of Thyroid Volume: A Comparative Analysis of Ultrasonographic Measurements and Specimen Volumes," *Journal of Clinical Medicine*, vol. 12, p. 6619, Oct. 2023.

[9] M. Krönke, C. Eilers, D. Dimova, M. Köhler, G. Buschner, L. Schweiger, L. Konstantinidou, M. Makowski, J. Nagarajah, N. Navab, W. Weber, and T. Wendler, "Tracked 3D ultrasound and deep neural network-based thyroid segmentation reduce interobserver variability in thyroid volumetry," *PLOS ONE*, vol. 17, p. e0268550, July 2022.

[10] M. H. Mozaffari and W.-S. Lee, "Freehand 3-D Ultrasound Imaging: A Systematic Review," *Ultrasound in Medicine & Biology*, vol. 43, pp. 2099–2124, Oct. 2017.

[11] F. Mohamed, C. V. Siang, F. Mohamed, and C. V. Siang, "A Survey on 3D Ultrasound Reconstruction Techniques," in *Artificial Intelligence - Applications in Medicine and Biology*, IntechOpen, Apr. 2019.

[12] R. W. Prager, A. Gee, and L. Berman, "Stradx: real-time acquisition and visualization of freehand three-dimensional ultrasound," *Medical Image Analysis*, vol. 3, pp. 129–140, June 1999.

[13] C. Peng, Q. Cai, M. Chen, and X. Jiang, "Recent Advances in Tracking Devices for Biomedical Ultrasound Imaging Applications," *Micromachines*, vol. 13, p. 1855, Oct. 2022.

[14] S.-W. Chung, C.-C. Shih, and C.-C. Huang, "Freehand three-dimensional ultrasound imaging of carotid artery using motion tracking technology," *Ultrasonics*, vol. 74, pp. 11–20, Feb. 2017.

[15] A. M. Aurand, J. S. Dufour, and W. S. Marras, "Accuracy map of an optical motion capture system with 42 or 21 cameras in a large measurement volume," *Journal of Biomechanics*, vol. 58, pp. 237–240, June 2017.

[16] F. Poulin and L. P. Amiot, "Interference during the use of an electromagnetic tracking system under OR conditions," *Journal of Biomechanics*, vol. 35, pp. 733–737, June 2002.

[17] H. Sadjadi, K. Hashtrudi-Zaad, and G. Fichtinger, "Simultaneous Electromagnetic Tracking and Calibration for Dynamic Field Distortion Compensation," *IEEE Transactions on Biomedical Engineering*, vol. 63, pp. 1771–1781, Aug. 2016.

[18] N. M. Bekedam, L. H. E. Karssemakers, M. J. A. van Alphen, R. L. P. van Veen, L. E. Smeele, and M. B. Karakullukcu, "Comparison of image quality of 3D ultrasound: motorized acquisition versus freehand navigated acquisition, a phantom study," *International Journal of Computer Assisted Radiology and Surgery*, vol. 18, pp. 1649–1663, Sept. 2023.

[19] A. M. Priester, S. Natarajan, and M. O. Culjat, "Robotic ultrasound systems in medicine," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 60, pp. 507–523, Mar. 2013.

[20] X. Bao, S. Wang, L. Zheng, R. J. Housden, J. V. Hajnal, and K. Rhode, "A Novel

Ultrasound Robot With Force/Torque Measurement and Control for Safe and Efficient Scanning," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–12, 2023.

[21] F. von Haxthausen, S. Böttger, D. Wulff, J. Hagenah, V. García-Vázquez, and S. Ipsen, "Medical Robotics for Ultrasound Imaging: Current Systems and Future Trends," *Current Robotics Reports*, vol. 2, pp. 55–71, Mar. 2021.

[22] Z. Jiang, S. E. Salcudean, and N. Navab, "Robotic ultrasound imaging: State-of-the-art and future perspectives," *Medical Image Analysis*, vol. 89, p. 102878, Oct. 2023.

[23] X.-X. Lin, M.-D. Li, S.-M. Ruan, W.-P. Ke, H.-R. Zhang, H. Huang, S.-H. Wu, M.-Q. Cheng, W.-J. Tong, H.-T. Hu, D.-N. He, R.-F. Lu, Y.-D. Lin, M. Kuang, M.-D. Lu, L.-D. Chen, Q.-H. Huang, and W. Wang, "Autonomous robotic ultrasound scanning system: a key to enhancing image analysis reproducibility and observer consistency in ultrasound imaging," *Frontiers in Robotics and AI*, vol. 12, Feb. 2025.

[24] Y. Bi, Z. Jiang, F. Duelmer, D. Huang, and N. Navab, "Machine Learning in Robotic Ultrasound Imaging: Challenges and Perspectives," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 7, pp. 335–357, July 2024.

[25] G.-Z. Yang, J. Cambias, K. Cleary, E. Daimler, J. Drake, P. E. Dupont, N. Hata, P. Kazanzides, S. Martel, R. V. Patel, V. J. Santos, and R. H. Taylor, "Medical robotics-Regulatory, ethical, and legal considerations for increasing levels of autonomy," *Science Robotics*, vol. 2, p. eaam8638, Mar. 2017.

[26] F. Ferraguti, C. Talignani Landi, L. Sabattini, M. Bonfè, C. Fantuzzi, and C. Secchi, "A variable admittance control strategy for stable physical human–robot interaction," *The International Journal of Robotics Research*, vol. 38, pp. 747–765, May 2019.

[27] Q. Huang, B. Wu, J. Lan, and X. Li, "Fully Automatic Three-Dimensional Ultrasound Imaging Based on Conventional B-Scan," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 12, pp. 426–436, Apr. 2018.

[28] Q. Huang, J. Lan, and X. Li, "Robotic Arm Based Automatic Ultrasound Scanning for Three-Dimensional Imaging," *IEEE Transactions on Industrial Informatics*, vol. PP, pp. 1–1, Sept. 2018.

[29] K. Su, J. Liu, X. Ren, Y. Huo, G. Du, W. Zhao, X. Wang, B. Liang, D. Li, and P. X. Liu, "A fully autonomous robotic ultrasound system for thyroid scanning," *Nature Communications*, vol. 15, p. 4004, May 2024.

[30] J. T. Kaminski, K. Rafatzand, and H. K. Zhang, "Feasibility of Robot-Assisted Ultrasound Imaging with Force Feedback for Assessment of Thyroid Diseases," *Proceedings of the International Society for Optical Engineering*, vol. 11315, p. 113151D, Feb. 2020.

[31] R. Kojcev, A. Khakzar, B. Fuerst, O. Zettinig, C. Fahkry, R. DeJong, J. Richmon,

R. Taylor, E. Sinibaldi, and N. Navab, "On the reproducibility of expert-operated and robotic ultrasound acquisitions," *International Journal of Computer Assisted Radiology and Surgery*, vol. 12, pp. 1003–1011, June 2017.

[32] Q. Huang, B. Gao, and M. Wang, "Robot-Assisted Autonomous Ultrasound Imaging for Carotid Artery," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–9, 2024.

[33] Y. Bi, Z. Jiang, Y. Gao, T. Wendler, A. Karlas, and N. Navab, "VesNet-RL: Simulation-Based Reinforcement Learning for Real-World US Probe Navigation," *IEEE Robotics and Automation Letters*, vol. 7, pp. 6638–6645, July 2022.

[34] K. AlMuhanna, M. M. Hossain, L. Zhao, J. Fischell, G. Kowalewski, M. Dux, S. Sikdar, and B. K. Lal, "Carotid plaque morphometric assessment with three-dimensional ultrasound imaging," *Journal of Vascular Surgery*, vol. 61, pp. 690–697, Mar. 2015.

[35] H. Jiang, A. Zhao, Q. Yang, X. Yan, T. Wang, Y. Wang, N. Jia, J. Wang, G. Wu, Y. Yue, S. Luo, H. Wang, L. Ren, S. Chen, P. Liu, G. Yao, W. Yang, S. Song, X. Li, K. He, and G. Huang, "Towards expert-level autonomous carotid ultrasonography with large-scale learning-based robotic system," *Nature Communications*, vol. 16, p. 7893, Aug. 2025.

[36] Z. Wang, Y. Han, B. Zhao, H. Xie, L. Yao, B. Li, M. Q.-H. Meng, and Y. Hu, "Autonomous Robotic System for Carotid Artery Ultrasound Scanning With Visual Servo Navigation," *IEEE Transactions on Medical Robotics and Bionics*, vol. 6, pp. 1436–1447, Nov. 2024.

[37] G. Faoro, S. Maglio, S. Pane, V. Iacovacci, and A. Menciassi, "An Artificial Intelligence-Aided Robotic Platform for Ultrasound-Guided Transcarotid Revascularization," *IEEE Robotics and Automation Letters*, vol. 8, pp. 2349–2356, Apr. 2023.

[38] G. Ning, H. Liang, X. Zhang, and H. Liao, "Inverse-Reinforcement-Learning-Based Robotic Ultrasound Active Compliance Control in Uncertain Environments," *IEEE Transactions on Industrial Electronics*, vol. 71, pp. 1686–1696, Feb. 2024.

[39] D. Dall'Alba, L. Busellato, T. R. Savarimuthu, Z. Cheng, and I. Iturrate, "Imitation Learning of Compression Pattern in Robotic-Assisted Ultrasound Examination Using Kernelized Movement Primitives," *IEEE Transactions on Medical Robotics and Bionics*, vol. 6, pp. 1567–1580, Nov. 2024.

[40] L. Howell, M. Waterston, T. M. Wah, J. H. Chandler, and J. R. McLaughlan, "A Unified Platform and Quality Assurance Framework for 3D Ultrasound Reconstruction with Robotic, Optical, and Electromagnetic Tracking," Mar. 2026. arXiv:2603.20077 [cs].

[41] M. K. Lee, D. G. Na, L. Joo, J. Y. Lee, E. J. Ha, J.-H. Kim, S. L. Jung, and J. H. Baek, "Standardized Imaging and Reporting for Thyroid Ultrasound: Korean Society

of Thyroid Radiology Consensus Statement and Recommendation," *Korean Journal of Radiology*, vol. 24, pp. 22–30, Jan. 2023.

[42] A. Lasso, T. Heffter, A. Rankin, C. Pinter, T. Ungi, and G. Fichtinger, "PLUS: Open-source toolkit for ultrasound-guided intervention systems," *IEEE Transactions on Biomedical Engineering*, vol. 61, pp. 2527–2537, 2014.

[43] G. Carbajal, A. Lasso, A. Gomez, and G. Fichtinger, "Improving N-wire phantom-based freehand ultrasound calibration," *International journal of computer assisted radiology and surgery*, vol. 8, July 2013.

[44] M.-P. Sopuschek, M. Freesmeyer, T. Winkens, C. Kühnel, M. Petersen, F. Gühne, A. Werner, and P. Seifert, "Standard operating procedure (SOP) for cervical ultrasound cine loop video sequences in the follow-up of differentiated thyroid carcinoma (DTC)," *Endocrine*, vol. 87, pp. 635–647, Feb. 2025.

[45] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," Jan. 2017.

[46] S. A. Taghanaki, Y. Zheng, S. Kevin Zhou, B. Georgescu, P. Sharma, D. Xu, D. Comaniciu, and G. Hamarneh, "Combo loss: Handling input and output imbalance in multi-organ segmentation," *Computerized Medical Imaging and Graphics*, vol. 75, pp. 24–33, July 2019.

[47] L. Howell, N. Ingram, R. Lapham, A. Morrell, and J. R. McLaughlan, "Deep learning for real-time multi-class segmentation of artefacts in lung ultrasound," *Ultrasonics*, vol. 140, p. 107251, May 2024.

[48] M. Bozorgi and F. Lindseth, "GPU-based multi-volume ray casting within VTK for medical applications," *International Journal of Computer Assisted Radiology and Surgery*, vol. 10, pp. 293–300, Mar. 2015.

[49] W. Shabana, E. Peeters, and M. De Maeseneer, "Measuring Thyroid Gland Volume: Should We Change the Correction Factor?," *American Journal of Roentgenology*, vol. 186, pp. 234–236, Jan. 2006.

[50] J. Brunn, U. Block, G. Ruf, I. Bos, W. P. Kunze, and P. C. Scriba, "[Volumetric analysis of thyroid lobes by real-time ultrasound (author's transl)]," *Deutsche Medizinische Wochenschrift (1946)*, vol. 106, pp. 1338–1340, Oct. 1981.

[51] A. Savitzky and M. J. E. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures.," *Analytical Chemistry*, vol. 36, pp. 1627–1639, July 1964.

[52] S. Balasubramanian, A. Melendez-Calderon, A. Roby-Brami, and E. Burdet, "On the analysis of movement smoothness," *Journal of NeuroEngineering and Rehabilitation*, vol. 12, p. 112, Dec. 2015.

[53] Universal robots, "Technical Specifications UR3e," Aug. 2025.

[54] T. Haidegger, "Autonomy for Surgical Robots: Concepts and Paradigms," *IEEE Transactions on Medical Robotics and Bionics*, vol. 1, pp. 65–76, May 2019.

[55] K. Li, Y. Xu, and M. Q.-H. Meng, "An Overview of Systems and Techniques for Autonomous Robotic Ultrasound Acquisitions," *IEEE Transactions on Medical Robotics and Bionics*, vol. 3, pp. 510–524, May 2021.

[56] P. Poudel, A. Illanes, D. Sheet, and M. Friebe, "Evaluation of Commonly Used Algorithms for Thyroid Ultrasound Images Segmentation and Improvement Using Machine Learning Approaches," *Journal of Healthcare Engineering*, vol. 2018, p. 8087624, Sept. 2018.

[57] S. Virga, R. Göbl, M. Baust, N. Navab, and C. Hennersperger, "Use the force: deformation correction in robotic 3D ultrasound," *International Journal of Computer Assisted Radiology and Surgery*, vol. 13, pp. 619–627, May 2018.

[58] C.-L. Cao, Q.-L. Li, J. Tong, L.-N. Shi, W.-X. Li, Y. Xu, J. Cheng, T.-T. Du, J. Li, and X.-W. Cui, "Artificial intelligence in thyroid ultrasound," *Frontiers in Oncology*, vol. 13, May 2023.

[59] L. Howell, A. Zarei, T. M. Wah, J. H. Chandler, S. Karthik, Z. Court, H. Ng, and J. R. McLaughlan, "RADEX: a rule-based clinical and radiology data extraction tool demonstrated on thyroid ultrasound reports," *European Radiology*, vol. 35, pp. 4506–4517, Aug. 2025.

[60] C. Mortensen, H. Lockyer, and E. Loveday, "The incidence and morphological features of pyramidal lobe on thyroid ultrasound," *Ultrasound: Journal of the British Medical Ultrasound Society*, vol. 22, pp. 192–198, Nov. 2014.

[61] H. Dou, L. Han, Y. He, J. Xu, N. Ravikumar, R. Mann, A. F. Frangi, P.-T. Yap, and Y. Huang, "Localizing the Recurrent Laryngeal Nerve via Ultrasound with a Bayesian Shape Framework," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022* (L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, eds.), Lecture Notes in Computer Science, (Cham), pp. 258–267, Springer Nature Switzerland, 2022.

[62] Y. Wang, Q. Yang, L. Drukker, A. Papageorghiou, Y. Hu, and J. A. Noble, "Task model-specific operator skill assessment in routine fetal ultrasound scanning," *International Journal of Computer Assisted Radiology and Surgery*, vol. 17, no. 8, pp. 1437–1444, 2022.

[63] K. Evans, S. Roll, and J. Baker, "Work-Related Musculoskeletal Disorders (WRMSD) Among Registered Diagnostic Medical Sonographers and Vascular Technologists: A Representative Sample," *Journal of Diagnostic Medical Sonography*, vol. 25, pp. 287–299, Nov. 2009.

[64] Z. Wu, H. Sun, Z. Li, S. Chen, L. Han, Y. Wang, W. Duan, and W. Zhang, "Application of visualization three-dimensional medical imaging technology based on mobile

platform in surgery of giant thyroid tumors: a case report," *Journal of Medical Case Reports*, vol. 19, p. 496, Oct. 2025.

[65] S. Merouche, L. Allard, E. Montagnon, G. Soulez, P. Bigras, and G. Cloutier, "A Robotic Ultrasound Scanner for Automatic Vessel Tracking and Three-Dimensional Reconstruction of B-Mode Images," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 63, pp. 35–46, Jan. 2016.

[66] Y. Wang, G. Wang, and C.-q. Gao, "Ultrasound-guided cannulation of the internal jugular vein in robotic cardiac surgery," *Chinese Medical Journal*, vol. 126, pp. 2414–2417, July 2013.

[67] T. Boers, S. J. Braak, W. M. Brink, M. Versluis, and S. Manohar, "3D ultrasound guidance for radiofrequency ablation in an anthropomorphic thyroid nodule phantom," *European Radiology Experimental*, vol. 8, p. 115, Oct. 2024.

[68] P. Chatelain, A. Krupa, and N. Navab, "3D ultrasound-guided robotic steering of a flexible needle via visual servoing," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2250–2255, May 2015.

[69] K. B. Bader, S. A. Hendley, and V. Bollen, "Assessment of Collaborative Robot (Cobot)-Assisted Histotripsy for Venous Clot Ablation," *IEEE Transactions on Biomedical Engineering*, vol. 68, pp. 1220–1228, Apr. 2021.

[70] C. Y. An, J. H. Syu, C. S. Tseng, and C.-J. Chang, "An Ultrasound Imaging-Guided Robotic HIFU Ablation Experimental System and Accuracy Evaluations," *Applied Bionics and Biomechanics*, vol. 2017, p. e5868695, Apr. 2017.

[71] S. Bazzani, S. Puliatti, S. Ferretti, G. Bianchi, C. Secchi, and F. Ferraguti, "Ultrasonographic-guided robotic-assisted percutaneous nephrolithotomy," *Communications Engineering*, vol. 4, p. 120, July 2025.

Chapter 6

Discussion

This thesis presents developments in imaging-guidance technology for Ultrasound-guided High-Intensity Focused Ultrasound (USgHIFU), introducing a novel co-robotic system for patient-specific 3D anatomical reconstruction. The development of lightweight DL models enabled real-time delineation of structures in ultrasound and subsequent 3D US reconstruction, as demonstrated in phantoms. This improves spatial understanding and has the potential to streamline the HIFU procedure by reducing operator dependence, improving safety and efficacy, and shortening treatment duration. Beyond HIFU, this work also has wide-ranging implications for other diagnostic and interventional applications, including radioiodine therapy, US-guided biopsy, Radiofrequency Ablation (RFA), and Microwave Ablation (MWA) (Section 6.1). Despite this, several challenges remain which would need to be addressed before translating these methods from the lab to clinical settings (Section 6.2). These challenges form the basis for future research outside the scope of this thesis (Section 6.3).

6.1 Implications

The methods and findings presented in this thesis support the development of ultrasound systems for a variety of diagnostic and treatment guidance applications. For example, in Chapter 2, applying multiclass DL to Lung Ultrasound (LUS) enabled the development of a tool for automatic monitoring of pneumonia severity, demonstrating the potential to support clinical decision-making in acute care settings. More generally, the use of AI segmentation has implications for addressing the well-documented operator dependence of ultrasound [72, 93, 103]. This is particularly relevant in low- and middle-income countries, where access to highly trained sonographers is limited [104].

In particular, developments in lightweight AI models for real-time segmentation and 3D reconstruction are useful as they provide immediate anatomical visualisation without introducing visible latency. This is important given the clinical use of ultrasound as a dynamic

imaging tool, where processing delays could disrupt feedback, causing anatomy to appear to lag behind probe motion and reducing the clinical usability. For robotic or image-guided procedures, excessive latency could also introduce errors in the spatial correspondence between the displayed anatomy and the probe's true position, thereby affecting targeting accuracy.

The data extraction tool developed in Chapter 3 contributes to the broader medical AI and research landscape by enabling scalable extraction of structured information from free-text clinical reports in settings where hardware limitations, explainability, or reproducibility are concerns. Curation of large, high-quality datasets is important not only for training and validating robust AI models but also for cohort selection in clinical trials, clinical audits, and other research applications [105]. The Leeds Thyroid Ultraosund Dataset (LTUS) dataset is a good example, as one of the largest thyroid datasets (40,882 images from 1,650 cases), with detailed clinical annotations, including nodule grading, thyroid indications, and surgical history. This richness enables the development of clinically meaningful models for HIFU treatment planning and other clinical decision support systems in future work (Section 6.3).

