The role of deep learning in structural and functional lung imaging

Joshua Russell Astley

BEng (Hons)

A thesis submitted for the degree of Doctor of Philosophy

Department of Oncology & Metabolism

Faculty of Medicine

The University of Sheffield

October 2022

Abstract

Background: Structural and functional lung imaging are critical components of pulmonary patient care. Image analysis methods, such as image segmentation, applied to structural and functional lung images, have significant benefits for patients with lung pathologies, including the computation of clinical biomarkers. Traditionally, machine learning (ML) approaches, such as clustering, and computational modelling techniques, such as CT-ventilation imaging, have been used for segmentation and synthesis, respectively. Deep learning (DL) has shown promise in medical image analysis tasks, often outperforming alternative methods.

Purpose: To address the hypothesis that DL can outperform conventional ML and classical image analysis methods for the segmentation and synthesis of structural and functional lung imaging via:

i. development and comparison of 3D convolutional neural networks (CNNs) for the segmentation of ventilated lung using hyperpolarised (HP) gas MRI.
ii. development of a generalisable, multi-centre CNN for segmentation of the lung cavity using ^{1}H-MRI.
iii. the proposal of a framework for estimating the lung cavity in the spatial domain of HP gas MRI.
iv. development of a workflow to synthesise HP gas MRI from multi-inflation, non-contrast CT.
v. the proposal of a framework for the synthesis of fully-volumetric HP gas MRI ventilation from a large, diverse dataset of non-contrast, multi-inflation ^{1}H-MRI scans.

Methods:

i. A 3D CNN-based method for the segmentation of ventilated lung using HP gas MRI was developed and CNN parameters, such as architecture, loss function and pre-processing were optimised.
ii. A 3D CNN trained on a multi-acquisition dataset and validated on data from external centres was compared with a 2D alternative for the segmentation of the lung cavity using ^{1}H-MRI.
iii. A dual-channel, multi-modal segmentation framework was compared to single-channel approaches for estimation of the lung cavity in the domain of HP gas MRI.
iv. A hybrid data-driven and model-based approach for the synthesis of HP gas MRI ventilation from CT was compared to approaches utilising DL or computational modelling alone.
v. A physics-constrained, multi-channel framework for the synthesis of fully-volumetric ventilation surrogates from ^{1}H-MRI was validated using five-fold cross-validation and an external test data set .

Results:

i. The 3D CNN, developed via parameterisation experiments, accurately segmented ventilation scans and outperformed conventional ML methods.
ii. The 3D CNN produced more accurate segmentations than its 2D analogues for the segmentation of the lung cavity, exhibiting minimal variation in performance between centres, vendors and acquisitions.
iii. Dual-channel, multi-modal approaches generate significant improvements compared to methods which use a single imaging modality for the estimation of the lung cavity.
iv. The hybrid approach produced synthetic ventilation scans which correlate with HP gas MRI.
v. The physics-constrained, 3D multi-channel synthesis framework outperformed approaches which did not integrate computational modelling, demonstrating generalisability to external data.

Conclusion: DL approaches demonstrate the ability to segment and synthesise lung MRI across a range of modalities and pulmonary pathologies. These methods outperform computational modelling and classical ML approaches, reducing the time required to adequately edit segmentations and improving the modelling of synthetic ventilation, which may facilitate the clinical translation of DL in structural and functional lung imaging.

Acknowledgements

First and foremost, I owe my profound gratitude to my principal supervisor, Dr Bilal Tahir, for his extraordinary commitment and relentless support, of which I am a beneficiary. His passion for research and his unshakable belief in me, have equipped me with the skills required to succeed in an academic research environment and have been instrumental in my development as a research scientist. I also thank him for his conscious efforts to provide training and opportunities, including providing funding to attend several conferences and workshops, in which I was able to connect with fellow researchers during an extraordinarily isolating period. The fruitful friendship we developed was the necessary catalyst for the number of publications contained herein. Furthermore, I would like to express my gratitude and appreciation to my secondary supervisor, Professor Jim Wild, for his guidance and technical support throughout my PhD. I also find inspiration in his hard work and ruthless desire for progress, which I wish to carry forth into my independent research career. He deserves yet further praise, for putting up with my jibes during a miraculous promotional season for the Mighty Reds.

In addition, I would like to thank the whole Polarised Lung and Respiratory Imaging Systems (POLARIS) team for welcoming me to their group. The support I received during my studies is only comparable to the intellect of those with whom I was thankful to share a laboratory. In no particular order, I would like to thank Helen Marshall, Graham Norquay, Guilhem Collier, Andy Swift, Laura Saunders, Jimmy Ball, Paul Hughes, Fung Chan, Neil Stewart, Jemima Pilgrim-Morris, Laurie Smith, Nick Weatherley, James Eaden, Martin Brook, Madhwesha Rao and Alberto Biancardi for their immense contribution to my work and development.

Lastly, I would not have been able to complete this PhD without the unwavering support of my family and friends. I thank my parents for their love and support throughout. My mother for enduring my long, rambling phone calls and my father for his determination and advice which has been invaluable throughout my life. I would also like to thank Elizabeth Waddilove for her advice, encouragement and sunny disposition, without which, my studies would have been much worse.

Joshua R. Astley

Contents

Abstract.....i

Acknowledgements.....ii

Contents.....iii

List of Figures.....viii

List of Tables.....xiii

Abbreviations.....1

Chapter 1 Thesis overview.....2

1.1 Thesis aims and objectives.....2

1.2 Thesis organisation.....3

Chapter 2 Background and theory.....5

2.1 Lung function and disease.....5

2.2 Imaging of the lungs.....6

2.2.1 Structural lung imaging.....7

2.2.2 Contrast-based functional lung imaging.....10

2.2.3 Image registration-based non-contrast functional lung imaging.....15

2.2.4 Image segmentation-based regional functional lung imaging ventilation biomarkers.....20

2.3 Deep learning theory.....22

2.3.1 Artificial neural networks.....23

2.3.2 Convolutional neural networks.....27

2.3.3 Fully convolutional neural networks.....31

2.3.4 Patch-based sampling.....32

2.3.5 Memory and computational constraints.....34

2.4 Deep learning training strategies.....35

2.4.1 Transfer learning.....36

2.4.2 Data augmentation.....36

2.4.3 Regularisation.....37

2.5 Deep learning validation strategies.....38

2.5.1 K-fold cross validation.....38

2.5.2 Stratified cross-validation.....38

2.5.3 Monte-Carlo cross validation.....39

2.5.4 External validation.....40

2.6 Related works.....40

Chapter 3 Deep learning in structural and functional lung image analysis: a review.....44

3.1 Preface.....45

3.1.1 Author contributions.....45

3.2 Introduction 45
3.3 Deep learning theory...46
3.3.1 Artificial neural networks 46
3.3.2 Pre-processing 48
3.4 Validation 51
3.4.1 Evaluation metrics..52
3.4.2 Validation techniques 52
3.5 Methods 53
3.6 Results and discussion 54
3.6.1 Study selection..54
3.6.2 Segmentation 56
3.6.3 Registration 68
3.6.4 Reconstruction 73
3.6.5 Synthesis 74
3.7 Future research directions 79
3.8 Conclusion 80
Chapter 4 Large-scale investigation of deep learning approaches for ventilated lung segmentation using multi-nuclear hyperpolarised gas MRI...81
4.1 Preface..82
4.1.1 Author contributions 83
4.2 Introduction 83
4.3 Materials and Methods...86
4.3.1 Hyperpolarised gas MRI acquisition 86
4.3.2 Dataset...86
4.3.3 Parameterisation experiments 89
4.3.4 Convolutional neural network...96
4.3.5 DL experimental methods 98
4.3.6 Comparison to conventional methods...98
4.3.7 Evaluation metrics...99
4.3.8 Statistical analysis...100
4.4 Results 101
4.4.1 Qualitative results 101
4.4.2 Quantitative evaluation 101
4.4.3 Validation on 2D ^{3}He hyperpolarised gas MRI scans 105
4.4.4 Ventilated volume 106
4.4.5 Comparison to machine learning approaches 107
4.5 Discussion...107
4.6 Conclusion 111

Chapter 5 Implementable deep learning for multi-sequence proton MRI lung segmentation: a multi-centre, multi-vendor, and multi-disease study 112

5.1 Preface......113

5.1.1 Author contributions......113

5.2 Introduction......113

5.3 Materials and methods......115

5.3.1 Patient data......115

5.3.2 ^{1}H-MRI acquisition......116

5.3.3 ^{1}H-MRI segmentations......117

5.3.4 Convolutional neural network......119

5.3.5 Conventional approach: Spatial fuzzy c-means......120

5.3.6 Quantitative evaluation......121

5.3.7 Statistical analysis......122

5.4 Results......123

5.4.1 Qualitative evaluation......123

5.4.2 Centre 1 evaluation......125

5.4.3 External Validation......132

5.5 Discussion......134

5.5.1 Limitations......136

5.6 Conclusion......138

Chapter 6 A dual-channel deep learning approach for lung cavity estimation from hyperpolarised gas and proton MRI......139