In Chapter 4, the development of an open-source platform for tracked ultrasound and a phantom-based Quality Assurance (QA) pipeline has important implications for the reproducibility of tracked ultrasound research. Providing tools to assess geometric reconstruction supports the development and evaluation of existing systems and has the potential to accelerate research not only in HIFU but also in other image-guided procedures such as US-guided biopsy [106], MWA, and RFA [107, 108]. The suite of metrics provides a comprehensive overview of geometric uncertainty under controlled conditions. This is important since, in practice, the clinical requirements vary with the end use case, and metric requirements should be defined based on the application in advance of testing. For example, acceptable DSC and HD could be defined based on the size of treatment margins, which depends on the organ site, proximity to critical structures, and precision required for the intervention. For example, high-precision interventions in neurosurgery may require smaller HD values than ablation of solid tumours in organs such as the liver, where larger margins may be acceptable. The Ultrasound Quality Assurance in 3-Dimensions (US-QA-3D) approach enables transparent standardised reporting and improves confidence in the reliability of 3D US systems, which is important for their safe and effective translation into clinical practice.

In Chapter 5, these methods were applied to thyroid ultrasound data, demonstrating that thyroid volumetry based on 3D reconstruction could be more accurate than the clinical standard approach. This difference has direct implications for radioiodine therapy planning in Graves' disease, which is used to treat approximately 10,000 patients each year in the United Kingdom (UK) [109] and relies on an accurate determination of thyroid volume [110]. 3D vessel reconstruction also enabled measurement of the lumen cross-sectional

area, vessel curvature, and tortuosity, which have clinical value for vascular interventions and plaque assessment in the carotid artery [79, 90, 111] and internal jugular vein [112]. Most significantly, 3D US with real-time DL-based segmentation enabled accurate multi-organ segmentation and addressed key limitations of existing HIFU planning systems that rely on manual segmentation and 2D ultrasound. This automated approach has the potential to reduce the gap between US- and MRI-guided HIFU in terms of planning and guidance, improving safety through better identification of critical structures and efficacy through more precise targeting and coverage. The use of co-robotics rather than fully automated scanning ensures that the clinician retains full control of the transducer's position and addresses ethical and regulatory concerns associated with autonomous systems, thereby lowering barriers to clinical translation and supporting integration into existing workflows [96, 97].

For benign symptomatic thyroid nodules, the total number of patients treated with HIFU worldwide increased 10x between 2024 and 2025, from approximately 235 to 2,620 [113]. This represents part of a growing trend towards minimally and non-invasive image-guided therapies worldwide [114]. The technical contributions of this work directly address current limitations of HIFU treatment planning and guidance systems for thyroid nodules (Chapter 1), which have been identified as barriers to wider clinical adoption. By enabling more reliable and automated planning, this research supports the development of HIFU as a viable non-invasive, non-ionising alternative to thyroid surgery, with the potential to reduce complications, recovery time, and costs.

These advances are a step towards more reproducible semi-automated HIFU treatment planning: automatically assessing the size and location of the target, then planning the optimal positions and durations of HIFU exposures based on the proximity to vessels and other sensitive structures. These capabilities could reduce the overall treatment time, improve consistency between operators and centres, and facilitate large-scale clinical studies by standardising workflows. In the longer term, this would support the generation of robust clinical and technical evidence required for regulatory approval and guideline endorsement by bodies such as National Institute for Health and Care Excellence (NICE) [25]. This would enable wider adoption of HIFU for thyroid nodules and other challenging tumours, including those of the liver, kidney, breast, and prostate, thereby improving patient outcomes.

6.2 Limitations

Despite these advances, several limitations remain to be addressed. Chapter 2 developed the lightweight U-Net architecture and dynamic training regimes, which were suitable for demonstrating multi-class segmentation in phantoms. However, small dataset size and high image homogeneity limited translation to clinical LUS data. This study also did not include an assessment of intra-annotator variability, as no commonly labelled images

were available. Consequently, the uncertainty associated with data annotation could not be quantified, limiting the ability to interpret segmentation performance against baseline annotation uncertainty. Lessons learned from this early work informed subsequent work, including Chapter 3, which extracted a large thyroid US dataset, and included inter-annotator assessment of diagnostic labels using Cohen's Kappa. This was used as a basis for curating the LTUS dataset, where inter-annotator uncertainty associated with segmentation was evaluated using the pair-wise DSC (Appendix E.4).

This thesis also primarily centres on experimental work conducted in phantoms, including a lung phantom (Chapter 2) and thyroid phantom (Chapter 5). While these phantoms were useful models for US development, they represent a single subject with simplified anatomical features. They do not capture the full complexity of in vivo imaging, including patient-specific anatomical variability, tissue deformation, motion artefacts, acoustic shadowing, variable coupling, or pathological heterogeneity. As a result, model performance and workflow feasibility demonstrated in these settings may overestimate performance in clinical deployment. Furthermore, extending the segmentation to include additional structures, such as the trachea, oesophagus, muscles, and lymph nodes, may be beneficial for HIFU planning, but was not possible within the constraints of the phantom study [115]. Therefore, work towards clinical translation presents a major challenge for future work.

Similar data-related limitations apply to the free-text data extraction models developed in Chapter 3. All reports were collected from a single National Health Service (NHS) trust, with institution-specific reporting conventions, terminology, and formatting. As such, the wider generalisability of these models to other hospitals, reporting systems, or healthcare settings could not be assessed. In future, external validation across multiple institutions and scanner manufacturers will be necessary to determine robustness and identify domain adaptation requirements. In the RADEX study, 200 reports were used for validation and 202 for testing the regular expression models. Manual annotation was time-consuming given the large number of classes, and annotation of additional reports for testing was limited by the availability of clinical expert time for the study. As suggested in the paper, future studies could also consider comparing Radiology Data Extraction Tool (RADEX) with other methods, such as Large Language Models (LLMs), in which power calculations are used to determine sample size and formal hypothesis testing is used to assess the performance of using the tool relative to other techniques. This was not possible in the RADEX study due to access controls within the secure NHS environment.

Along this theme, the work presented in this thesis (Chapters 2, 4, and 5) used a single ultrasound system and one or two transducers (linear and curvilinear). Consequently, the model's robustness to manufacturer-specific beamforming, image post-processing pipelines, transducer geometries, and imaging presets was not evaluated, despite this being a known limitation of many ultrasound segmentation models [116]. In practice, a broad range of transducers and US systems are used, making it important for models to be robust to

device- and setting-related variability in order to ensure reliability across imaging environments. For example, in the LTUS clinical dataset, there are images from at least four US manufacturers (GE, Siemens, Philips, ATL) and a wide variety of imaging settings (frequency, dynamic range, acoustic output, power, grey map). Future work should therefore prioritise assessment of model performance across different systems and settings and consider domain generalisation strategies to improve robustness [116].

Finally, in Chapter 5 we demonstrated that 3D anatomical reconstruction enables more accurate thyroid volumetry than 2D approaches. However, determining whether this improves the safety and efficacy of HIFU compared to the current 2D planning approach remains unexplored. In terms of the full therapy pipeline, exposure placement and thermal modelling are important next steps, but remain out of scope of this thesis. Subsequent delivery of therapy, including automated robotic control, is also important but was unexplored in this thesis. Further, while segmentation models were demonstrated as live overlays on B-mode to aid interpretation, using the segmentation masks to inform closed-loop HIFU control remains unexplored and presents an avenue for future research.

6.3 Future directions

This thesis demonstrates advances in AI-integrated HIFU, particularly in tracked ultrasound acquisition, anatomical US segmentation, and 3D reconstruction for pre-treatment planning and real-time guidance. However, several areas remain unexplored and should form the basis of future research. This includes optimising exposure placement, thermal dose estimation, and closed-loop feedback for HIFU therapy delivery (Figure 6.1). The proposed tracked US system and treatment planning and visualisation tools set out the groundwork for these next steps. Future work should build on the tools presented here to address outstanding limitations, develop systems for delivering therapy, and investigate next steps to translate results into more clinically relevant settings.

6.3.1 Exposure placement

As discussed in Chapter 1, the complex planning and long treatment time for HIFU present major barriers to clinical uptake [49]. While co-robotic 3D US would reduce treatment planning time by eliminating the need for manual segmentation, improvements in exposure placement could further enhance the planning process. Current clinical approaches use raster-based methods to place rows of lesions that fully cover the tumour, before the operator manually selects and removes exposures that exceed the safety margins based on their interpretation of B-mode images [9]. Figure 6.2 shows a simple raster plan with planned HIFU exposures to treat a thyroid 3D reconstruction, demonstrating how future work could use 3D US volumes and visualisation techniques developed in this thesis to automate this process, accounting for thyroid margins and proximity to major vessels.

Future research could also consider approaches that reduce the number of exposures and

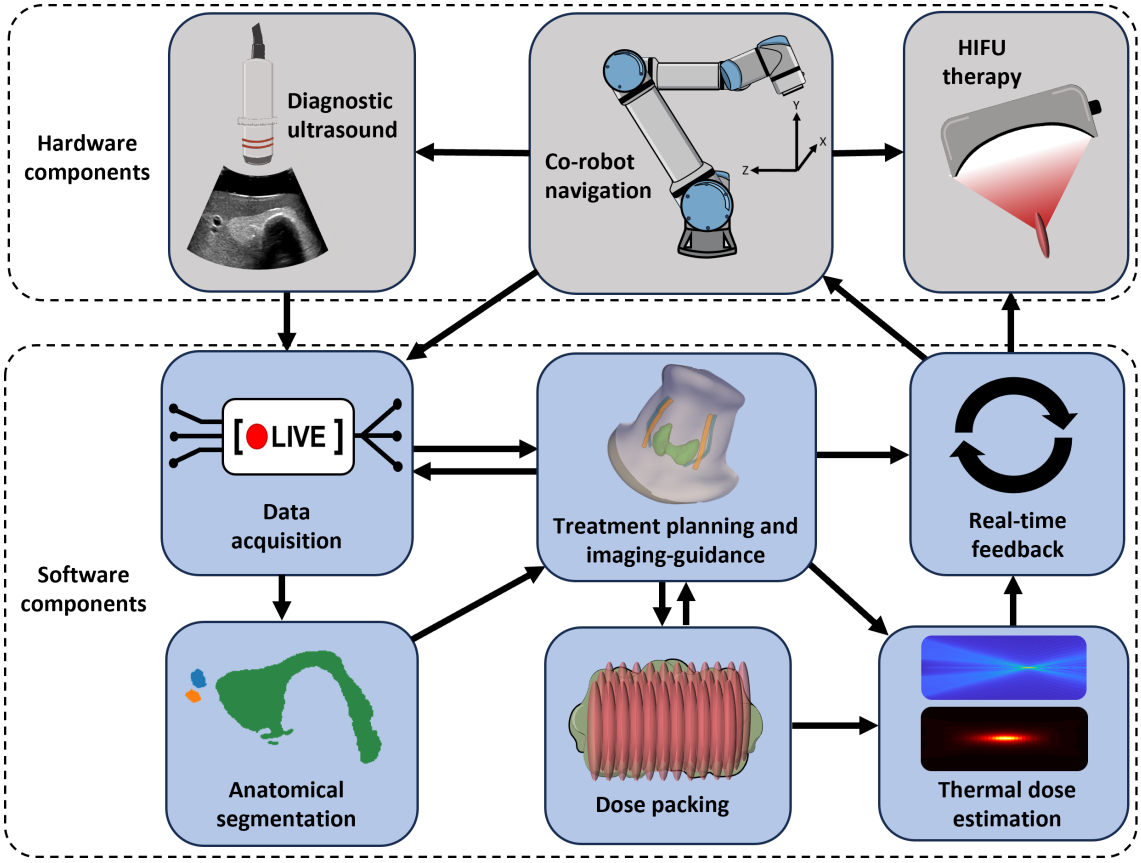


Figure 6.1: Schematic showing research avenues in AI-integrated USgHIFU, with links between software and hardware components. Data acquisition, anatomical segmentation, and treatment planning and imaging-guidance have been investigated in this thesis.

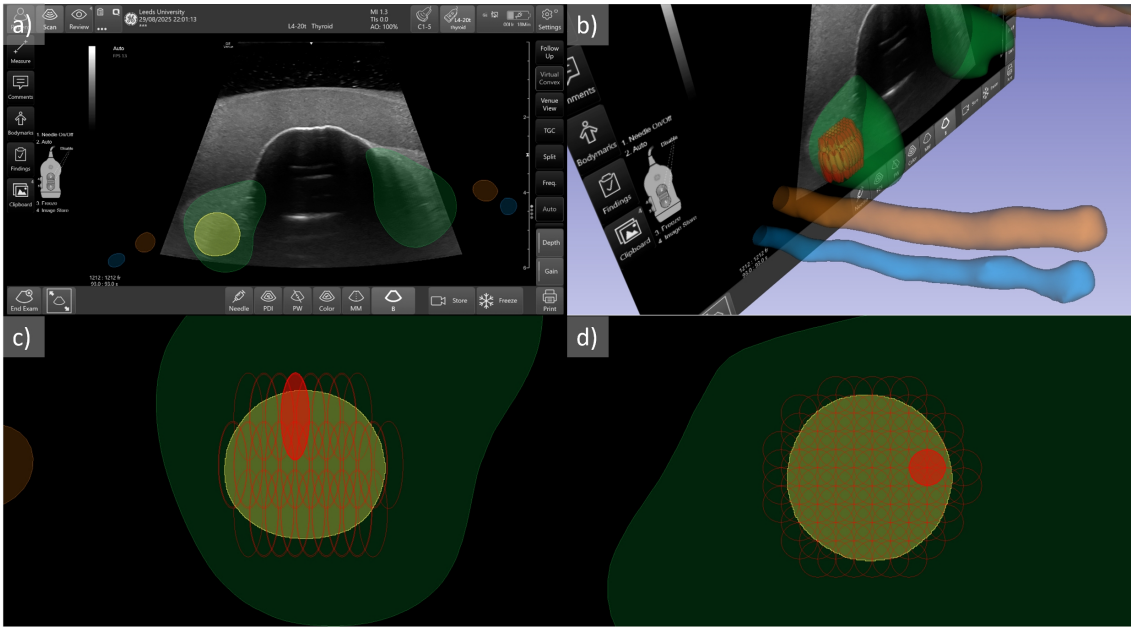


Figure 6.2: HIFU exposure placement (raster packing) for a thyroid nodule based on 3D reconstruction. The thyroid (green), carotid arteries (orange), and jugular veins (blue) were reconstructed using co-robotic 3D US with a thyroid phantom. The 'nodule' (yellow) was a reconstructed sphere from the US-QA-3D phantom, artificially added to the left thyroid lobe to demonstrate HIFU planning. 3DSlicer allows live visualisation of a) the US imaging with a segmented overlay, b) 3D view of the anatomy and planned treatment volume, c) transverse view, and d) coronal view of the planned exposures (red outlines) and first exposure (red solid).

the cooling time between HIFU exposures [117], using the 3D segmented volumes as input for planning. At the moment, the 'on-beam' time for a 3 cm thyroid nodule ranges from 30 to 45 minutes, almost twice as long compared with other established ablation techniques such as RFA [51]. The treatment time could be reduced by increasing treatment power or shortening the cooling period between treatment pulses [49], potentially making the treatment of larger nodules and multinodular goitre feasible in a single session.

In the existing Theraclion Beammotion planning software, minimisation of cooling time is considered [8, 42], but therapy planning is restricted to a plane perpendicular to the nodule, making it impossible to treat some areas such as the retrosternal part of the nodule [8]. Most studies also only treat a single layer along the nodule midplane, which results in undertreatment, especially in nodules with a larger anteroposterior diameter [11, 42]. In the future, combining 3D planning with Deep Reinforcement Learning (DRL) methods could optimise exposure placement and ordering to achieve more accurate tumour ablation in minimal time [118].

6.3.2 Thermal dose estimation

Using the spatial information provided by DL-segmented ultrasound images also enables further integration of AI into the treatment planning pathway, including acoustic and thermal modelling. Segmentation masks could be used to automatically determine path lengths across different tissue types, which affect ultrasound propagation [119]. Further, the positions of vessels relative to the treatment site could be determined. This is important because vessels act as heat sinks, drawing heat away from the treatment area and increasing the energy required to ablate the tissue [120].

Thermal dose estimation is a challenging area for future research, as solving the non-linear ultrasound propagation models needed for HIFU modelling is computationally expensive and existing techniques are not yet fast enough for planning or treatment monitoring within reasonable timeframes [121]. Other approaches, such as Physics Informed Neural Networks (PINNs), have been proven for fast computation of problems governed by partial differential equations [122]. Using these feed-forward neural networks to model acoustic propagation and subsequent bio-heat transfer in tissue may enable measurement of the cumulative thermal dose to plan or monitor ablation [123]. However, these models often lack generalisability, and they have not yet been explored for HIFU modelling [122].

In the future, these models could be integrated with treatment planning and guidance systems, such as the one presented in this thesis, to enable improved temperature-monitoring overlays in USgHIFU.

6.3.3 Robotic-guided ablation

Once the treatment plan is in place and approved by the clinical operator, the robotic system positions the HIFU transducer so that the focal point aligns with the location

specified in the treatment plan (Figure 6.3). The exposure is then delivered, and the robot moves to the next HIFU exposure locations until the treatment is complete. The co-robotic image-guided system developed in Chapters 4 and 5 supports a Sonics Concepts 1.1 MHz H-102 HIFU transducer with a central opening to allow coaxially aligned B-mode imaging, but further development and testing would be needed to evaluate the precision of therapy, as well as address challenges such as reliable coupling, response to movement, and tissue deformation [94, 124].

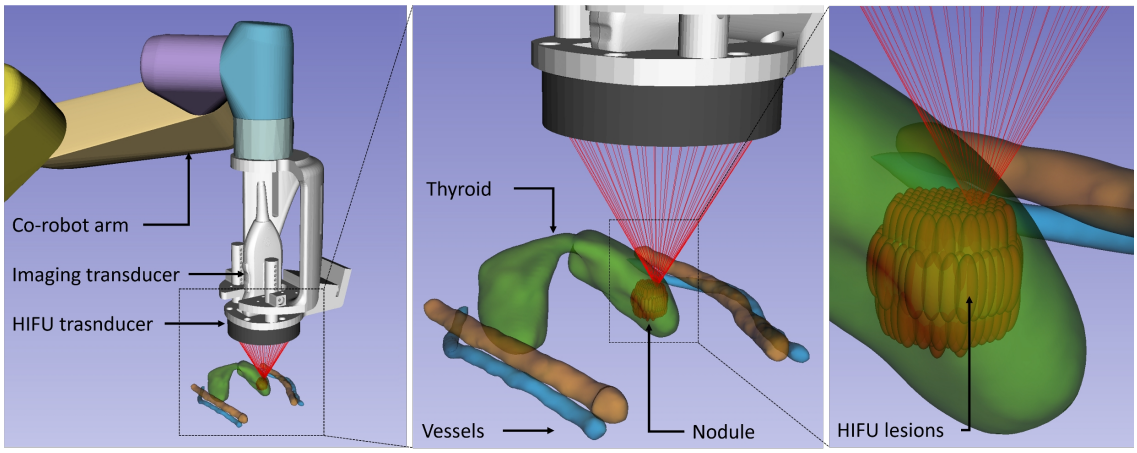


Figure 6.3: Robotic HIFU ablation based on 3D US phantom anatomy, showing the imaging and therapy assembly, thyroid and vessels, focal therapy beam, and treatment of a thyroid nodule with lesion locations.

Future work should conduct ablation experiments with anthropomorphic phantoms that simulate real anatomy as closely as possible, such as that proposed by Boers *et al.*, which was based on MRI data and contained a model trachea, thyroid, thyroid nodules, carotid arteries, internal jugular veins, recurrent laryngeal nerves, and vagal nerves [125]. This phantom also used a continuous-flow pump with 37°C water to mimic the heat-sink effect around vessels. 3D Reconstruction should be used to identify the treatment volume and calculate the treatment plan, allowing comparison with current approaches. Phantom materials such as polyacrylamide gels with thermochromic ink allow for measurement of the thermal ablation coverage within the nodule [126], or using fibre Bragg grating sensors or MRI thermometry to measure temperature more precisely [125].

6.3.4 Patient motion and real-time feedback

Patient motion is a challenge in both 3D ultrasound reconstruction and HIFU treatment planning, as even sub-millimetre movements can introduce registration errors. Respiratory motion and other involuntary motions introduce displacement errors between tracked frames, which can result in reconstruction artefacts such as blurring, geometric distortion, and misregistration between the imaging and planning coordinate systems. In HIFU, this could cause off-target heating or incomplete ablation. Organ movement is particularly important in treating intra-abdominal tumours, such as those in the liver or kidney,

and active motion compensation requires an understanding of robot and tissue dynamics, patient-dependent organ motion and characterisation, and tracking and feedback of these systems [1].

In clinical trials, general anaesthesia and ventilator-controlled apnoea have been widely used to control organ motion during treatment and ensure accurate targeting and procedural safety [127–130]. While this is effective, general anaesthesia is associated with increased procedural complexity, longer treatment times, and additional clinical risk [128]. In organs less susceptible to motion (such as the thyroid and prostate), sedation and physical immobilisation may be suitable, especially if the treatment is to be performed on an outpatient basis [1]. In these cases, computational motion-correction techniques such as respiratory gating (rejecting frames by breathing phase), image-based registration (deformable alignment of sequential ultrasound volumes), and deformable modelling of soft-tissue motion could help to maintain spatial consistency across acquisitions [131–133].

Closed-loop feedback based on real-time segmentation would represent a significant advancement over conventional open-loop USgHIFU workflows, enabling the system to adapt to motion during scanning. Previous approaches have incorporated feedback mechanisms based on detecting changes in tissue echogenicity [134] and on patient motion [135], but segmentation-based feedback remains largely unexplored [136]. For example, Groen *et al.* showed that monitoring tissue echogenicity can detect increases in backscatter resulting from cavitation or boiling, detecting potentially unsafe conditions and triggering power reduction [137]. In HIFU for thyroid nodules, laser-gated motion detectors have been used to trigger immediate power interruption when the patient moves or swallows during ablation, improving system safety [49].

Real-time anatomical segmentation of live ultrasound data could further improve safety by allowing the system to maintain up-to-date registration with the planning scene [138]. This would allow continuous verification of the alignment between the planned focal point and the actual target location, and small deviations arising from tissue deformation, probe pressure, or patient motion could be quantified and corrected through micro-adjustments in probe pose or electronic beam steering, potentially improving the accuracy and consistency of energy delivery.

Live segmentation of US-guidance would also allow the system to enforce safety, preventing unsafe delivery of therapy if the focal point is outside the treatment area [136]. Image-gated therapy could adapt continuously if the anatomy moves, reducing the risk of off-target heating and unintended tissue damage [135].

Segmentation-driven feedback could also gate therapy delivery based on the quality of US-guidance images and the presence of artefacts. For example, if coupling is poor or bone or air prevents visualisation of the treatment target on B-mode, the ablation is also unlikely to succeed [3]. These approaches could improve HIFU's safety and efficacy in the future.

6.3.5 Dataset curation

A primary barrier to the development of image-guided technologies is the availability of suitable clinical datasets. As discussed in Chapter 2 in the context of LUS, large datasets exist, but are often limited by poor imaging quality, incomplete annotations, or restricted access [139]. Similar challenges are presented in thyroid US. Although several datasets are available [140], many lack the image diversity and annotation detail needed to train generalisable models suitable for image-guided interventions such as HIFU.

Chapter 3 proposes a new UK dataset as a first step towards addressing these limitations. This dataset, called the Leeds Thyroid Ultrasound Dataset (LTUS), will provide a general-purpose foundation dataset for thyroid ultrasound research and contains 40,882 images from 1,650 patients across 5 centres. Current work towards this goal has included the extraction and linkage of anonymised US imaging studies with radiology report data, Fine Needle Aspiration (FNA) results, biopsy results, and pathology results (Appendix E). This linked data provides the necessary context for data annotation and offers a more accurate ground truth for thyroid nodule pathology than the nodule gradings in previous datasets. It also enabled curation of a dataset with varied pathologies and reduced class imbalance, which is important for training robust and generalisable AI models.

However, annotation remains a significant challenge. High-quality segmentation of thyroid structures and associated anatomy is time-intensive and requires clinical expertise and an understanding of US. While increasing the number of annotators or adopting crowdsourcing approaches could accelerate dataset development, these approaches introduce inter-annotator variability. The inter-annotator study presented in the Appendix E.4 provides an initial assessment of both annotation time and variability, highlighting the trade-off between scalability and consistency. These findings suggest that while multi-annotator strategies are feasible, they require careful quality control, potentially through consensus annotations or probabilistic label fusion techniques.

6.3.6 Clinical translation

Perhaps the most important challenge for future work is translating the methods and findings of this thesis from phantom experiments to clinical applications. While the phantom studies provide a necessary baseline for developing and validating segmentation and 3D reconstruction systems, they obscure the real-world challenges seen in clinical studies [94, 124].

Practical considerations for ultrasound-guided HIFU systems, such as workspace limitations imposed by the HIFU transducer and the effects of the coupling bag on imaging quality, would also need to be addressed. Feasibility studies should assess whether the transducer assembly's larger footprint limits the available range of motion for ultrasound scanning, particularly when imaging anatomically constrained regions for 3D reconstruction. Furthermore, maintaining adequate acoustic coupling between the imaging and ther-

apy transducers imposes additional constraints on probe orientation and access, and the robot arm may cause collisions or further limit the range of motion. Future studies should consider the mechanical design of combined imaging-therapy systems to account for these workspace limitations, ensuring sufficient imaging coverage and clinical usability.

Training an AI model for real-time clinical US segmentation would pose additional challenges, particularly regarding accuracy across diverse pathologies and anatomical variability. Models must also be robust to differences in tissue appearance, imaging artefacts, and transducer orientation. Recent advances, such as transformer-based segmentation [141] and hybrid feature learning [142], offer potential solutions by enabling models to learn high-level transferable feature representations. As shown in Chapter 2, transfer learning provides a method for refining existing models for new clinical tasks. Additionally, semi-supervised and weakly supervised learning approaches may be useful when only partially labelled data is available. However, care must be taken to avoid amplifying label noise, which may be challenging in US segmentation. Finally, the requirement for real-time inference constrains model complexity, requiring a trade-off between accuracy and efficiency.

Once a segmentation model is established, the next step is to conduct a feasibility study with healthy volunteers. Reconstructed 3D ultrasound volumes could be compared with MRI using the DSC and Hausdorff Distance (HD) to assess volumetric agreement, as in Chapter 5. This study would provide insight into the proposed system's real-world performance under clinically relevant conditions. System-level evaluation would also be useful to assess whether co-robotic assistance enhances or inhibits the operator's natural scanning technique, how this is affected by the operator's level of experience, and whether it improves accuracy and consistency compared to the existing approach. In the longer term, extending trials to patients with thyroid pathologies would also be necessary, although ethics approval would be required.

Chapter 7

Conclusions

HIFU has long been recognised as a technology with the potential to transform patient care. As the only non-invasive and non-ionising treatment modality, it offers clear advantages over conventional surgical and other interventional radiology approaches. However, despite decades of development, technical and workflow challenges have limited its widespread clinical adoption. Recent advances in AI, imaging guidance, and robotic control are beginning to address these challenges.

This project contributes to this by presenting an AI-integrated system for USgHIFU that incorporates DL-based segmentation, multimodal tracking, and real-time 3D reconstruction. This enabled improved volumetric assessment and has the potential to enhance procedural efficiency, reduce treatment times, and increase treatment confidence.

Clinically, benign thyroid nodules represent a promising short-term target for HIFU, offering clear patient benefits compared to surgery. More broadly, ongoing work in the liver, kidney, pancreas, and other challenging sites demonstrates HIFU's true potential to treat deep-seated or otherwise inoperable disease. Trials such as CAIN highlight HIFU's ability to provide life-extending or even life-saving options where conventional treatments are limited.

Ultimately, USgHIFU represents part of a paradigm shift towards non-invasive precision medicine. As Professor Wah often says in interviews, histotripsy is a "Star Wars–era" technology [143], but HIFU is not fictional and holds very real potential to improve lives. Developments in DL–based segmentation, treatment planning systems, and robotic control algorithms will help achieve this vision: to establish HIFU as a mainstream therapeutic option for patients of the future.

Appendix A

Appendix for Chapter 2: AI for real-time ultrasound segmentation

A.1 Comparison with the state-of-the-art

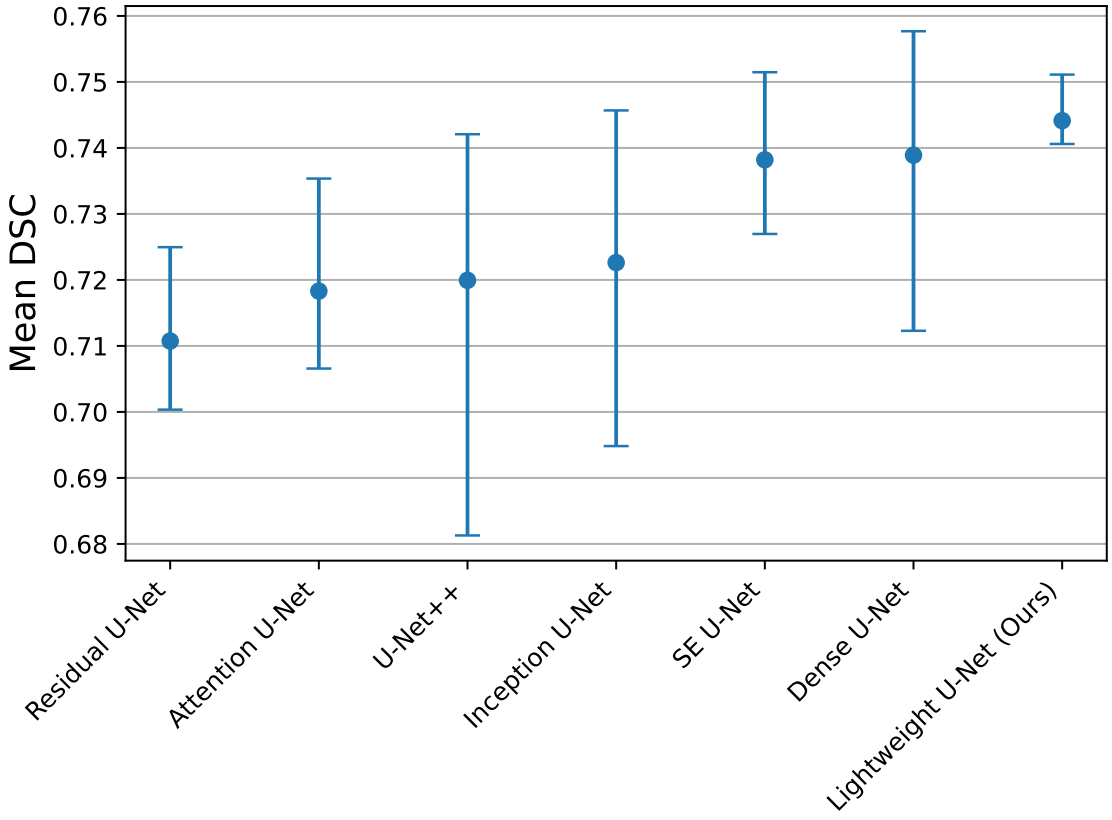


Figure A.1: Comparison of the Dice similarity coefficient for various U-Net variants, trained on the phantom dataset (450 images). Error bars show the best and worst scores over five training repeats.

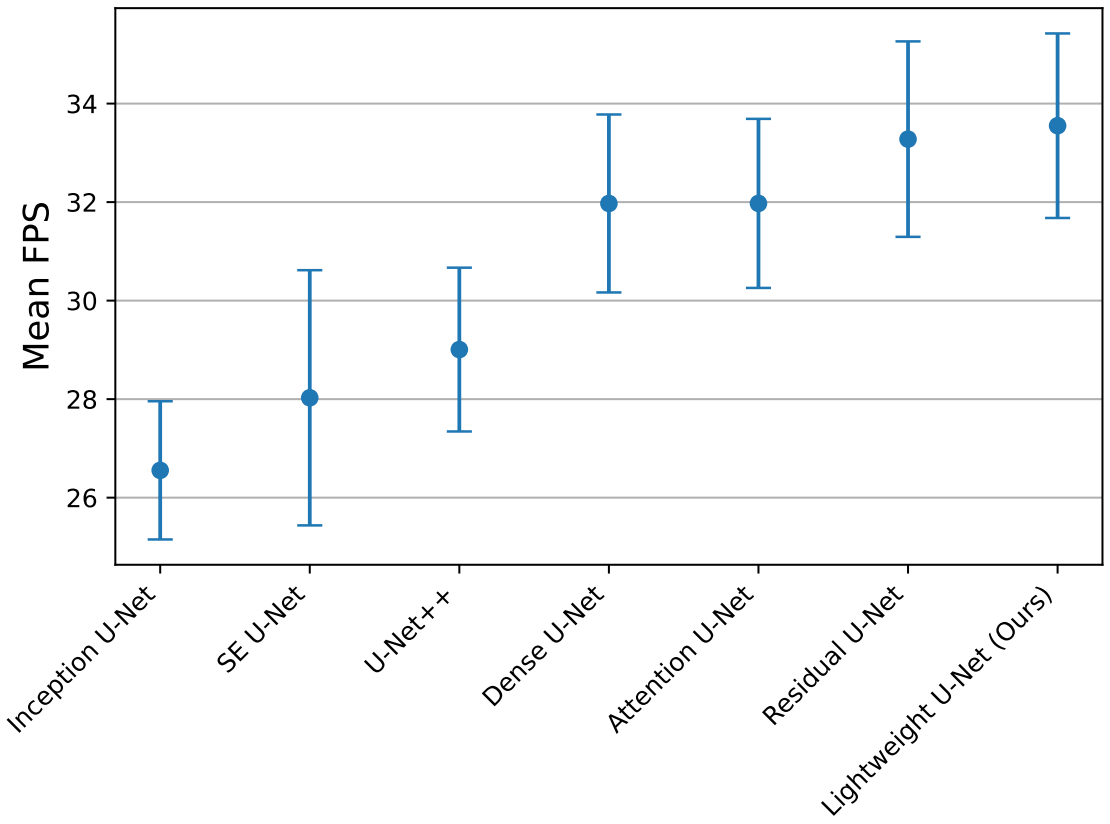


Figure A.2: Comparison of the benchmark framerate for inference of various U-Net variants, trained on the phantom dataset (450 images). Error bars show the best and worst scores over five training repeats.

Appendix B

Appendix for Chapter 4: Ultrasound 3D reconstruction and QA

B.1 Supplementary material

B.1.1 US-QA-3D phantom fabrication

Silicone moulds of reference objects were created by casting negative impressions of wooden masters. Moulds were made of platinum-cure silicone (Smooth-On, Dragon Skin 10 Medium) in 3D-printed moulds with pour and vent channels (Fig. B.1a). These were then filled with an agar mixture with no acoustic scatterers to form solid inclusions which appear anechoic on B-mode US. These objects were removed from the moulds, the sprues trimmed, and the objects were measured to obtain ground-truth dimensions (Fig. B.1b). Four reference objects were created with varying geometric and symmetry properties: a sphere (Fig. B.1c), an ellipsoid (spheroid) (Fig. B.1d), a right circular cylinder (Fig. B.1e), and an equilateral triangular prism (Fig. B.1f). The inclusions were then set in a larger agar block, this time adding glass beads to introduce speckle noise in B-mode images (mimicking heterogeneous tissue), thereby creating contrast with the anechoic inclusions. A flat layer approximately 3 cm deep (one-third of the agar) was poured, allowed to cool, and mostly solidify, while the remainder was kept warm. The anechoic inclusions were then placed on top of the first layer, and a second layer was poured to surround and cover the inclusions (remaining two-thirds). Once fully set, the phantom was vacuum-sealed for longevity, and the appearance of inclusions was confirmed on B-mode. The recipe for making the agar mixtures is below and can be scaled to the required volume. STL files for 3D-printed parts are available at https://github.com/SMRTUS/US-QA-3D.

For ∼250ml of agar:

- 250 mL deionised water,
- 12.5 g agar-agar,

- 25 mL glycerol ≥99% (increases speed of sound),
- 2.5 mL phenoxyethanol EHG (preservative),
- *10 g glass beads (soda lime glass microspheres, 3-6 μm diameter, introduces heterogeneity/speckle).

1. Transfer deionised water into a Büchner flask and heat to 70–75 °C.
2. Slowly add agar-agar, stirring continuously to prevent clumping. Add the glycerol. *Optionally add the glass beads for a tissue-mimicking material, exclude for anechoic material.
3. Apply a vacuum and degass the mixture for 20–30 min.
4. Slowly release the vacuum, then carefully pour into the mould(s). Allow to cool and set at room temperature.

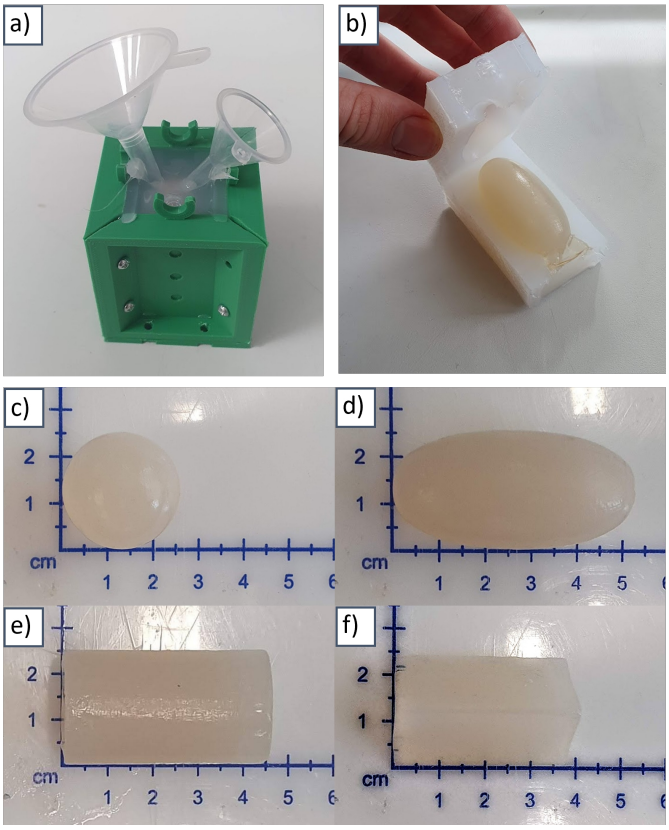


Figure B.1: US-QA-3D phantom construction showing a) fabrication of silicone mould of reference object in 3D-printed container, b) silicone mould with agar inclusion (before trimming the sprue), c-f) agar sphere, ellipsoid, cylinder, and triangular prism reference objects.

Appendix C

Ultrasound spatial and temporal calibration

Ultrasound spatial calibration was used to determine the ImageToProbe transform between the image coordinate system and probe coordinate system (Figure C.1). Fiducial registration techniques have been developed to standardise this calibration, including the use of multiple N-wire phantoms with wire targets arranged in a known geometric configuration [144]. Finding this transform is nontrivial, and several calibration steps are required to ensure that tracked components and reference geometries are accurately defined within a common coordinate system. The fCal module within the PLUS toolkit was used for calibration, and the main steps are outlined below [145].

For probe calibration, a tracking system (NDI Aurora EM field generator, and 3 tracking markers), water tank (degassed water), and tracked calibration objects (stylus, N-wire phantom, and ultrasound transducer) were used (Figure C.2a).

First, pivot calibration was used to find the StylusToStylusTip static transform between the tracking marker and the stylus tip. During pivot calibration, the stylus tip was held at a fixed point in space. At the same time, the handle was slowly rotated, and the tracking system captured the tracking marker's poses to determine the offset via a least-squares approach. The calibrated stylus was then used to capture accurate point-based measurements with the stylus tip.

Following this, landmark detection was used to find the PhantomToReference transform, which maps the phantom model and wire positions into the tracker coordinate system. Known registration points (defined in the N-wire Computer-Aided Design (CAD) coordinate system) were physically touched with the stylus tip, allowing their positions to be recorded in the tracker coordinate frame. The corresponding point sets in the phantom and tracker space were then used to compute a rigid registration, defining the spatial relationship between the known geometry of the N-wire targets and the global tracking

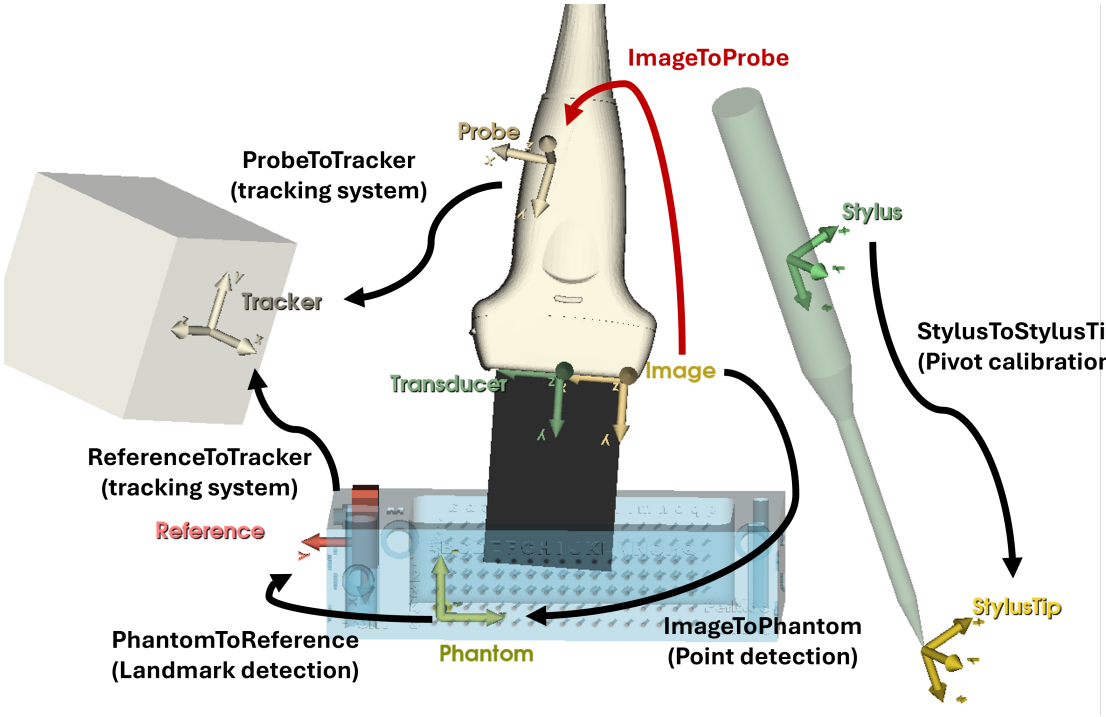


Figure C.1: Coordinate definitions for ultrasound spatial calibration. The ImageToProbe transform was calculated using the N-wire technique, in which wires in the phantom are imaged and detected to enable registration of the transducer, phantom, and image within a unified tracker coordinate system.

frame.