6.1 Preface......140

6.1.1 Author contributions......140

6.2 Introduction......141

6.3 Materials and methods......143

6.3.1 Patient data......143

6.3.2 Image acquisition......143

6.3.3 Image quality assessment......144

6.3.4 Lung cavity estimation segmentations......147

6.3.5 Deep learning frameworks......148

6.3.6 Quantitative evaluation......150

6.3.7 Clinical evaluation......151

6.3.8 Statistical analysis......151

6.4 Results......152

6.4.1 Quantitative evaluation......152

6.4.2 Clinical evaluation......154

6.4.3 Image quality assessment......156

6.5 Discussion....158

6.5.1 Limitations....161

6.6 Conclusion....162

Chapter 7 A hybrid model- and deep learning-based framework for functional lung image synthesis from multi-inflation CT and hyperpolarised gas MRI....163

7.1 Preface....164

7.1.1 Author contributions....164

7.2 Introduction....165

7.3 Materials and Methods....167

7.3.1 Dataset....167

7.3.2 Image acquisition....168

7.3.3 Image segmentation....171

7.3.4 Image registration....171

7.3.5 CT ventilation modelling....172

7.4 Deep learning experiments and evaluation....173

7.4.1 CNN architecture configurations....173

7.4.2 CNN training parameters....174

7.4.3 Quantitative evaluation....175

7.4.4 Statistical analysis....176

7.5 Results....177

7.5.1 Image registration....177

7.5.2 Qualitative and quantitative evaluation....179

7.5.3 VLP evaluation....182

7.6 Discussion....183

7.7 Conclusion....188

Chapter 8 PhysVENeT: A physics-informed deep learning-based framework for the synthesis of 3D hyperpolarised gas MRI ventilation....189

8.1 Preface....190

8.1.1 Author contributions....190

8.2 Introduction....191

8.3 Materials and methods....193

8.3.1 Dataset....193

8.3.2 Image acquisition....194

8.3.3 Image segmentation....194

8.3.4 Image registration....194

8.3.5 ^{1}H-MRI ventilation modelling....195

8.4 Deep learning experiments and evaluation....196

8.4.1 CNN architecture configurations....196

8.4.2 CNN training parameters 198
8.4.3 Data split 198
8.4.4 Quantitative evaluation 200
8.4.5 Statistical analysis 201
8.5 Results 201
8.5.1 Qualitative evaluation 201
8.5.2 Quantitative evaluation 203
8.5.3 External validation 206
8.6 Discussion 207
8.6.1 Limitations 210
8.7 Conclusion 211
Chapter 9 Novel contributions and future research directions 212
9.1 Novel contributions 212
9.2 Potential clinical applications 214
9.2.1 Segmentation 214
9.2.2 Synthesis 215
9.3 Future research directions 216
9.3.1 Multi-centre evaluation and federated learning 216
9.3.2 Unsupervised learning 217
9.3.3 Vision transformers 218
9.3.4 Demystifying deep learning 219
9.3.5 Ethics of AI: opinions on medical applications 220
Appendix A Computation of evaluation metrics 223
A.1 Segmentation evaluation metrics 223
A.1.1 Code 223
A.2 Synthesis evaluation metrics 228
A.2.1 Code 229
Appendix B Publications resulting from this thesis 234
B.1 Accepted Journal articles 234
B.2 Invited review articles 235
B.3 Invited talks 235
B.4 Conference proceedings 236
Bibliography 240

List of Figures

Figure 2.1 Example coronal slice of a non-contrast CT scan from a lung cancer patient. 7

Figure 2.2 3D coronal slice of a spoiled gradient-echo (SPGR) lung MRI scan for a patient with cystic fibrosis acquired using a 1.5T GE HDx scanner with an 8-element cardiac coil at an isotropic resolution of 3mm^3. A TR/TE of 1.8/0.7 milliseconds with a flip angle of 3° and bandwidth of ±166.6kHz was used. 9

Figure 2.3 3D coronal slice of a spoiled gradient-echo (SPGR) lung MRI scan for a patient with Covid-19 acquired using a 3T Philips Ingenia scanner with a body coil at a resolution of 2x2x5mm^3. A TR/TE of 1.9/0.6 milliseconds with a flip angle of 3° and bandwidth of ±321.4kHz was used. 9