The transducer was then used to acquire a sequence of images at different transducer orientations, while simultaneously recording the corresponding tracker poses. On B-mode, the wire targets appear as bright point reflectors, allowing the positions to be automatically detected (Figure C.2b). Point correspondences between the image space and the known 3D wire geometry were calculated in fCal, and the rigid transform that best aligns these point sets was estimated using least-mean-squares optimisation. This allowed calculation of the ImageToProbe transform, allowing the transducer, image, and phantom to be registered in a common coordinate system (Figure C.2c).

Temporal calibration was needed to estimate the offset between data streams acquired by imaging and tracking devices. This was needed because each hardware device has an independent time reference (clock) that is not typically synchronised. The temporal calibration algorithm measures the motion of the reflection artefact from the bottom of the water tank in B-mode, imaged perpendicular to the transducer face. The bottom of the tank was imaged while moving the transducer in a continuous quasi-periodic motion. The normalised position (amplitude) was compared between the reflection in the image and the tracker to find the time offset that yields the highest correlation between the signals (Figure C.3) [145].

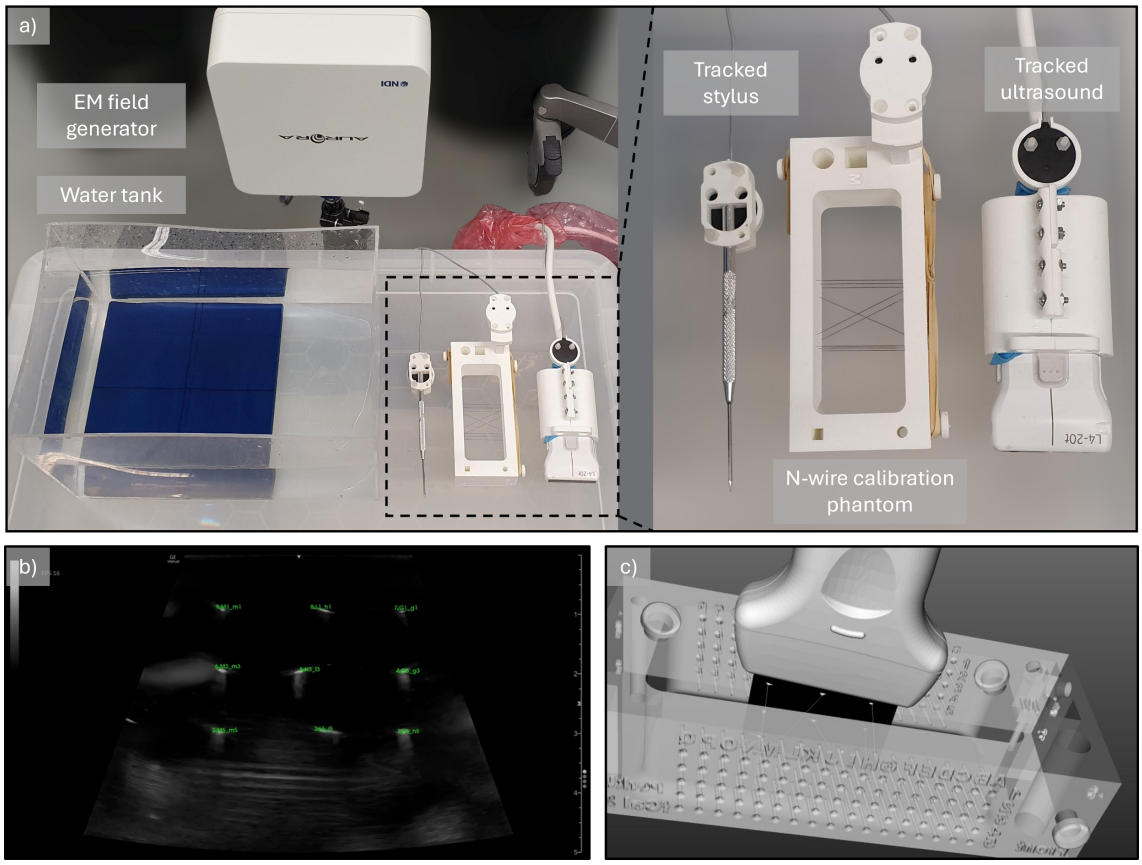


Figure C.2: Experimental setup for ultrasound spatial calibration using the N-wire phantom technique. a) EM-tracking system, water tank for imaging, and tracked tools for calibration (stylus, phantom, and ultrasound transducer). b) B-mode image of the wire targets, which appear as bright points and are detected by fCal. d) Calibrated system showing the image, transducer, and wire phantom in a common coordinate system.

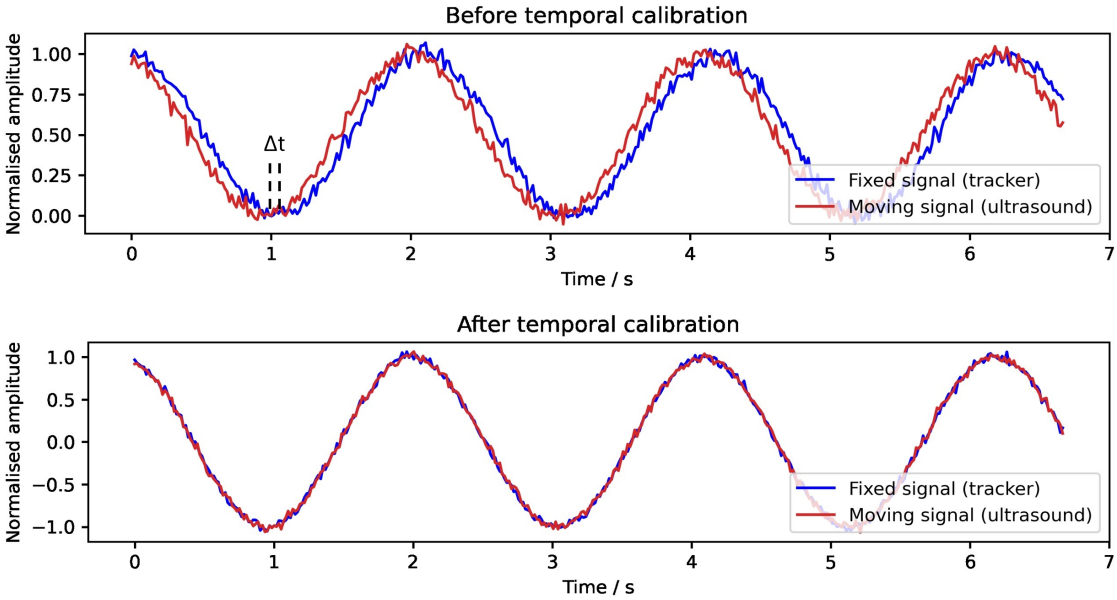


Figure C.3: Ultrasound temporal calibration between fixed and moving signals. The normalised position (amplitude) of the US image was captured alongside the tracked position over several cycles, allowing the offset between data streams (Δt) to be estimated.

C.1 Experimental setup for 3D reconstruction

The experimental setup for US-QA-3D experiments is shown in Figure C.4, including the robotic arm, water tank, ultrasound probe, and integrated tracking systems (robot, electromagnetic, and optical). This setup enabled repeatable, trajectory-based scanning of the geometric phantom under consistent imaging conditions.

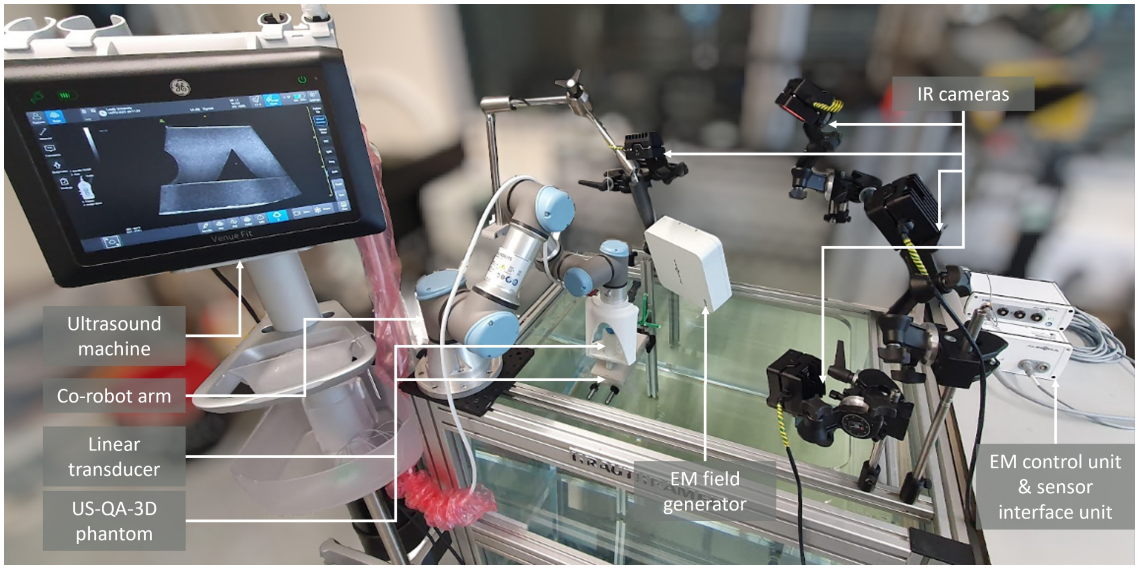


Figure C.4: Experimental setup for US-QA-3D experiments. The co-robot arm was positioned over the phantom, which was submerged in a tank of degassed water. Optical and EM tracking systems capture the transducer's position and orientation. The ultrasound machine was used to capture an imaging sweep across the phantom.

Appendix D

Thyroid ultrasound datasets

The development of accurate AI systems for thyroid US is fundamentally dependent on the availability of high-quality annotated datasets. These datasets are needed to train and evaluate models for tasks such as detection, segmentation, and classification. Detection aims to identify structures such as thyroid nodules that may otherwise be overlooked; segmentation enables precise delineation for radiomic analysis and volumetry; and classification supports malignancy risk stratification, informing whether FNA is needed.

Data-driven approaches must be robust and generalisable across varying patient populations and imaging conditions. Therefore, an effective thyroid US dataset should satisfy several key requirements:

1) The dataset should be sufficiently large and diverse, ideally incorporating multicentre data to capture variability in anatomy, pathology, ultrasound systems, acquisition settings, and image quality. The amount of data needed varies widely by application, with the best AI models trained on tens to hundreds of thousands of images. It has also been widely shown that clinical AI models often lack generalisability, a limitation that can be mitigated by using multicentre data and sampling to increase image diversity.

2) The dataset should contain high-quality annotations, with a clear quantification of uncertainty based on inter-operator agreement. Pixel-level segmentation masks provide the richest data in terms of object localisation, enabling a wider range of downstream applications compared to object detection or image-level classification, albeit at a higher annotation cost (Fig. D.1). Thyroid nodules and the thyroid gland must be annotated to enable assessment of focal and diffuse thyroid disease. Annotation of other structures, such as the major vessels, muscles, trachea, and oesophagus, is also useful for applications like treatment planning and guidance for minimally invasive procedures. If neck ultrasound images are included, lymph node annotations are also useful for biopsy and surgical planning.

3) The dataset should include contextual clinical information, radiological findings related

to diffuse and focal thyroid pathologies, nodule gradings, as well as labels from diagnostic tests. The incorporation of imaging metadata (e.g., probe orientation, laterality, and acquisition parameters) also improves the dataset's utility.

4) Datasets should use established standard formats (such as PASCAL VOC) and be publicly available to support reproducibility, benchmarking, and wider adoption within the research community.

Another important consideration is trade-off between time and cost for annotation vs dataset richness. In Fig. D.1, dataset richness increases when moving downwards and to the right, with multi-class anatomical segmentation providing the most detailed spatial and semantic information, and therefore enabling the widest range of downstream applications. Conversely, moving upwards or to the left reduces annotation complexity and burden. While simpler annotation types can generally be derived from more detailed ones (e.g. segmentation to detection to classification), the reverse is not generally possible without additional information.

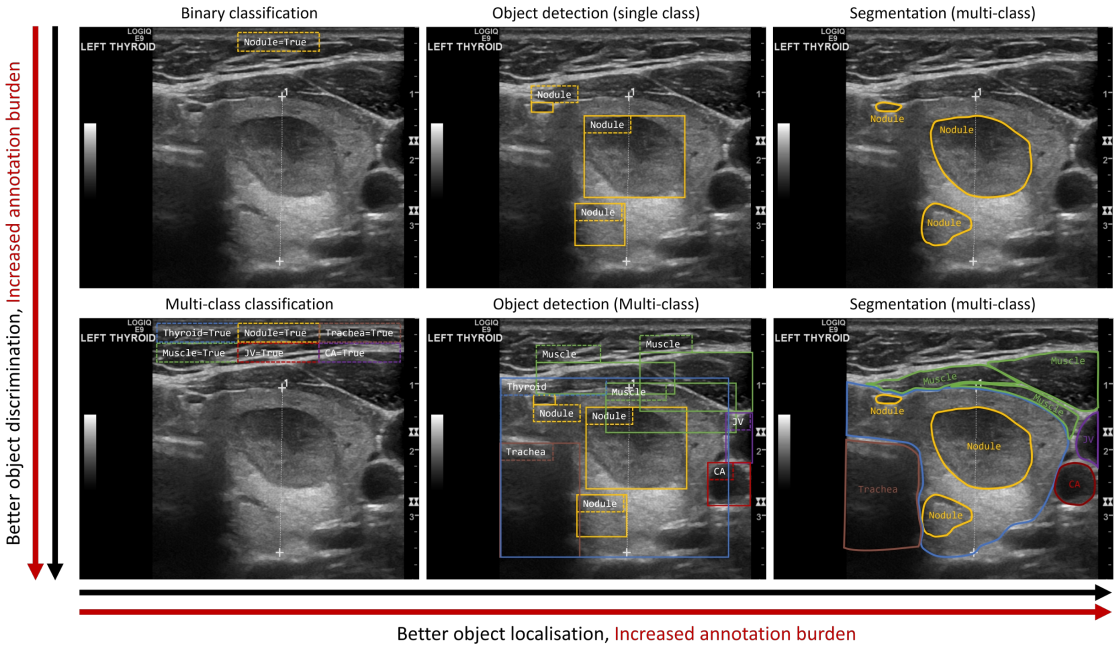


Figure D.1: Data annotation types for thyroid ultrasound. Improved discrimination and localisation contribute to richer data and enable more applications. As the number of classes increases, the annotation burden increases. Similarly, image-level classification has a lower annotation burden than object detection, which has a lower annotation burden than segmentation.

D.1 Overview of existing datasets

A number of thyroid US datasets have been proposed in the literature, each targeting specific tasks such as nodule segmentation, classification, or thyroid gland delineation

(Table D.1). However, no single dataset satisfies all the requirements for a multi-purpose foundation dataset in thyroid US (Table D.2). Early datasets such as DDTI [146] have been widely used to train computer-aided diagnosis systems, but they contain only a small number of images and lack pathology diversity, which can lead to overfitting and poor generalisability in models. TN3K [147] is also widely used and has a larger number of images and cases, but it lacks diagnostic information. Similarly, TN-SCUI2020 [148] has been used in prior studies but is unfortunately not publicly available.

Table D.1: Overview of thyroid ultrasound datasets, showing the data types, number of cases, data points, and annotated anatomical features. The anatomy column determines the scope of annotated features: TG = thyroid gland, TN = thyroid nodule, TR = trachea, CA = carotid artery, JV = jugular vein, M = muscle, O = oesophagus. Dataset annotations still in progress†

Dataset	Data type	Cases	Examples	Anatomy
UltrasoundCases [149]	Images	91	518	TG, TN
DDTI [146]	Images	390	637	TN
TG3K [150]	Video	16	16	TG
Li et al. dataset [151]	Images	45,644	332,785	TN
TN-SCUI2020 [148]	Images	–	4554	TN
TN3K [147]	Images	2421	3493	TN
Stanford Cine-clip [152]	Video	167	192	TN
SegThy (1) [78]	Video	28	504	TG, (TR, CA, JV)†
SegThy (2) [78]	Video	186	1116	TG
ThyUS2Path [153]	Images	842	8508	TN
TN5000 [140]	Images	–	5000	TN
LTUS (Ours)	Images	40,882	1,650	TG, TN, TR, CA, JV, M, O†

Table D.2: Suitability of thyroid ultrasound datasets, showing the possible tasks Seg. = Segmentaion, Det. = (Object) Detection, Cla. = Classificaiton. *Datasets are not easily accessible or may be unavailable. ‡dataset is not yet publicly available, but it has a planned public release date.

Dataset	Seg.	Det.	Cla.	Multicentre	TI-RADS	FNA	Public
UltrasoundCases [149]	–	–	✓	–	–	–	✓
DDTI [146]	✓	–	✓	–	✓	✓	✓*
TG3K [150]	✓	–	–	–	–	–	✓
Li et al. dataset [151]	–	–	✓	✓	–	✓	–
TN-SCUI2020 [148]	✓	–	✓	✓	–	✓	–
TN3K [147]	✓	–	–	–	–	–	✓
Stanford Cine-clip [152]	✓	–	✓	–	✓	✓	✓
SegThy (1) [78]	✓	–	–	–	–	–	✓
SegThy (2) [78]	✓	–	–	✓	–	–	✓*
ThyUS2Path [153]	–	–	✓	–	–	✓	✓
TN5000 [140]	–	✓	✓	–	–	✓	✓
LTUS (Ours)	✓	✓	✓	✓	✓	✓	✓‡

More recent efforts have produced larger datasets. The TN5000 [140] dataset is one of the largest publicly available thyroid datasets and includes bounding-box annotations for nodules, with associated malignancy labels derived from FNA. However, it lacks segmen-

tation annotations and provides minimal contextual or metadata beyond binary nodule classification. The number of included cases is also unclear. A large dataset reported by Li et al. [151] is also frequently cited but is not publicly accessible for research purposes.

Several datasets include video data, such as TG3K [150], Stanford Cine-clip [152], and SegThy [78], which may enable 3D segmentation and reconstruction. However, among these datasets, only a small number of videos have been sparsely annotated (16 in TG3K, 28 in SegThy (part 1), and 192 in Stanford CINE-clip), likely due to the heavy annotation burden of video segmentation. On this theme, the SegThy (part 2) dataset may eventually contain more annotated videos, but work on the dataset has been ongoing since 2022 [78]. Among all publicly available datasets, only SegThy includes annotations of adjacent structures such as the trachea, carotid artery, and jugular vein.

Appendix E

The Leeds thyroid ultrasound dataset

To address the limitations of previous datasets, we introduce the LTUS dataset (Table E.1), designed as a general-purpose foundation dataset for thyroid ultrasound research. The dataset provides 40,882 images from 1,650 cases, across 5 centres, with accompanying patient-level radiological findings, nodule gradings (British Thyroid Association (BTA) and Thyroid Imaging Reporting and Data System (TI-RADS)), malignancy labels from FNA, Core Needle Biopsy (FNA), or post-operative histopathology, and image metadata (laterality, scanning angle, imaging parameters). This enables a variety of tasks, including thyroid nodule detection, segmentation, and classification for FNA assessment; thyroid segmentation for automated measurements, volumetry, and radioiodine therapy planning; and treatment planning and guidance for minimally invasive therapies and HIFU.

Table E.1: Summary of the Leeds Thyroid Ultrasound dataset, including the dataset size, image details, aggregated demographics, and a summary of image metadata.

Category	Summary
Dataset size	40,882 images, 1650 cases, 5 centres
Image format	DICOM, PNG
Image size	720×960 (median), 576×768–970×1552 (range)
Modalities	B-mode 79.8%, Doppler 20.2%
Manufacturers	GE, Siemens, Phillips, ATL
Demographics	Ages 18-89 (median 54), 76% women, 24% men.
Images per case	25 ± 10.6 (median ± SD)
Images by anatomy	Thyroid (12,553); salivary glands (4158); lymph nodes (lvl I–V: 1.6–2.2k)
Scan plane	Transverse 54.4%, Longitudinal 45.6%
Laterality	Left 45.7%, Right 50.3%, Bilateral 4.0%
Imaging frequency	11.0 MHz (30.1%), 9.0 MHz (25.5%), 13.0 MHz (24.1%), others <10%

The dataset includes ultrasound images retrieved from examinations tagged Ultrasound Neck (UNECK), Ultrasound thyroid (UTHYD) and Ultrasound thyroid and parathyroid (UTHPY) between 01/01/2015 and 31/12/2019 within Leeds Teaching Hospital Trust (LTHT), UK. To obtain the date, entries were first extracted from Computerised Radiology Information System (CRIS) using the clinical record interactive search, which returned

47,566 entries. These entries were merged by event identifier to combine multipart examinations into a single entry, yielding 22,793 unique events. Paediatric patients (Age <18) were excluded, further reducing the entries to 20,336, of which 16,246 unique patients were present (4,090 entries referred to repeat or follow-up examinations).

The RADEX tool was then used to select the cohort, extracting anonymised labels from free-text radiology reports for features such as the presence of isolated or multiple thyroid nodules, diffuse thyroid disease, thyroid enlargement, and a history of thyroid surgery [154]. This information was tabularised, providing the context needed to inform downstream annotation and enable patient-level classification. These features were also used for stratified sampling of the final dataset of 1,650 cases, with representative splits across diffuse and focal pathologies, history of thyroid surgery, and nodules graded BTA U2-U5.

For each patient in the sampled cohort, all images in the study were then extracted and de-identified, with pseudoidentifiers assigned to enable case-level aggregation. Alongside the main image dataset, nodule gradings and, where present, sample results from FNA, FNA, and post-surgical pathology are included to enable accurate determination of nodule malignancy. The composition of cases in the final dataset is shown in Table E.2.

Table E.2: Characterisation of the Leeds Thyroid Ultrasound dataset, including the distribution of labels, nodule gradings, and sample types.

Category		Value
Pathology distribution (cases)	Healthy	281
	Single nodule cases	600
	Multiple nodules	554
	Diffuse thyroid disease	751
	Post-thyroidectomy	328
	Abnormal lymph nodes	136
BTA	U2	32.7%
	U3	32.6%
	U4	27.4%
	U5	7.3%
TI-RADS	TR1	23.6%
	TR2	33.0%
	TR3	12.8%
	TR4	22.9%
	TR5	7.7%
Malignancy	Cytology-confirmed (BTA U3-U5)	85%
	Benign (%)	68.9%
	Malignant (%)	31.1%
Sample types	FNA	542
	CNB	161
	Post-surgical pathology	415

E.1 Thyroid nodule grading

Thyroid nodule grading is assigned based on ultrasound assessment and indicates the risk of malignancy, informing the decision on whether to sample the nodule (FNA/FNA). In the LTUS dataset, nodule grading labels were extracted from the radiology reports at the patient level. BTA U gradings were extracted directly using RADEX, as this is the grading system used in the UK. In the BTA system, the grading is assigned based on the nodule's overall appearance, where U2 indicates benign features (such as cystic or spongiform appearance) and no FNA cytology is needed, U3 represents an equivocal/indeterminant nodule, U4 represents a suspicious nodule, and U5 represents a highly suspicious/malignant nodule, with FNA recommended for U3-U5 nodules [155].

In addition to the BTA grading system, several other grading systems are used internationally [156]. For example, the American College of Radiologists (ACR) TI-RADS system, used in the United States (U.S.), aims to standardise thyroid ultrasound reporting through a points-based approach, compared to the description-based BTA approach used in the UK. To allow comparison between the two grading systems, the LTUS dataset retrospectively assigned TI-RADS categories based on descriptive features of thyroid nodule composition, echogenicity, shape, margins, and echogenic foci.

The search strategy and ACR TI-RADS points are shown in Table E.2, where the sum of points from all categories determines the ACR TI-RADS level: 0 points = TR1 (Benign, No FNA), 1-2 points = TR2 (Benign, No FNA), 3 points = TR3 (Mildly Suspicious, FNA if $\geq$2.5 cm), 4-6 points = TR4 (Moderately Suspicious, FNA if $\geq$1.5 cm), and $\geq$7 points - TR5 (Highly Suspiciousm FNA if $\geq$1.0 cm [157].

The relation between paired BTA and TI-RADS gradings is shown in boxplots (Figure E.1. As expected, the median TI-RADS grading increases with increasing BTA category, indicating general agreement in overall risk stratification. However, the interquartile ranges and range bars indicate large variability, suggesting that each BTA grading is associated with a range of TI-RADS gradings.

E.2 Cytology/pathology

Cytology and pathology reports were also examined for all patients with one or more thyroid nodules. Manual chart review was conducted by two resident doctors, who identified relevant reports for all thyroid nodules sampled within $\pm$1 year of the imaging study. The sample type, laterality, diagnosis, and, in the case of cancers, the Tumour Nodes Metastasis (TNM) staging was recorded. All follow-up samples within the date range were included, since multiple samples were often taken. For example, if there were multiple suspicious nodules and prior samples were inadequate for diagnosis, or if a nodule was sampled using FNA, the malignancy was later confirmed on resection. Up to 7 samples were recorded, with a median of 2 samples for BTA U2-U5 thyroid nodules.

Table E.3: Sonographic features used for calculating TIRADS and the RADEX search terms used. Each feature has a corresponding number of points that contribute to the overall grading.

Sonographic features	RADEX search terms	Points
Composition		
Cystic	nodule∼∼2cystic	0
Spongiform	nodule∼∼2spongiform | micro?cyst*	0
Mixed solid-cystic	nodule∼∼2mix* | (nodule∼∼2cyst & nodule∼∼2solid) | solid component∼∼2nodule | cystic component∼∼2nodule | partly solid∼∼2nodule | part solid∼∼2nodule | part* cystic∼∼2nodule	1
Solid	nodule∼∼2solid EXCEPT 'Mixed solid-cystic'	2
Echogenicity		
Anechoic	nodule∼∼2anechoic	0
Hyperechoic or isoechoic	nodule∼∼2iso?echoic | nodule∼∼2hyper?echoic	1
Hypoechoic	nodule∼∼2hypo?echoic | echo?poor∼∼2nodule	2
Very hypoechoic	very∼∼2hypo?echoic | marked*∼∼2hypo?echoic | predominant*∼∼2hypo?echoic | extremely∼∼2hypo?echoic | hypo?echoic∼∼2muscle | marked*∼∼2echo?poor	3
Shape		
Wider than tall	wider∼∼2tall | oblate	0
Taller than wide	taller∼∼2wide | prolate	3
Margin		
Smooth	smooth∼∼2nodule | well?defined∼∼2nodule	0
Ill-defined	ill?defined | poorly?defined	0
Lobulated or irregular	lobulated | protru* | irreg* | jagged | sharp∼∼2margin? | spiculated	2
Extra-thyroidal extension	extra?thyroid* | extra?capsular | exten*∼∼2beyond | exten*∼∼2outside | exten*∼∼3capsule | infil* | inva*	3
Echogenic Foci		
None or comet-tail artefacts	no∼2foci | no foci | comet* | nodule ring?down | colloid*	0
Macrocalcification	macro?calcification?∼∼2nodule? | large∼∼2calcification? | coarse∼∼2calcification?	1
Peripheral (rim) calcifications	peripheral∼∼2calcification? | egg?shell | edge∼∼2calcification? | rim∼∼2calcification?	2
Punctate echogenic foci	punctate | echo?gen* foc* | echo?gen* fleck?	3

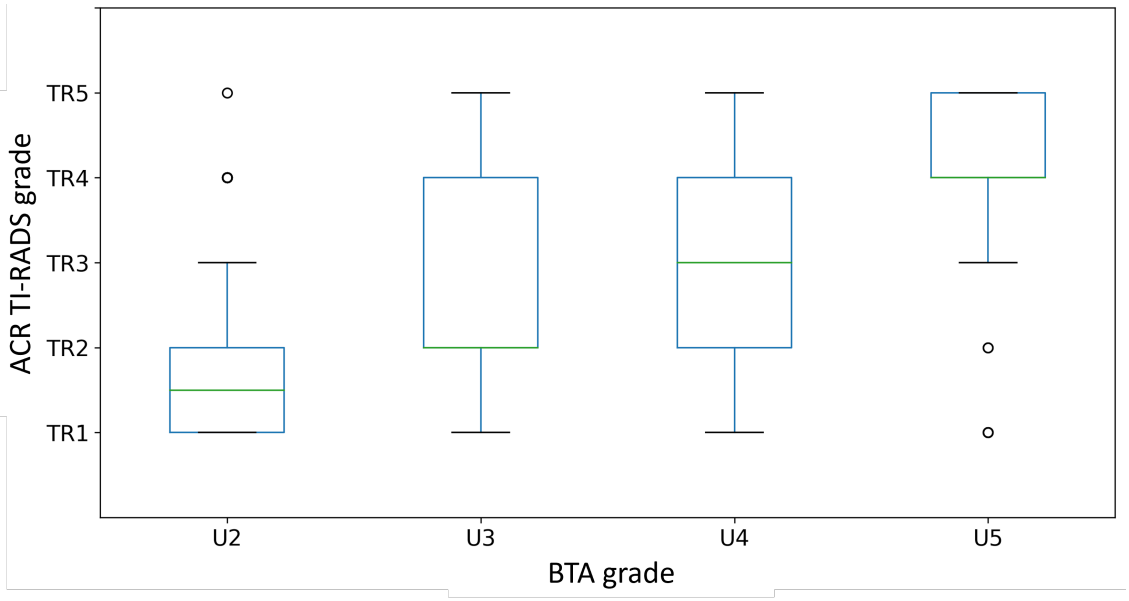


Figure E.1: Boxplot for BTA and ACR TI-RADS thyroid nodule gradings.

Of the three sample types, post-resection histopathology is the gold standard for determining malignancy, followed by FNA, then FNA in terms of sensitivity [158, 159]. This priority was therefore used to assign malignancy labels where multiple samples were present for the same nodule. This enables accurate determination of diagnostic labels for most thyroid nodules in the dataset (Figure E.2). As expected, U2 nodules had significantly fewer samples than U3-U5 and a low number of malignant nodules overall. U3-U5 nodules had an increasing percentage of malignant samples. This is expected as the BTA U systems have been shown to be highly sensitive in selecting malignant nodules for sampling [160].

Across repeat samples, a large number of diagnostic and non-diagnostic results were recorded (Figure E.3). In line with guidance, sample 1 was almost always FNA [161], with cytology 'Thy' categories recorded. These categorise the nodule sampled cells as non-diagnostic (Thy1), non-diagnostic cystic (Thy1c), benign (Thy2), benign cystic (Thy2c), atypical but indeterminate (Thy3a), indeterminate follicular (Thy3f), suspicious of cancer (Thy4), and malignant (Thy5) [161, 162]. However, a large number of results were non-diagnostic (Thy1) or indeterminate (Thy3a/Thy3f), requiring repeat sampling. Sample 2-6 included examples of benign and malignant disease from histopathology results from FNA or post-resection tissue. Benign diseases included Multinodular Goitre (MNG), colloid nodules, hyperplastic nodules, follicular adenomas, and hurthle cell adenomas. Malignant diagnoses included papillary carcinomas, follicular variant papillary thyroid carcinomas, follicular carcinomas, metastatic disease, non-invasive follicular thyroid neoplasms (NIFTP), hurthle cell carcinomas, and medullary carcinomas. Secondary findings of microcarcinoma were also present in several samples.

For histopathology results, TNM provides a framework for describing disease extent in

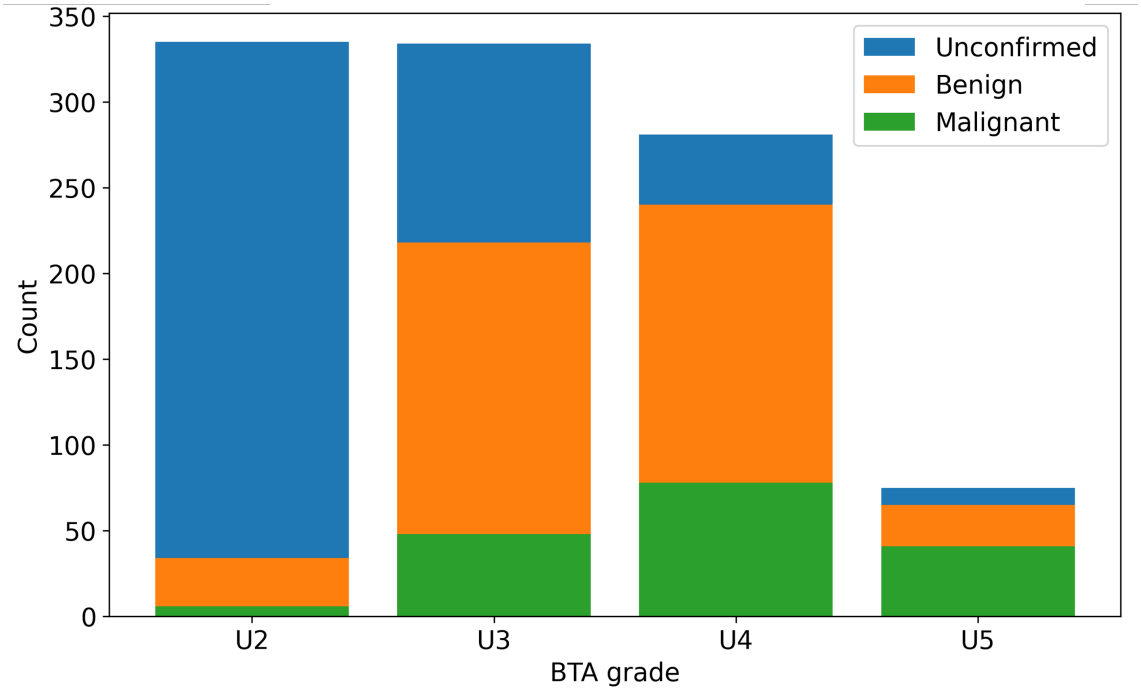


Figure E.2: Total number of nodules and sample-confirmed nodules grouped by BTA grade in the LTUS dataset.

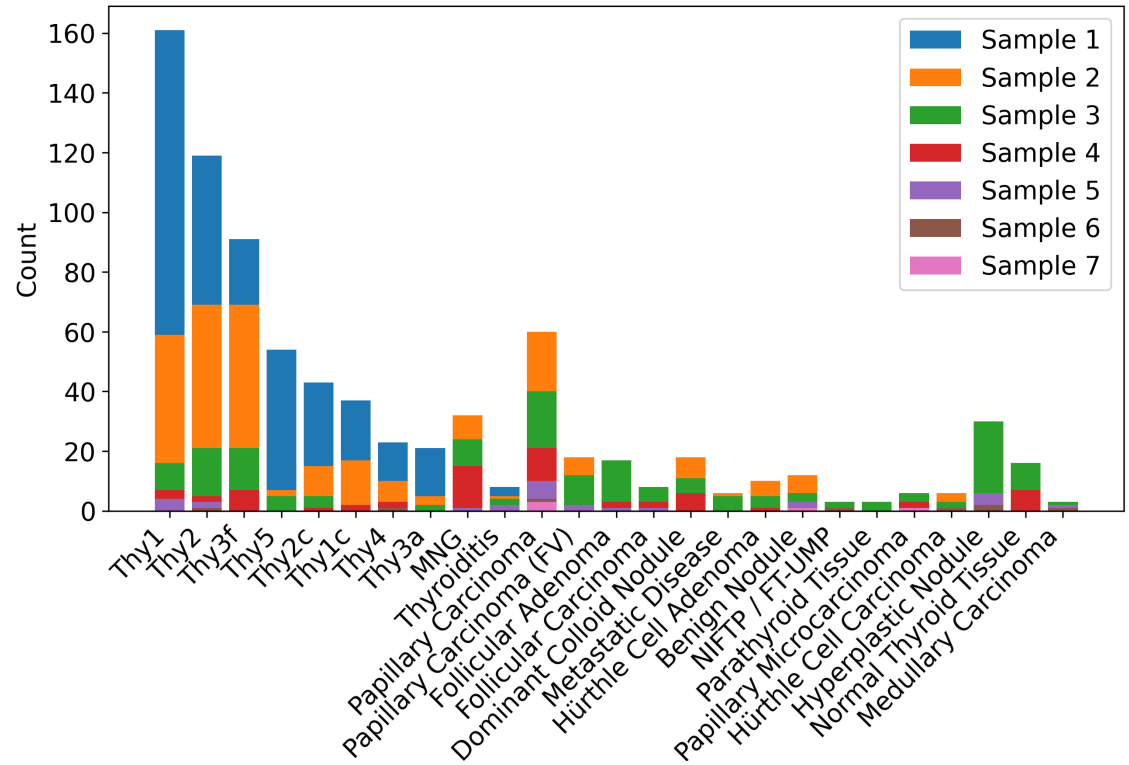


Figure E.3: Value counts for cytology and pathology labels for thyroid nodules in the LTUS dataset. Results are grouped by sample number, with increasing numbers indicating chronological follow-up or repeat samples.

thyroid cancer, based on primary tumour size and extent (pT), regional lymph node involvement (pN), and distant metastasis (M) [163]. TNM staging was also recorded, adding clinical value by contextualising disease severity and prognosis and enabling analysis of specific tumour stages. The frequency of TNM stagings included in the dataset is shown in Figure E.4.

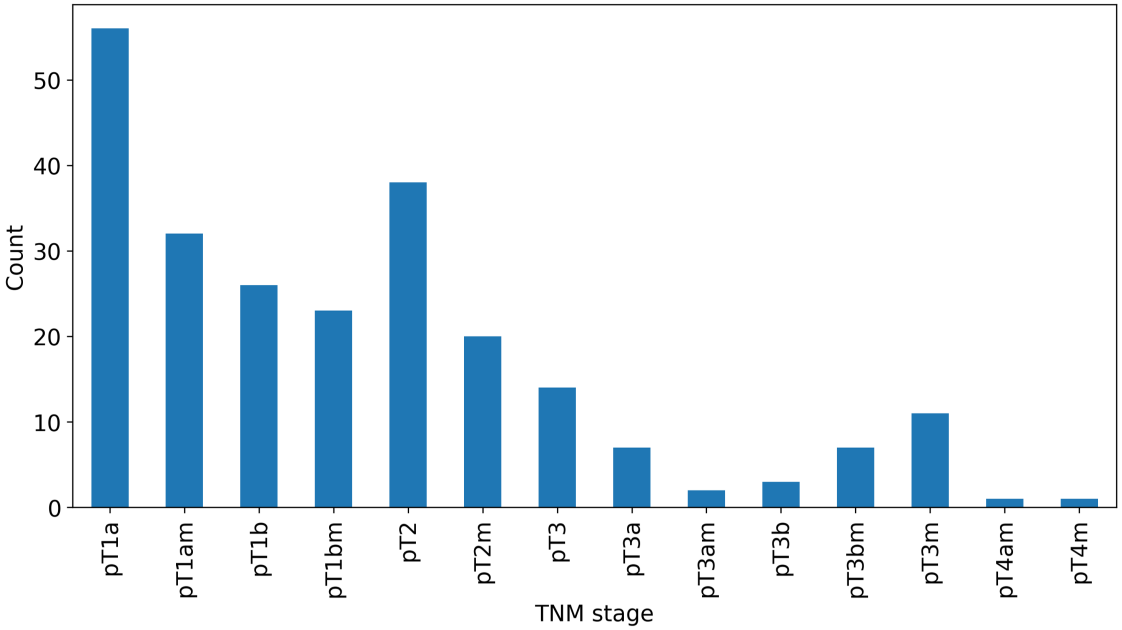


Figure E.4: TNM staging for malignant thyroid tumours in the LTUS dataset.

E.3 Imaging metadata

In addition to radiological findings and diagnostic information, the dataset includes imaging metadata to facilitate stratification by imaging settings. This is important for assessing generalisability across different imaging systems and settings, and for simplifying data-loading pipelines by providing information such as image dimensions upfront.

Optical Character Recognition (OCR) was used to extract the radiologist context label from the image (i.e. Thyroid Left LS), as well as US settings from the parameter box burned into the image itself. For each of the 23 unique image shapes in the dataset (representing images obtained from various US machines), bounding boxes were drawn around the context label and US settings. Binary thresholding was then used to separate the text from the background, thereby improving OCR accuracy in areas where text overlapped the B-mode image. Pytesseract with page segmentation mode (PSM) 6 was used to convert the text in the images to a machine-readable string, assuming a single, uniform block of English text [164].