Figure 2.4 3D coronal slice of an ultrashort echo time (UTE) lung MRI scan of a patient with interstitial lung disease acquired using a 1.5T GE HDx scanner with an 8-element cardiac coil at an isotropic resolution of 1.5mm^3. A TR/TE of 2.8/0.078 milliseconds with a flip angle of 4° and bandwidth of ±125kHz was used.....10

Figure 2.5 ^{99m}Tc-diethylenetriaminepentaacetic acid (^{99m}Tc DTPA) free-breathing SPECT ventilation scan from a lung cancer patient. 11

Figure 2.6 ^{68}Ga-aerosol (Galligas) free-breathing PET ventilation scan from a lung cancer patient. . 12

Figure 2.7 3D coronal balanced steady-state free precession (bSSFP) hyperpolarised ^{129}Xe lung MRI for a lung cancer patient acquired using a 1.5T GE HDx scanner with a flexible quadrature radiofrequency coil for transmission and reception of MR signals at the Larmor frequency of ^{129}Xe at a resolution of 4x4x10mm^3. A TR/TE of 6.7/2.2 milliseconds with a flip angle of 9° and bandwidth of ±8kHz was used. ^{129}Xe was polarised on site to approximately 25% by using an in-house developed rubidium spin-exchange polariser.................14

Figure 2.8 3D coronal balanced steady-state free precession (bSSFP) hyperpolarised ^{3}He lung MRI for a lung cancer patient acquired using a 1.5T GE HDx scanner with a flexible quadrature radiofrequency coil for transmission and reception of MR signals at the Larmor frequency of ^{3}He at a resolution of 4x4x5mm^3. A TR/TE of 1.9/0.6 milliseconds with a flip angle of 10° and bandwidth of ±166.6kHz was used. ^{3}He was polarised on site to approximately 25% by using an in-house developed rubidium spin-exchange polariser.................15

Figure 2.9 Coronal slice of a CT^{HU} ventilation scan at expiratory geometry generated from inspiratory and expiratory breath-hold CT scans for a patient with lung cancer. 18

Figure 2.10 Coronal slice of ^{1}H-MRI ventilation scan at expiratory geometry generated from inspiratory and expiratory breath-hold isotropic SPGR ^{1}H-MRI scans. 19

Figure 2.11 Comparison of traditional machine learning and DL algorithms for a lung segmentation task, contrasting the hand-crafted feature selection and the end-to-end approach of machine learning and DL, respectively...23

Figure 2.12 Common activation functions, including the Sigmoid, ReLU and Leaky ReLU functions. The example Leaky ReLU employs an α of 0.025. 25

Figure 2.13 Hyperpolarised gas MRI scan with example overlapping and non-overlapping patch-based sampling methods for a patch size of 64 × 64 pixels. 33

Figure 2.14 Diagram of underfitting, overfitting and optimal performance of a classification algorithm. 35

Figure 2.15 Visual representation of K-fold cross-validation where K=5. 39

Figure 2.16 Visual representation of Monte-Carlo cross-validation for N number of iterations with data splits of 80/20% and 75/25%. 40

Figure 3.1 Simplified diagrams of the processes of forward propagation (left) and backpropagation (right) for a neural network with two hidden layers. The neural network is represented as a series of nodes, each of which contains a weight and bias. The weight and bias are combined using the activation function to produce an activation that impacts the strength of connections within the network. Once an input has been passed through the network, it is compared to a desired output, such as an expert segmentation of an anatomical region of interest, to produce a loss. This loss is used to propagate changes to weights and biases, hence, changing the strength of connections for the subsequent example..48

Figure 3.2 Illustration of three common types of deep learning architectures used in medical imaging: a) convolutional neural network (CNN), b) recurrent neural network (RNN) and c) generative adversarial network (GAN). In the lung image analysis examples given, the CNN and RNN are used for image segmentation while the GAN is used for image synthesis..49

Figure 3.3 The search strategy used on Scopus, Web of Science and PubMed to identify relevant studies for inclusion in the review. Further studies that met the selection criteria were identified by handsearching references and through the authors' input. 54

Figure 3.4 PRISMA flowchart of studies identified, screened, assessed for eligibility and included in the literature review analysis. 55

Figure 3.5 Graphical overview of the number of studies per year for the four image analysis applications considered in this review. 2020 values calculated up to 1 April 2020. 55