The text was then cleaned to remove non-alphanumeric characters (except spaces, newlines, the per cent sign, forward slash, and periods) and extra whitespace. Additional post-

processing rules were then applied to standardise the text and exclude differences arising from minor OCR errors. Finally, a function split the context label into references of laterality (left, right, midline), transducer orientation (transverse, longitudinal), and image anatomy (thyroid, salivary glands, lymph node levels), the counts for which are shown in Figure E.5. A separate function extracted key-value pairs from the US settings, returning the setting name and corresponding value as a dictionary. This allowed the extraction of information such as the imaging frequency, gain, dynamic range, grey map, resolution, noise suppression, framerate, and speckle reduction. For example, the most common imaging frequencies were 11.0 MHz, 9.0 MHz, 13.0 MHz, and 15.0 MHz, which together made up 86% of the data.

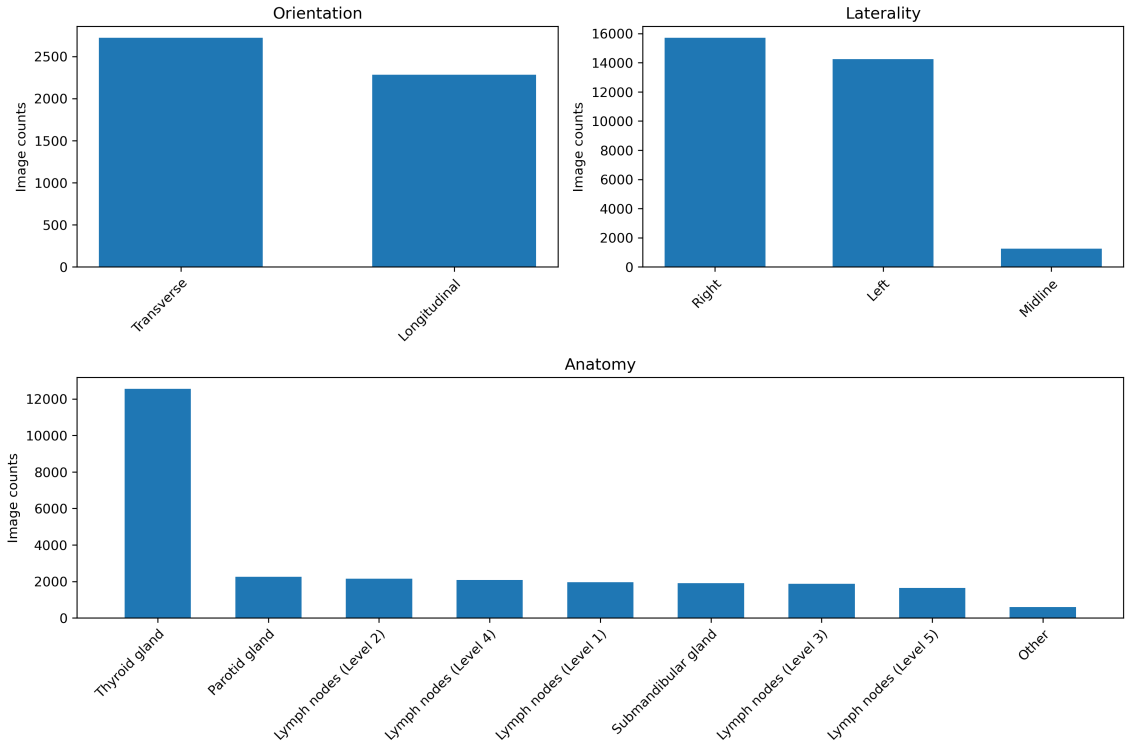


Figure E.5: Metadata labels for the LTUS dataset, showing the image orientation, laterality, and anatomy from OCR.

E.4 Inter-annotator study

To assess the cost and predicted uncertainty of annotation, an interannotator study was conducted with 7 annotators of varying experience levels. Two consultant-level experts, with >10 years of experience in head and neck ultrasound (IDs 1 and 2), two medium-experience resident-level clinicians with ∼2 years of experience in ultrasound (IDs 3 and 4), and three non-clinical researchers with mixed-level experience in ultrasound research (IDs 5, 6, and 7). This enabled an assessment of the level of experience required to precisely annotate thyroid ultrasound images, a largely unexplored topic in the literature.

Fifty images were sampled from the full dataset for the interannotator study. These images were selected as samples to represent the main pathological variations: ten with normal anatomy, ten with a single thyroid nodule, ten with multiple thyroid nodules, ten with a history of thyroid surgery, and ten with diffuse thyroid disease. Images were randomly selected based on the patient-level annotations and hosted for annotation using the tool LabelStudio. Patient-level classification labels were available to guide annotation, but only a single image per patient was annotated.

Seven classes were selected for annotation: thyroid gland, thyroid nodule, internal jugular vein, common carotid artery, trachea, oesophagus, and major superficial neck muscles (strap muscles and sternocleidomastoid). All seven annotators were given the same instructions on using the annotation tool and the same general guidelines for annotation:

- **Multiple objects:** Annotate all visible instances of each class (e.g., multiple nodules or muscles in the same image).
- **Overlapping objects:** Objects may overlap (e.g., thyroid and thyroid nodules). To create overlapping labels, hold Ctrl/Cmd while annotating.
- **Missing classes:** Not all classes will be present in every image. Only annotate objects that are visible.
- **Partially visible or obscured objects:** Annotate only the visible portion of the object. Do not estimate or extend boundaries beyond what is visible. If an object extends beyond the image border, trace the boundary up to the image edge where it exits the frame.
- **Unclear boundaries:** When boundaries are indistinct, annotate as accurately as possible using best judgment. Place points along the most probable object boundary, even if edges are poorly defined.
- **Annotation quality:** Aim for smooth, natural contours. Avoid sharp or jagged polygons; use sufficient points to closely follow the object boundary while maintaining smooth edges.
- **Review before submission:** Check for missed objects, accidental label overlap errors, incomplete polygons, or stray points prior to submission.

In addition, all annotators had access to training materials produced by a clinical expert that showed how to annotate the relevant anatomical features, with examples illustrating the annotations. All annotators independently labelled all 50 images, and the complete annotations were compared using the class-wise DSC to assess inter-annotator agreement. Metadata also included the lead time for each annotation and the number of control points used per image, offering insight into the annotation process. Example images (randomly selected) from the inter-annotator study are shown in Figure E.6.

Comparing these annotations using paired DSC provides a robust, intuitive assessment of

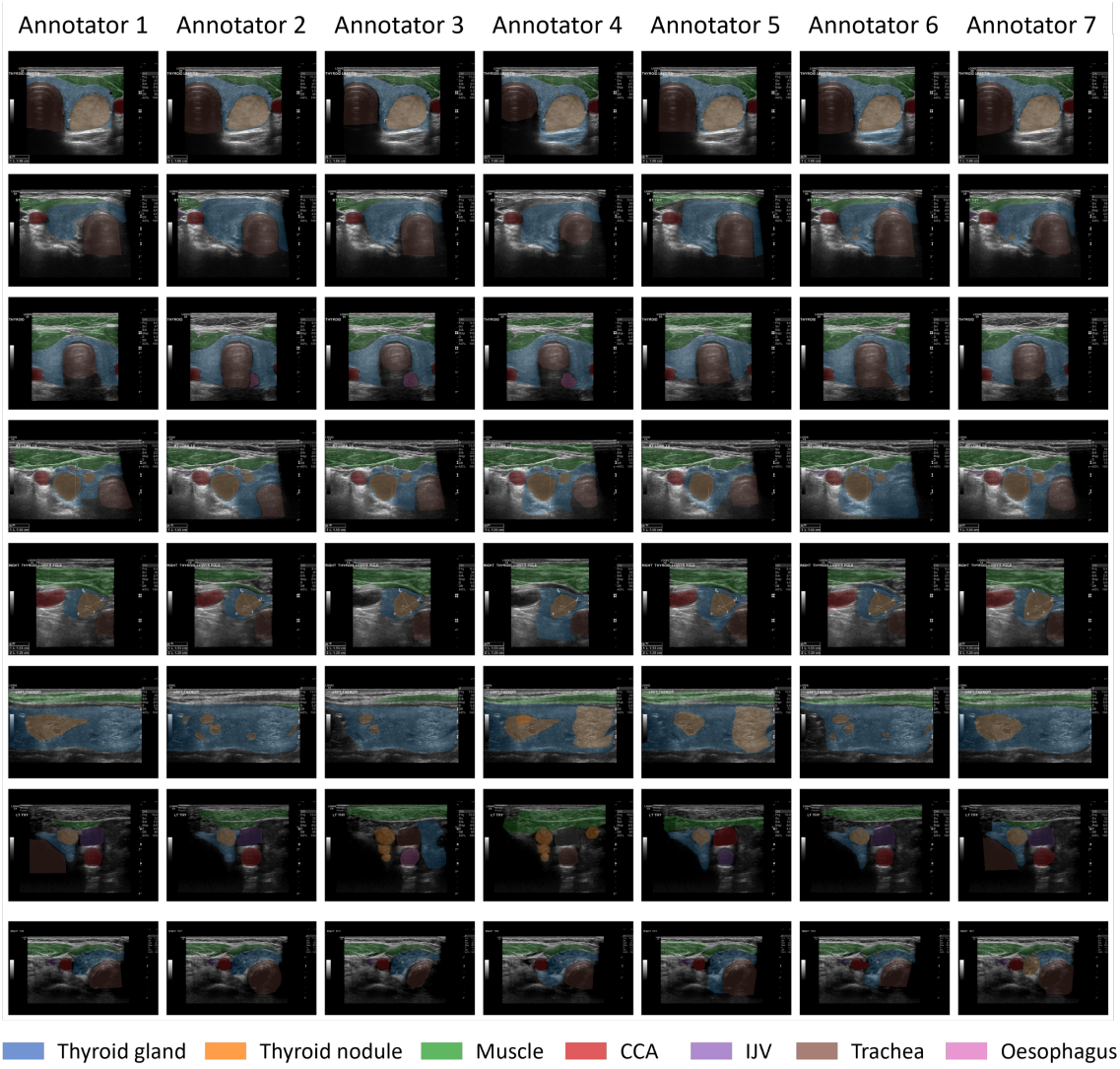


Figure E.6: Example segmented images from the inter-annotator study. Annotators 1 and 2 had high-level experience in thyroid ultrasound imaging; annotators 3 and 4 had mid-level experience; and annotators 5, 6, and 7 had lower-level experience.

the average pixel-wise agreement between the annotators for each class (and an overall average). (Figure E.7). In the thyroid and muscle classes, the paired DSC was generally high across all annotator pairs, with DSC values >0.7 indicating a good agreement. This suggests that all annotators in the study were able to annotate the thyroid and muscles with a high degree of agreement, especially the expert annotators (1 and 2), the medium-experience annotator 3, and the less-experienced annotator 7. In contrast, thyroid nodules were more difficult to annotate, as reflected in a lower average inter-annotator DSC. Agreement between the expert annotators was reasonable (DSC = 0.7), but generally poorer among the less experienced annotators, with moderate-to-poor agreement at medium and lower experience levels, as well as between these groups and the experienced annotators. The macro-average across all 7 classes showed reasonable paired agreement between annotators with the same experience levels, as well as overall. This suggests that while inexperienced annotators may be able to label some classes (such as the thyroid and muscles), other classes must be annotated, or at least reviewed, by experienced annotators.

Across the 7 annotators, the median number of control points used to annotate objects was 126 per image, with a maximum of 364. Each control point was manually placed (and sometimes adjusted). The mean lead time, measured as the cumulative time spent annotating each image, was 3.5 min/image (excluding outliers > 20 min), indicating that annotating 50 images took just under 3 hours. This highlights the annotation burden associated with dense segmentation, which becomes a problem at scale: 40,000 images would require >2,300 hours, which is infeasible for a small group of annotators, especially since expert input is needed. Upscaling to a larger pool of annotators could ease this burden, with crowdsourced annotation presenting a possible option for large-scale annotation, but robust data quality checking with input from experts would be needed. Alternatively, pre-annotation with trained DL models, or interactive annotation with Segment Anything Models (SAM) such as medical SAM [165] or SAMUS [166], may reduce the time required to annotate each image.

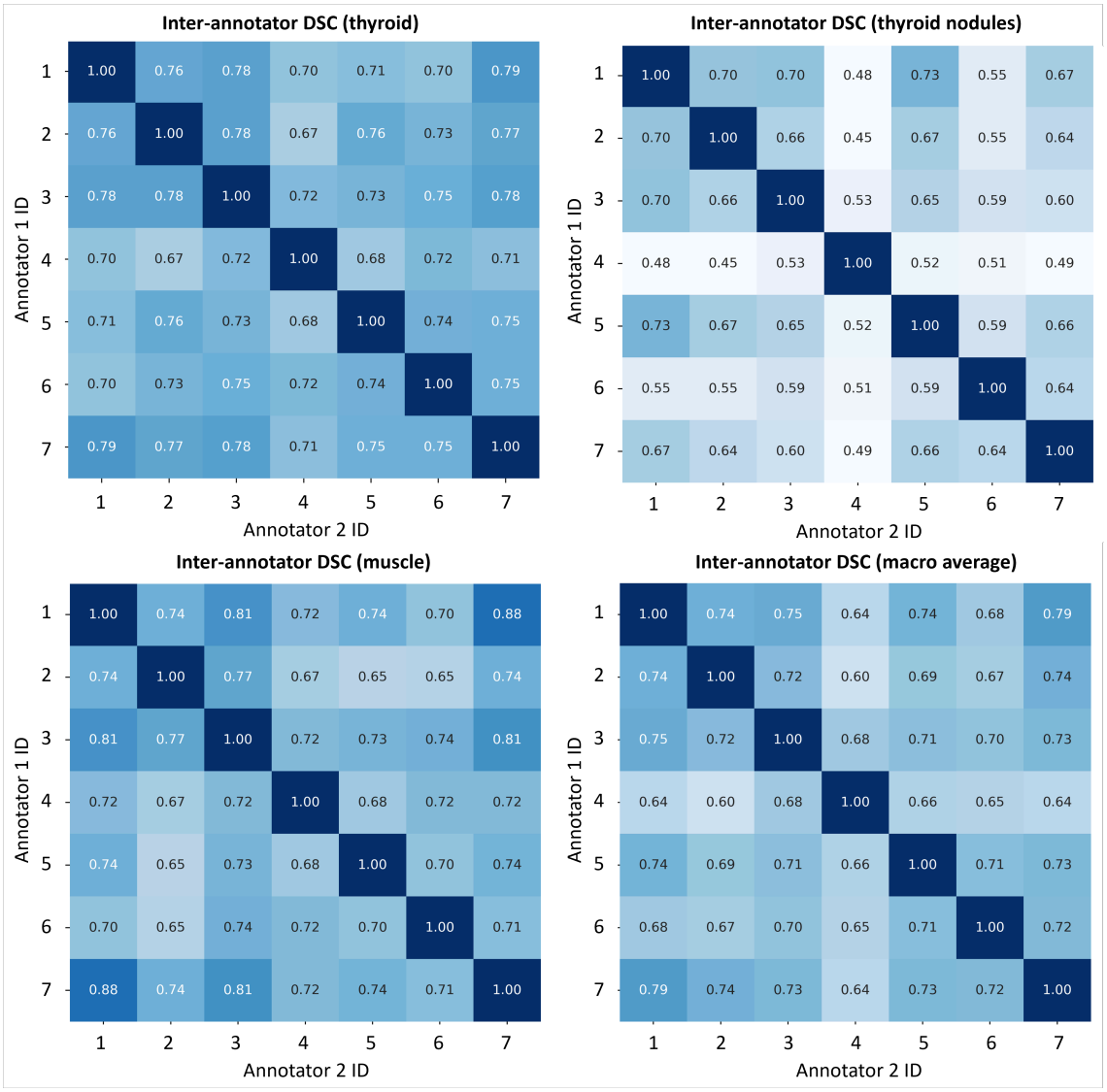


Figure E.7: Matrices showing paired DSC for inter-annotator study. The inter-annotator DSC is calculated per class and reported for thyroid, thyroid nodules, and muscle, as well as the macro-average across all 7 classes.

Bibliography

[1] A. Gunderman, R. Montayre, A. Ranjan, and Y. Chen, "Review of Robot-Assisted HIFU Therapy," *Sensors*, vol. 23, p. 3707, Jan. 2023. Number: 7.

[2] J. E. Kennedy, "High-intensity focused ultrasound in the treatment of solid tumours," *Nature Reviews. Cancer*, vol. 5, pp. 321–327, Apr. 2005.

[3] Z. Izadifar, Z. Izadifar, D. Chapman, and P. Babyn, "An Introduction to High Intensity Focused Ultrasound: Systematic Review on Principles, Devices, and Clinical Applications," *Journal of Clinical Medicine*, vol. 9, p. 460, Feb. 2020. Number: 2.

[4] H. Su, K.-W. Kwok, K. Cleary, I. Iordachita, M. C. Cavusoglu, J. P. Desai, and G. S. Fischer, "State of the Art and Future Opportunities in MRI-Guided Robot-Assisted Surgery and Interventions," *Proceedings of the IEEE*, vol. 110, pp. 968–992, July 2022.

[5] Q. Xiao, R. Monfaredi, M. Musa, K. Cleary, and Y. Chen, "MR-Conditional Actuations: A Review," *Annals of Biomedical Engineering*, vol. 48, pp. 2707–2733, Dec. 2020.

[6] D. Schlesinger, S. Benedict, C. Diederich, W. Gedroyc, A. Klibanov, and J. Larner, "MR-guided focused ultrasound surgery, present and future," *Medical Physics*, vol. 40, no. 8, p. 080901, 2013.

[7] R. D. Kovatcheva and K. Zaletel, "High-intensity focused ultrasound for thyroid nodule ablation: the evidence to date," *Reports in Medical Imaging*, vol. 10, pp. 9–16, Jan. 2017.

[8] I. Pałyga, R. Pałyga, J. Młynarczyk, J. Kopczyński, S. Góźdź, and A. Kowalska, "The current state and future perspectives of high intensity focused ultrasound (HIFU) ablation for benign thyroid nodules," *Gland Surgery*, vol. 9, pp. S95–S9S104, Feb. 2020. Number: Suppl 2.

[9] T. Williamson, S. Everitt, and S. Chauhan, "Automated geometric optimization for robotic HIFU treatment of liver tumors," *Computers in Biology and Medicine*, vol. 96, pp. 1–7, May 2018.

[10] B. H.-H. Lang and A. L. H. Wu, "High intensity focused ultrasound (HIFU) ablation of benign thyroid nodules – a systematic review," *Journal of Therapeutic Ultrasound*, vol. 5, p. 11, May 2017.

[11] R. D. Kovatcheva, J. D. Vlahov, J. I. Stoinov, and K. Zaletel, "Benign Solid Thyroid Nodules: US-guided High-Intensity Focused Ultrasound Ablation—Initial Clinical Outcomes," *Radiology*, vol. 276, pp. 597–605, Aug. 2015.

[12] S. Secinaro, D. Calandra, A. Secinaro, V. Muthurangu, and P. Biancone, "The role

of artificial intelligence in healthcare: a structured literature review," *BMC Medical Informatics and Decision Making*, vol. 21, p. 125, Apr. 2021.

[13] Y. Fu, Y. Lei, T. Wang, W. J. Curran, T. Liu, and X. Yang, "A review of deep learning based methods for medical image multi-organ segmentation," *Physica Medica*, vol. 85, pp. 107–122, May 2021.

[14] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27–48, Apr. 2016.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, eds.), Lecture Notes in Computer Science, (Cham), pp. 234–241, Springer International Publishing, 2015.

[16] Z. Wang, "Deep Learning in Medical Ultrasound Image Segmentation: a Review," 2021.

[17] R. Yao, J. Hu, W. Zhao, Y. Cheng, and C. Feng, "A review of high-intensity focused ultrasound as a novel and non-invasive interventional radiology technique," *Journal of Interventional Medicine*, vol. 5, pp. 127–132, Aug. 2022.

[18] A. L. Klibanov and J. A. Hossack, "Ultrasound in Radiology: from Anatomic, Functional, Molecular Imaging to Drug Delivery and Image-Guided Therapy," *Investigative radiology*, vol. 50, pp. 657–670, Sept. 2015.

[19] P. R. Hoskins, K. Martin, and A. Thrush, eds., *Diagnostic Ultrasound: Physics and Equipment*. Cambridge: Cambridge University Press, 2 ed., 2010.

[20] C. R. Hill, I. Rivens, M. G. Vaughan, and G. R. Ter Haar, "Lesion development in focused ultrasound surgery: A general model," *Ultrasound in Medicine & Biology*, vol. 20, pp. 259–269, Jan. 1994.

[21] Z. Xu, T. L. Hall, E. Vlaisavljevich, and F. T. Lee Jr., "Histotripsy: the first noninvasive, non-ionizing, non-thermal ablation technique based on ultrasound," *International Journal of Hyperthermia*, vol. 38, pp. 561–575, Jan. 2021.

[22] W. Kim, Y. L. Yu, J. Ryu, Y. J. Ju, and S. Kang, "How Do Patients Value the Benefit of Minimally Invasive Surgery in Cancer Treatment?," *Value in Health*, vol. 25, pp. 1212–1217, July 2022.

[23] Y.-F. Zhou, "High intensity focused ultrasound in clinical tumor ablation," *World Journal of Clinical Oncology*, vol. 2, pp. 8–27, Jan. 2011.

[24] S. Dason, N. C. Wong, C. B. Allard, J. Hoogenes, W. Orovan, and B. Shayegan, "High-Intensity Focused Ultrasound (HIFU) as salvage therapy for radio-recurrent

prostate cancer: predictors of disease response," *International Brazilian Journal of Urology*, vol. 44, no. 2, pp. 248–257, 2018.

[25] NICE, "High-intensity focused ultrasound for symptomatic benign thyroid nodules," NICE Guidelines IPG643, NICE, Feb. 2019.

[26] C. Brace, "Thermal Tumor Ablation in Clinical Use," *IEEE pulse*, vol. 2, no. 5, pp. 28–38, 2011.

[27] G. T. Haar and C. Coussios, "High intensity focused ultrasound: physical principles and devices," *International Journal of Hyperthermia*, vol. 23, pp. 89–104, Mar. 2007.

[28] M. Nikfarjam, V. Muralidharan, and C. Christophi, "Mechanisms of Focal Heat Destruction of Liver Tumors," *Journal of Surgical Research*, vol. 127, pp. 208–223, Aug. 2005.

[29] S. A. Sapareto and W. C. Dewey, "Thermal dose determination in cancer therapy," *International Journal of Radiation Oncology, Biology, Physics*, vol. 10, pp. 787–800, June 1984.

[30] M. R. Bailey, V. A. Khokhlova, O. A. Sapozhnikov, S. G. Kargl, and L. A. Crum, "Physical mechanisms of the therapeutic effect of ultrasound (a review)," *Acoustical Physics*, vol. 49, pp. 369–388, July 2003.

[31] J. McLaughlan, I. Rivens, T. Leighton, and G. t. Haar, "A Study of Bubble Activity Generated in Ex Vivo Tissue by High Intensity Focused Ultrasound," *Ultrasound in Medicine and Biology*, vol. 36, pp. 1327–1344, Aug. 2010.

[32] Focused Ultrasound Foundation, "2023 State of the Field Report," tech. rep., Focused Ultrasound Foundation, July 2023.

[33] HistoSonics, "FDA Awards HistoSonics Clearance of its First-of-a-Kind Edison® Histotripsy System," Oct. 2023.

[34] NICE, "Interventional procedure overview of high-intensity focused ultrasound for symptomatic benign thyroid nodules," Interventional Procedures Programme IP 1718 [IPG643], NICE, July 2018.

[35] J. Coon, N. Todd, and R. Roemer, "HIFU treatment time reduction through heating approach optimisation," *International Journal of Hyperthermia*, vol. 28, pp. 799–820, Dec. 2012.

[36] C. M. C. Tempany, N. J. McDannold, K. Hynynen, and F. A. Jolesz, "Focused Ultrasound Surgery in Oncology: Overview and Principles," *Radiology*, vol. 259, p. 39, Apr. 2011.

[37] C. Y. An, J. H. Syu, C. S. Tseng, and C.-J. Chang, "An Ultrasound Imaging-Guided

Robotic HIFU Ablation Experimental System and Accuracy Evaluations," *Applied Bionics and Biomechanics*, vol. 2017, p. e5868695, Apr. 2017.

[38] S. Chauhan and G. ter Haar, "FUSBOTUS: Empirical Studies Using a Surgical Robotic System for Urological Applications," *AIP Conference Proceedings*, vol. 911, pp. 117–121, May 2007.

[39] T. Tang, T. Azuma, T. Iwahashi, H. Takeuchi, E. Kobayashi, and I. Sakuma, "A High-Precision US-Guided Robot-Assisted HIFU Treatment System for Breast Cancer," *Engineering*, vol. 4, pp. 702–713, Oct. 2018.

[40] C. Y. An, Y. L. Hsu, and C. S. Tseng, "An Ultrasound-Guided Robotic HIFU Ablation System with Respiration Induced Displacement and Time Delay Compensation," *Journal of Medical and Biological Engineering*, vol. 39, pp. 796–805, Oct. 2019.

[41] L. Daunizeau, A. Nguyen, M. Le Garrec, J. Y. Chapelon, and W. A. N'Djin, "Robot-assisted ultrasound navigation platform for 3D HIFU treatment planning: Initial evaluation for conformal interstitial ablation," *Computers in Biology and Medicine*, vol. 124, p. 103941, Sept. 2020.

[42] B. H. Lang and A. L. H. Wu, "The efficacy and safety of high-intensity focused ultrasound ablation of benign thyroid nodules," *Ultrasonography (Seoul, Korea)*, vol. 37, pp. 89–97, Apr. 2018.

[43] S. Guth, U. Theune, J. Aberle, A. Galach, and C. M. Bamberger, "Very high prevalence of thyroid nodules detected by high frequency (13 MHz) ultrasound examination," *European Journal of Clinical Investigation*, vol. 39, no. 8, pp. 699–706, 2009.

[44] Z. Cheng and P. Liang, "Advances in ultrasound-guided thermal ablation for symptomatic benign thyroid nodules," *Advances in clinical and experimental medicine*, vol. 29, pp. 1123–1129, Sept. 2020.

[45] C. Durante, G. Costante, G. Lucisano, R. Bruno, D. Meringolo, A. Paciaroni, E. Puxeddu, M. Torlontano, S. Tumino, M. Attard, L. Lamartina, A. Nicolucci, and S. Filetti, "The natural history of benign thyroid nodules," *JAMA*, vol. 313, pp. 926–935, Mar. 2015.

[46] A. diMarco and F. Palazzo, "Goitre and thyroid cancer," *Medicine*, vol. 45, pp. 517–522, Aug. 2017.

[47] B. R. Haugen, "2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer: What is new and what has changed?," *Cancer*, vol. 123, no. 3, pp. 372–381, 2017.

[48] A. Bergenfelz, S. Jansson, A. Kristoffersson, H. Mårtensson, E. Reihnér, G. Wallin, and I. Lausen, "Complications to thyroid surgery: results as reported in a database

from a multicenter audit comprising 3,660 patients," *Langenbeck's Archives of Surgery*, vol. 393, pp. 667–673, Sept. 2008.

[49] N. Kotewall and B. H. H. Lang, "High-intensity focused ultrasound ablation as a treatment for benign thyroid diseases: the present and future," *Ultrasonography*, vol. 38, pp. 135–142, Apr. 2019.

[50] K.-Y. Fan, E.-W. Loh, and K.-W. Tam, "Efficacy of HIFU for the treatment of benign thyroid nodules: a systematic review and meta-analysis," *European Radiology*, vol. 34, pp. 2310–2322, Apr. 2024.

[51] B. H. H. Lang, Y.-C. Woo, and K. W.-H. Chiu, "Two-year efficacy of single-session high-intensity focused ultrasound (HIFU) ablation of benign thyroid nodules," *European Radiology*, vol. 29, pp. 93–101, Jan. 2019.

[52] B. H. H. Lang, Y.-C. Woo, and K. W.-H. Chiu, "Sequential high intensity focused ultrasound (HIFU) ablation in the treatment of benign multinodular goitre: an observational retrospective study," *European Radiology*, vol. 28, pp. 3237–3244, Aug. 2018.

[53] L. Mascagni, A. Posa, A. R. Scrofani, M. Lippi, V. Lancellotta, R. Castellana, P. Cornacchione, L. Tagliaferri, C. Capotondi, and R. Iezzi, "Thermal ablation compared to surgery for symptomatic benign thyroid nodules: a systematic review," *British Journal of Radiology*, vol. 98, pp. 2056–2065, Dec. 2025.

[54] Y. Korkusuz, D. Gröner, N. Raczynski, O. Relin, Y. Kingeter, F. Grünwald, and C. Happel, "Thermal ablation of thyroid nodules: are radiofrequency ablation, microwave ablation and high intensity focused ultrasound equally safe and effective methods?," *European Radiology*, vol. 28, pp. 929–935, Mar. 2018.

[55] E. H. Benaim, C. Nieri, M. Mamidala, M. J. Herr, A. Sheyn, and M. B. Gillespie, "High-intensity focused ultrasound for benign thyroid nodules: Systemic review and meta-analysis," *American Journal of Otolaryngology*, vol. 44, p. 103999, Nov. 2023.

[56] B. H. H. Lang, C. K. H. Wong, and E. P. M. Ma, "Single-session high intensity focussed ablation (HIFU) versus open cervical hemithyroidectomy for benign thyroid nodule: analysis on early efficacy, safety and voice quality," *International Journal of Hyperthermia*, vol. 33, pp. 868–874, Dec. 2017.

[57] A. Nugroho, R. Hidayat, and H. A. Nugroho, "Thyroid Ultrasound Image Segmentation: A Review," in *2019 5th International Conference on Science and Technology (ICST)*, vol. 1, pp. 1–6, July 2019.

[58] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," June 2021. arXiv:2010.11929 [cs].

[59] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. Roth, and D. Xu, "UNETR: Transformers for 3D Medical Image Segmentation," Oct. 2021. arXiv:2103.10504 [eess].

[60] X. Nie, X. Zhou, T. Tong, X. Lin, L. Wang, H. Zheng, J. Li, E. Xue, S. Chen, M. Zheng, C. Chen, and M. Du, "N-Net: A novel dense fully convolutional neural network for thyroid nodule segmentation," *Frontiers in Neuroscience*, vol. 16, 2022.

[61] N. Yadav, R. Dass, and J. Virmani, "Objective assessment of segmentation models for thyroid ultrasound images," *Journal of Ultrasound*, Oct. 2022.

[62] T. Zheng, H. Qin, Y. Cui, R. Wang, W. Zhao, S. Zhang, S. Geng, and L. Zhao, "Segmentation of thyroid glands and nodules in ultrasound images using the improved U-Net architecture," *BMC Medical Imaging*, vol. 23, p. 56, Apr. 2023.

[63] N. S. Narayan, P. Marziliano, J. Kanagalingam, and C. G. L. Hobbs, "Speckle Patch Similarity for Echogenicity-Based Multiorgan Segmentation in Ultrasound Images of the Thyroid Gland," *IEEE journal of biomedical and health informatics*, vol. 21, pp. 172–183, Jan. 2017.

[64] H. Gong, J. Chen, G. Chen, H. Li, G. Li, and F. Chen, "Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules," *Computers in Biology and Medicine*, vol. 155, p. 106389, Mar. 2023.

[65] J. Sun, C. Li, Z. Lu, M. He, T. Zhao, X. Li, L. Gao, K. Xie, T. Lin, J. Sui, Q. Xi, F. Zhang, and X. Ni, "TNSNet: Thyroid nodule segmentation in ultrasound imaging using soft shape supervision," *Computer Methods and Programs in Biomedicine*, vol. 215, p. 106600, Mar. 2022.

[66] M. Yu, M. Han, X. Li, X. Wei, H. Jiang, H. Chen, and R. Yu, "Adaptive soft erasure with edge self-attention for weakly supervised semantic segmentation: Thyroid ultrasound image case study," *Computers in Biology and Medicine*, vol. 144, p. 105347, May 2022.

[67] Q. Kang, Q. Lao, Y. Li, Z. Jiang, Y. Qiu, S. Zhang, and K. Li, "Thyroid nodule segmentation and classification in ultrasound images through intra- and inter-task consistent learning," *Medical Image Analysis*, vol. 79, p. 102443, July 2022.

[68] K. Su, J. Liu, X. Ren, Y. Huo, G. Du, W. Zhao, X. Wang, B. Liang, D. Li, and P. X. Liu, "A fully autonomous robotic ultrasound system for thyroid scanning," *Nature Communications*, vol. 15, p. 4004, May 2024.

[69] R. Cunningham, M. B. Sánchez, and I. D. Loram, "Ultrasound Segmentation of Cervical Muscle during head motion: a Dataset and a Benchmark using Deconvolutional Neural Networks," Feb. 2019.

[70] K. Huang, Y. Zhang, H. Cheng, and P. Xing, "Trustworthy Breast Ultrasound Image

Semantic Segmentation Based on Fuzzy Uncertainty Reduction," *Healthcare*, vol. 10, p. 2480, Dec. 2022.

[71] Q. Huang and Z. Zeng, "A Review on Real-Time 3D Ultrasound Imaging Technology," *BioMed Research International*, vol. 2017, no. 1, p. 6027029, 2017.

[72] Y. Wang, Q. Yang, L. Drukker, A. Papageorghiou, Y. Hu, and J. A. Noble, "Task model-specific operator skill assessment in routine fetal ultrasound scanning," *International Journal of Computer Assisted Radiology and Surgery*, vol. 17, no. 8, pp. 1437–1444, 2022.

[73] C.-H. Yee, P. K.-F. Chiu, J. Y.-C. Teoh, C.-F. Ng, C.-K. Chan, and S.-M. Hou, "High-Intensity Focused Ultrasound (HIFU) Focal Therapy for Localized Prostate Cancer with MRI-US Fusion Platform," *Advances in Urology*, vol. 2021, no. 1, p. 7157973, 2021.

[74] F. Mohamed, C. V. Siang, F. Mohamed, and C. V. Siang, *A Survey on 3D Ultrasound Reconstruction Techniques*. IntechOpen, Apr. 2019. Publication Title: Artificial Intelligence - Applications in Medicine and Biology.

[75] M. Ingram, M. Antico, L. Cattani, M. Cvijic, L. Scheys, J. D'hooge, and E. Harris, "Three-dimensional ultrasound imaging: a review of the technology," *Physics in Medicine & Biology*, vol. 70, p. 22TR01, Nov. 2025.

[76] R. W. Prager, A. Gee, and L. Berman, "Stradx: real-time acquisition and visualization of freehand three-dimensional ultrasound," *Medical Image Analysis*, vol. 3, pp. 129–140, June 1999.

[77] C. Peng, Q. Cai, M. Chen, and X. Jiang, "Recent Advances in Tracking Devices for Biomedical Ultrasound Imaging Applications," *Micromachines*, vol. 13, p. 1855, Oct. 2022.

[78] M. Krönke, C. Eilers, D. Dimova, M. Köhler, G. Buschner, L. Schweiger, L. Konstantinidou, M. Makowski, J. Nagarajah, N. Navab, W. Weber, and T. Wendler, "Tracked 3D ultrasound and deep neural network-based thyroid segmentation reduce interobserver variability in thyroid volumetry," *PLOS ONE*, vol. 17, p. e0268550, July 2022.

[79] S.-W. Chung, C.-C. Shih, and C.-C. Huang, "Freehand three-dimensional ultrasound imaging of carotid artery using motion tracking technology," *Ultrasonics*, vol. 74, pp. 11–20, Feb. 2017.

[80] F. Poulin and L. P. Amiot, "Interference during the use of an electromagnetic tracking system under OR conditions," *Journal of Biomechanics*, vol. 35, pp. 733–737, June 2002.

[81] N. M. Bekedam, L. H. E. Karssemakers, M. J. A. van Alphen, R. L. P. van Veen,

L. E. Smeele, and M. B. Karakullukcu, "Comparison of image quality of 3D ultrasound: motorized acquisition versus freehand navigated acquisition, a phantom study," *International Journal of Computer Assisted Radiology and Surgery*, vol. 18, pp. 1649–1663, Sept. 2023.

[82] A. M. Priester, S. Natarajan, and M. O. Culjat, "Robotic ultrasound systems in medicine," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 60, pp. 507–523, Mar. 2013.

[83] X.-X. Lin, M.-D. Li, S.-M. Ruan, W.-P. Ke, H.-R. Zhang, H. Huang, S.-H. Wu, M.-Q. Cheng, W.-J. Tong, H.-T. Hu, D.-N. He, R.-F. Lu, Y.-D. Lin, M. Kuang, M.-D. Lu, L.-D. Chen, Q.-H. Huang, and W. Wang, "Autonomous robotic ultrasound scanning system: a key to enhancing image analysis reproducibility and observer consistency in ultrasound imaging," *Frontiers in Robotics and AI*, vol. 12, Feb. 2025.

[84] Q. Huang, B. Gao, and M. Wang, "Robot-Assisted Autonomous Ultrasound Imaging for Carotid Artery," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–9, 2024.

[85] Y. Bi, Z. Jiang, Y. Gao, T. Wendler, A. Karlas, and N. Navab, "VesNet-RL: Simulation-Based Reinforcement Learning for Real-World US Probe Navigation," *IEEE Robotics and Automation Letters*, vol. 7, pp. 6638–6645, July 2022.

[86] K. AlMuhanna, M. M. Hossain, L. Zhao, J. Fischell, G. Kowalewski, M. Dux, S. Sikdar, and B. K. Lal, "Carotid plaque morphometric assessment with three-dimensional ultrasound imaging," *Journal of Vascular Surgery*, vol. 61, pp. 690–697, Mar. 2015.

[87] H. Jiang, A. Zhao, Q. Yang, X. Yan, T. Wang, Y. Wang, N. Jia, J. Wang, G. Wu, Y. Yue, S. Luo, H. Wang, L. Ren, S. Chen, P. Liu, G. Yao, W. Yang, S. Song, X. Li, K. He, and G. Huang, "Towards expert-level autonomous carotid ultrasonography with large-scale learning-based robotic system," *Nature Communications*, vol. 16, p. 7893, Aug. 2025.

[88] D. Huang, Y. Bi, N. Navab, and Z. Jiang, "Motion Magnification in Robotic Sonography: Enabling Pulsation-Aware Artery Segmentation," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6565–6570, Oct. 2023.

[89] Z. Wang, Y. Han, B. Zhao, H. Xie, L. Yao, B. Li, M. Q.-H. Meng, and Y. Hu, "Autonomous Robotic System for Carotid Artery Ultrasound Scanning With Visual Servo Navigation," *IEEE Transactions on Medical Robotics and Bionics*, vol. 6, pp. 1436–1447, Nov. 2024.

[90] G. Faoro, S. Maglio, S. Pane, V. Iacovacci, and A. Menciassi, "An Artificial Intelligence-Aided Robotic Platform for Ultrasound-Guided Transcarotid Revascu-

larization," *IEEE Robotics and Automation Letters*, vol. 8, pp. 2349–2356, Apr. 2023. Conference Name: IEEE Robotics and Automation Letters.

[91] Q. Huang, B. Wu, J. Lan, and X. Li, "Fully Automatic Three-Dimensional Ultrasound Imaging Based on Conventional B-Scan," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 12, pp. 426–436, Apr. 2018.

[92] Q. Huang, J. Lan, and X. Li, "Robotic Arm Based Automatic Ultrasound Scanning for Three-Dimensional Imaging," *IEEE Transactions on Industrial Informatics*, vol. PP, pp. 1–1, Sept. 2018.

[93] J. T. Kaminski, K. Rafatzand, and H. K. Zhang, "Feasibility of Robot-Assisted Ultrasound Imaging with Force Feedback for Assessment of Thyroid Diseases," *Proceedings of the International Society for Optical Engineering*, vol. 11315, p. 113151D, Feb. 2020.

[94] R. Kojcev, A. Khakzar, B. Fuerst, O. Zettinig, C. Fahkry, R. DeJong, J. Richmon, R. Taylor, E. Sinibaldi, and N. Navab, "On the reproducibility of expert-operated and robotic ultrasound acquisitions," *International Journal of Computer Assisted Radiology and Surgery*, vol. 12, pp. 1003–1011, June 2017.

[95] J. Zielke, C. Eilers, B. Busam, W. Weber, N. Navab, and T. Wendler, "RSV: Robotic Sonography for Thyroid Volumetry," *IEEE Robotics and Automation Letters*, vol. 7, pp. 3342–3348, Apr. 2022.

[96] T. Haidegger, "Autonomy for Surgical Robots: Concepts and Paradigms," *IEEE Transactions on Medical Robotics and Bionics*, vol. 1, pp. 65–76, May 2019.

[97] Y. Bi, Z. Jiang, F. Duelmer, D. Huang, and N. Navab, "Machine Learning in Robotic Ultrasound Imaging: Challenges and Perspectives," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 7, pp. 335–357, July 2024.

[98] R. Finocchi, F. Aalamifar, T. Y. Fang, R. H. Taylor, and E. M. Boctor, "Co-robotic ultrasound imaging: a cooperative force control approach," in *Medical Imaging 2017: Image-Guided Procedures, Robotic Interventions, and Modeling*, vol. 10135, pp. 270–280, SPIE, Mar. 2017.

[99] G.-Z. Yang, J. Cambias, K. Cleary, E. Daimler, J. Drake, P. E. Dupont, N. Hata, P. Kazanzides, S. Martel, R. V. Patel, V. J. Santos, and R. H. Taylor, "Medical robotics-Regulatory, ethical, and legal considerations for increasing levels of autonomy," *Science Robotics*, vol. 2, p. eaam8638, Mar. 2017.

[100] F. Ferraguti, C. Talignani Landi, L. Sabattini, M. Bonfè, C. Fantuzzi, and C. Secchi, "A variable admittance control strategy for stable physical human–robot interaction," *The International Journal of Robotics Research*, vol. 38, pp. 747–765, May 2019.

[101] F. von Haxthausen, S. Böttger, D. Wulff, J. Hagenah, V. García-Vázquez, and

S. Ipsen, "Medical Robotics for Ultrasound Imaging: Current Systems and Future Trends," *Current Robotics Reports*, vol. 2, pp. 55–71, Mar. 2021.

[102] J. P. Yung, A. Shetty, A. Elliott, J. S. Weinberg, R. J. McNichols, A. Gowda, J. D. Hazle, and R. J. Stafford, "Quantitative comparison of thermal dose models in normal canine brain," *Medical Physics*, vol. 37, pp. 5313–5321, Oct. 2010.

[103] M. Correa, M. Zimic, F. P. Porras, R. Barrientos, A. Roman-Gonzalez, M. Pajuelo, C. Anticona, H. Mayta, A. Alva, L. Solis-Vasquez, D. Figueroa, M. Chavez, R. Lavarello, B. Castaneda, V. Paz-Soldán, W. Checkley, R. Gilman, and R. Oberhelman, "Automatic classification of pediatric pneumonia based on lung ultrasound pattern recognition," *PLOS ONE*, vol. 13, p. e0206410, Dec. 2018.

[104] S. Shah, B. A. Bellows, A. A. Adedipe, J. E. Totten, B. H. Backlund, and D. Sajed, "Perceived barriers in the use of ultrasound in developing countries," *Critical Ultrasound Journal*, vol. 7, p. 28, Dec. 2015.

[105] M. Chen, R. Ball, L. Yang, N. Moradzadeh, B. Chapman, D. Larson, C. Langlotz, T. Amrhein, and M. Lungren, "Deep learning to classify radiology free-text reports," *Radiology*, vol. 286, no. 3, pp. 845–852, 2018.

[106] P. Chatelain, A. Krupa, and N. Navab, "3D ultrasound-guided robotic steering of a flexible needle via visual servoing," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2250–2255, May 2015.

[107] D. Li, Z. Cheng, G. Chen, F. Liu, W. Wu, J. Yu, Y. Gu, F. Liu, C. Ren, and P. Liang, "A multimodality imaging-compatible insertion robot with a respiratory motion calibration module designed for ablation of liver tumors: a preclinical study," *International Journal of Hyperthermia*, vol. 34, pp. 1194–1201, Nov. 2018.

[108] T. Boers, S. J. Braak, W. M. Brink, M. Versluis, and S. Manohar, "3D ultrasound guidance for radiofrequency ablation in an anthropomorphic thyroid nodule phantom," *European Radiology Experimental*, vol. 8, p. 115, Oct. 2024.

[109] University College London Hospitals NHS Foundation Trust, "Radioiodine therapy for benign thyroid disease," May 2024.

[110] P. Szumowski, M. Mojsak, S. Abdelrazek, M. Sykała, A. Amelian-Fiłonowicz, D. Jurgilewicz, and J. Myśliwiec, "Calculation of therapeutic activity of radioiodine in Graves' disease by means of Marinelli's formula, using technetium (99mTc) scintigraphy," *Endocrine*, vol. 54, pp. 751–756, Dec. 2016.

[111] S. Merouche, L. Allard, E. Montagnon, G. Soulez, P. Bigras, and G. Cloutier, "A Robotic Ultrasound Scanner for Automatic Vessel Tracking and Three-Dimensional Reconstruction of B-Mode Images," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 63, pp. 35–46, Jan. 2016.

[112] Y. Wang, G. Wang, and C.-q. Gao, "Ultrasound-guided cannulation of the internal jugular vein in robotic cardiac surgery," *Chinese Medical Journal*, vol. 126, pp. 2414–2417, July 2013.

[113] E. White, S. Myhre, M. R. Serafini, M. Broad, A. Chesnut, D. Heishman, J. Knupp, and J. C.-y. Chao, "Focused Ultrasound Foundation: State of the field 2025," July 2025.

[114] M. A. Kisting, M. S. Jentink, M. G. Wagner, Z. Xu, J. L. Hinshaw, P. F. Laeseke, T. J. Ziemlewicz, E. Knavel Koepsel, and F. T. Lee Jr., "Imaging for Targeting, Monitoring, and Assessment After Histotripsy: A Non-invasive, Non-thermal Therapy for Cancer," *EMJ Radiology*, vol. 10, pp. 15–21, Mar. 2023.

[115] H. Dou, L. Han, Y. He, J. Xu, N. Ravikumar, R. Mann, A. F. Frangi, P.-T. Yap, and Y. Huang, "Localizing the Recurrent Laryngeal Nerve via Ultrasound with a Bayesian Shape Framework," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022* (L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, eds.), Lecture Notes in Computer Science, (Cham), pp. 258–267, Springer Nature Switzerland, 2022.

[116] D. Wu, D. Smith, B. VanBerlo, A. Roshankar, H. Lee, B. Li, F. Ali, M. Rahman, J. Basmaji, J. Tschirhart, A. Ford, B. VanBerlo, A. Durvasula, C. Vannelli, C. Dave, J. Deglint, J. Ho, R. Chaudhary, H. Clausdorff, R. Prager, S. Millington, S. Shah, B. Buchanan, and R. Arntfield, "Improving the Generalizability and Performance of an Ultrasound Deep Learning Model Using Limited Multicenter Data for Lung Sliding Artifact Identification," *Diagnostics*, vol. 14, p. 1081, May 2024.

[117] A. Coronato, M. Naeem, G. De Pietro, and G. Paragliola, "Reinforcement learning for intelligent healthcare applications: A survey," *Artificial Intelligence in Medicine*, vol. 109, p. 101964, Sept. 2020.

[118] H. van Hasselt, A. Guez, and D. Silver, "Deep Reinforcement Learning with Double Q-learning," Dec. 2015. arXiv:1509.06461 [cs].

[119] V. F. Humphrey, "Nonlinear propagation in ultrasonic fields: measurements, modelling and harmonic imaging," *Ultrasonics*, vol. 38, pp. 267–272, Mar. 2000.

[120] H. Kok, E. Cressman, W. Ceelen, C. Brace, R. Ivkov, H. Grüll, G. ter Haar, P. Wust, and J. Crezee, "Heating technology for malignant tumors: a review," *International Journal of Hyperthermia*, vol. 37, no. 1, pp. 711–741, 2020.

[121] A. L. Kapetanović, A. Šušnjara, and D. Poljak, "Numerical Solution and Uncertainty Quantification of Bioheat Transfer Equation Using Neural Network Approach," in *2020 5th International Conference on Smart and Sustainable Technologies (SpliTech)*, pp. 1–6, Sept. 2020.

[122] S. Cuomo, V. S. di Cola, F. Giampaolo, G. Rozza, M. Raissi, and F. Piccialli,

"Scientific Machine Learning through Physics-Informed Neural Networks: Where we are and What's next," June 2022. arXiv:2201.05624 [physics].

[123] M. K. Almekkaway, I. A. Shehata, and E. S. Ebbini, "Anatomical-based model for simulation of HIFU-induced lesions in atherosclerotic plaques," *International Journal of Hyperthermia*, vol. 31, pp. 433–442, May 2015.

[124] S. Virga, R. Göbl, M. Baust, N. Navab, and C. Hennersperger, "Use the force: deformation correction in robotic 3D ultrasound," *International Journal of Computer Assisted Radiology and Surgery*, vol. 13, pp. 619–627, May 2018.

[125] T. Boers, W. Brink, L. Bianchi, P. Saccomandi, J. van Hespen, G. Wennemars, S. Braak, M. Versluis, and S. Manohar, "An anthropomorphic thyroid phantom for ultrasound-guided radiofrequency ablation of nodules," *Medical Physics*, vol. 51, no. 2, pp. 826–838, 2024.

[126] T. Boers, S. J. Braak, M. Versluis, and S. Manohar, "Matrix 3D ultrasound-assisted thyroid nodule volume estimation and radiofrequency ablation: a phantom study," *European Radiology Experimental*, vol. 5, p. 31, July 2021.

[127] H. H. L. Ng, V. W.-S. Chan, L. Howell, T. Shiwani, J. Zhong, J. Brandon, A. Samson, J. Chandler, J. Mclaughlan, and T. M. Wah, "Early Adoption of Image-Guided Histotripsy Therapy in Interventional Oncology: Challenges and Opportunities in the UK," *British Journal of Radiology*, p. tqag047, Mar. 2026.

[128] J.-F. Aubry, K. B. Pauly, C. Moonen, G. T. Haar, M. Ries, R. Salomir, S. Sokka, K. M. Sekins, Y. Shapira, F. Ye, H. Huff-Simonin, M. Eames, A. Hananel, N. Kassell, A. Napoli, J. H. Hwang, F. Wu, L. Zhang, A. Melzer, Y.-S. Kim, and W. M. Gedroyc, "The road to clinical use of high-intensity focused ultrasound for liver cancer: technical and clinical consensus," *Journal of Therapeutic Ultrasound*, vol. 1, p. 13, 2013.

[129] C.-L. Yao, T. Trinh, G. T. C. Wong, and M. G. Irwin, "Anaesthesia for high intensity focused ultrasound (HIFU) therapy," *Anaesthesia*, vol. 63, no. 8, pp. 865–872, 2008.

[130] A. B. Holbrook, P. Ghanouni, J. M. Santos, C. Dumoulin, Y. Medan, and K. B. Pauly, "Respiration Based Steering for High Intensity Focused Ultrasound Liver Ablation," *Magnetic resonance in medicine*, vol. 71, pp. 797–806, Feb. 2014.

[131] P. Bour, V. Ozenne, F. Marquet, B. Denis de Senneville, E. Dumont, and B. Quesson, "Real-time 3D ultrasound based motion tracking for the treatment of mobile organs with MR-guided high-intensity focused ultrasound," *International Journal of Hyperthermia*, vol. 34, pp. 1225–1235, Nov. 2018. _eprint: https://doi.org/10.1080/02656736.2018.1433879.

[132] N. V. Ruiter, P. Pfistner, M. Zapf, Z. Lu, S. E. Kraft, and T. Hopp, "Image based

Motion Tracking for 3D Ultrasound Tomography," in *2025 IEEE International Ultrasonics Symposium (IUS)*, pp. 1–3, Sept. 2025. ISSN: 1948-5727.

[133] M. Pernot, M. Tanter, and M. Fink, "3-D real-time motion correction in high-intensity focused ultrasound therapy," *Ultrasound in Medicine & Biology*, vol. 30, pp. 1239–1249, Sept. 2004.

[134] M. Sadeghi-Goughari, H. Rajabzadeh, J.-w. Han, and H.-J. Kwon, "Artificial intelligence-assisted ultrasound-guided focused ultrasound therapy: a feasibility study," *International Journal of Hyperthermia*, vol. 40, p. 2260127, Dec. 2023.

[135] J. Krücker, S. Xu, A. Venkatesan, J. K. Locklin, H. Amalo, N. Glossop, and B. J. Wood, "Clinical Utility of Real-Time Fusion Guidance for Biopsy and Ablation," *Journal of vascular and interventional radiology : JVIR*, vol. 22, pp. 515–524, Apr. 2011.

[136] E. S. Ebbini and G. Ter Haar, "Ultrasound-guided therapeutic focused ultrasound: Current status and future directions," *International Journal of Hyperthermia*, vol. 31, pp. 77–89, Feb. 2015.

[137] M. H. A. Groen, F. J. B. Slieker, A. Vink, G. J. d. Borst, M. V. Simons, E. S. Ebbini, P. A. Doevendans, C. E. V. B. Hazenberg, and R. v. Es, "Safety and feasibility of arterial wall targeting with robot-assisted high intensity focused ultrasound: a preclinical study," *International Journal of Hyperthermia*, vol. 37, pp. 903–912, Jan. 2020.

[138] F. Jacso, A. Kouznetsov, and W. L. Smith, "Development and evaluation of an ultrasound-guided tracking and gating system for hepatic radiotherapy," *Medical Physics*, vol. 36, pp. 5633–5640, Dec. 2009.

[139] C.-L. Cao, Q.-L. Li, J. Tong, L.-N. Shi, W.-X. Li, Y. Xu, J. Cheng, T.-T. Du, J. Li, and X.-W. Cui, "Artificial intelligence in thyroid ultrasound," *Frontiers in Oncology*, vol. 13, May 2023.

[140] H. Zhang, Q. Liu, X. Han, L. Niu, and W. Sun, "TN5000: An Ultrasound Image Dataset for Thyroid Nodule Detection and Classification," *Scientific Data*, vol. 12, p. 1437, Aug. 2025.

[141] A. Ozcan, O. Tosun, E. Donmez, and M. Sanwal, "Enhanced-TransUNet for ultrasound segmentation of thyroid nodules," *Biomedical Signal Processing and Control*, vol. 95, p. 106472, Sept. 2024.

[142] M. Fang, Q. Ma, and B. Xu, "Automated multi-class classification of thyroid nodules in ultrasound imaging using transformer-based segmentation and hybrid feature learning," *Journal of Radiation Research and Applied Sciences*, vol. 18, p. 101782, Dec. 2025.

[143] Jamie Coulson, "Patient 'honoured' to be treated in Leeds Cancer Centre trial," *BBC News*, Oct. 2024.

[144] G. Carbajal, A. Lasso, A. Gomez, and G. Fichtinger, "Improving N-wire phantom-based freehand ultrasound calibration," *International journal of computer assisted radiology and surgery*, vol. 8, July 2013.

[145] A. Lasso, T. Heffter, A. Rankin, C. Pinter, T. Ungi, and G. Fichtinger, "PLUS: Open-source toolkit for ultrasound-guided intervention systems," *IEEE Transactions on Biomedical Engineering*, vol. 61, pp. 2527–2537, 2014.

[146] L. Pedraza, C. Vargas, F. Narváez, O. Durán, E. Muñoz, and E. Romero, "An open access thyroid ultrasound image database," in *10th International Symposium on Medical Information Processing and Analysis*, vol. 9287, pp. 188–193, SPIE, Jan. 2015.

[147] H. Gong, G. Chen, R. Wang, X. Xie, M. Mao, Y. Yu, F. Chen, and G. Li, "Multi-Task Learning For Thyroid Nodule Segmentation With Thyroid Region Prior," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 257–261, Apr. 2021. ISSN: 1945-8452.

[148] J. Zhou, X. Jia, D. Ni, A. Noble, R. Huang, T. Tan, and M. T. Van, "Thyroid Nodule Segmentation and Classification in Ultrasound Images," Mar. 2020.

[149] Taco Geertsma, "Thyroid gland | Ultrasound Cases," 2018.

[150] T. Wunderling, B. Golla, P. Poudel, C. Arens, M. Friebe, and C. Hansen, "Comparison of thyroid segmentation techniques for 3D ultrasound," *Proc. SPIE 10133, Medical Imaging 2017: Image Processing*, Feb. 2017.

[151] X. Li, S. Zhang, Q. Zhang, X. Wei, Y. Pan, J. Zhao, X. Xin, C. Qin, X. Wang, J. Li, F. Yang, Y. Zhao, M. Yang, Q. Wang, Z. Zheng, X. Zheng, X. Yang, C. T. Whitlow, M. N. Gurcan, L. Zhang, X. Wang, B. C. Pasche, M. Gao, W. Zhang, and K. Chen, "Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study," *The Lancet. Oncology*, vol. 20, pp. 193–201, Feb. 2019.

[152] Stanford Center for AI in Medicine & Imaging, "Thyroid Ultrasound Cine-clip," Apr. 2026.

[153] X. Hou, M. Hua, W. Zhang, J. Ji, X. Zhang, H. Jiang, M. Li, X. Wu, W. Zhao, S. Sun, L. Cao, and L. Wang, "An ultrasonography of thyroid nodules dataset with pathological diagnosis annotation for deep learning," *Scientific Data*, vol. 11, p. 1272, Nov. 2024.

[154] L. Howell, A. Zarei, T. M. Wah, J. H. Chandler, S. Karthik, Z. Court, H. Ng, and J. R. McLaughlan, "RADEX: a rule-based clinical and radiology data extraction tool

demonstrated on thyroid ultrasound reports," *European Radiology*, vol. 35, pp. 4506–4517, Aug. 2025.

[155] P. Perros, K. Boelaert, S. Colley, C. Evans, R. M. Evans, G. Gerrard BA, J. Gilbert, B. Harrison, S. J. Johnson, T. E. Giles, L. Moss, V. Lewington, K. Newbold, J. Taylor, R. V. Thakker, J. Watkinson, and G. R. Williams, "Guidelines for the management of thyroid cancer," *Clinical Endocrinology*, vol. 81, no. s1, pp. 1–122, 2014.

[156] J.-J. Wang, H.-Q. Huang, R.-T. Zheng, Z.-H. Lin, M.-M. Wu, S.-Z. Cai, X.-Y. Liao, D.-M. Guo, and Z. Chen, "Comparative study of different worldwide versions of the thyroid risk stratification system in patients with thyroid nodules in China," *Quantitative Imaging in Medicine and Surgery*, vol. 16, p. 67, Jan. 2026.

[157] F. N. Tessler, W. D. Middleton, E. G. Grant, J. K. Hoang, L. L. Berland, S. A. Teefey, J. J. Cronan, M. D. Beland, T. S. Desser, M. C. Frates, L. W. Hammers, U. M. Hamper, J. E. Langer, C. C. Reading, L. M. Scoutt, and A. T. Stavros, "ACR Thyroid Imaging, Reporting and Data System (TI-RADS): White Paper of the ACR TI-RADS Committee," *Journal of the American College of Radiology*, vol. 14, pp. 587–595, May 2017.

[158] J. Y. Sung, D. G. Na, K. S. Kim, H. Yoo, H. Lee, J.-H. Kim, and J. H. Baek, "Diagnostic accuracy of fine-needle aspiration versus core-needle biopsy for the diagnosis of thyroid malignancy in a clinical cohort," *European Radiology*, vol. 22, pp. 1564–1572, July 2012.

[159] L. Lan, Y. Luo, M. Zhou, L. Huo, H. Chen, Q. Zuo, and W. Deng, "Comparison of Diagnostic Accuracy of Thyroid Cancer With Ultrasound-Guided Fine-Needle Aspiration and Core-Needle Biopsy: A Systematic Review and Meta-Analysis," *Frontiers in Endocrinology*, vol. 11, p. 44, Feb. 2020.

[160] A. Weller, B. Sharif, M. H. Qarib, D. St Leger, H. S. De Silva, and R. K. Lingam, "British Thyroid Association 2014 classification ultrasound scoring of thyroid nodules in predicting malignancy: Diagnostic performance and inter-observer agreement," *Ultrasound*, vol. 28, pp. 4–13, Feb. 2020.

[161] P. Cross, A. Chandra, T. Giles, S. Johnson, G. Kocjan, D. Poller, and T. Stephenson, "Guidance on the reporting of thyroid cytology specimens," Tech. Rep. 2, Royal College of Pathologists, 2016.

[162] D. N. Poller, M. Bongiovanni, and P. Trimboli, "Risk of malignancy in the various categories of the UK Royal College of Pathologists Thy terminology for thyroid FNA cytology: A systematic review and meta-analysis," *Cancer Cytopathology*, vol. 128, no. 1, pp. 36–42, 2020.

[163] J. D. Brierley, M. Giuliani, B. O'Sullivan, B. Rous, and E. V. Eycken, *TNM Classification of Malignant Tumours*. Oxford, UK: Wiley, 9 ed., 2025.

[164] P. Madmaze, "pytesseract: Python-tesseract - A wrapper for Google's Tesseract-OCR. GitHub repository," 2017.

[165] J. Wu, Z. Wang, M. Hong, W. Ji, H. Fu, Y. Xu, M. Xu, and Y. Jin, "Medical SAM adapter: Adapting segment anything model for medical image segmentation," *Medical Image Analysis*, vol. 102, p. 103547, May 2025.

[166] X. Lin, Y. Xiang, L. Zhang, X. Yang, Z. Yan, and L. Yu, "SAMUS: Adapting Segment Anything Model for Clinically-Friendly and Generalizable Ultrasound Image Segmentation," Sept. 2023. arXiv:2309.06824 [cs].