Figure 3.6 Graphical overview of deep learning lung image analysis studies reviewed by a) disease present in patient cohorts, b) imaging modality and c) architecture. Absolute numbers of papers are provided in a) and b). 56

Figure 3.7 Example images from the authors' own work using deep learning for hyperpolarised gas MRI segmentation. The ^{129}Xe MR ventilation images are taken from three subjects in a testing set, a healthy volunteer, asthma patient and cystic fibrosis patient. The patient images selected are characterised by significant ventilation defects. These are compared to expert segmentations of the same image. Dice similarity coefficient (DSC) values are displayed for all images. 68

Figure 4.1 Results for the network architecture parameterisation experiments showing performance at (a) intervals of 5000 iterations and (b) 30000 iterations for four 3D CNNs in terms of DSC (left) and Avg HD (right). 91

Figure 4.2 Results for both the nn-UNet and VNet architectures using three common loss functions, namely, the binary cross-entropy loss, dice loss and Tversky loss evaluated with a) DSC and b) Avg HD metrics at intervals of 5000 iterations up to 30000 iterations. 92

Figure 4.3 a) Results for both the nn-UNet and VNet architectures using common pre-processing strategies evaluated with DSC and Avg HD at intervals of 5000 iterations up to 30000 iterations. b) Results for pre-processing experiments comparing performance at 30000 iterations in terms of DSC. 94

Figure 4.4 Comparison of performance for the nn-UNet denoised and VNet no pre-processing models using Avg HD (top) and DSC (bottom) at intervals of 5000 iterations up to 30000 iterations.........................95

Figure 4.5 Visual representation of the modified 3D nn-UNet network used in this work. The deconvolution side of the network is omitted as it follows the same structure as the convolutional path, however, with the addition of a 1x1x1 SoftMax layer. 97

Figure 4.6 Example coronal slices for a healthy subject and four cases with different pathologies for each DL experimental method. Individual, and median (range), DSC values are displayed........................102

Figure 4.7 Comparison of segmentation performance on 86 testing scans for five DL experimental methods using the DSC, Avg HD, HD95 and XOR metrics (left to right). P-values are displayed for Friedman tests with Bonferroni correction for multiple comparisons, comparing the combined ^{3}He and ^{129}Xe DL method to the other DL methods. Mean and standard deviation values are marked by a bold line and whiskers, respectively. 103

Figure 4.8 Comparison of DSC (top) and Avg HD (bottom) values for ^{129}Xe and ^{3}He testing scans for five DL methods. P-values between ^{129}Xe and ^{3}He using Mann-Whitney tests are shown. Mean and standard deviation values are marked by a bold line and whiskers, respectively. 104

Figure 4.9 Example coronal slices for an asthma and CF case with ground truth segmentations and the segmentations generated by the combined ^{3}He and ^{129}Xe DL model. DSC values are displayed 106

Figure 4.10 Pearson correlation and Bland-Altman analysis of lung volumes for 86 testing set cases compared to volumes derived from expert segmentations for the combined ^{3}He and ^{129}Xe DL. 106

Figure 4.11 Comparison of performance on testing scans between the combined ^{129}Xe and ^{3}He DL method and conventional segmentation methods (SFCM and K-means) with P-values for Friedman tests with Bonferroni correction for multiple comparisons. Mean and standard deviation values are marked by a bold line and whiskers, respectively. Individual DSC and Avg HD values for each method are displayed for three cases. 107

Figure 5.1 Example coronal slices showing the ^{1}H-MRI scans (row 1) and the ^{1}H-MRI scans overlaid with manual segmentations (row 2) and segmentations generated by the 3D UNet, 2D UNet and SFCM methods (rows 3-5) for six cases. DSC and Avg HD values are provided for each case. Example slices were left uncropped to display differences in field of view and arm position between acquisitions. 124

Figure 5.2 Example coronal slices showing ^{1}H-MRI scans that exhibit challenging features such as artifacts, ground glass opacities, consolidation, and bronchiectasis for five cases with corresponding expert, deep learning, and SFCM segmentations. DSC values are provided for each case and method 125

Figure 5.3 Comparison of segmentation performance for each of the methods using the a) DSC, b) Avg HD, and c) XOR metrics. Significances of differences between DL methods and SFCM as assessed by Friedman tests with Bonferroni correction are displayed for each metric. 127

Figure 5.4 (Top) 2D UNet and (bottom) 3D nn-UNet training performance and convergence. 129

Figure 5.5 Comparison of segmentation performance across acquisition protocols for the DSC, Avg HD and XOR metrics for a) 3D UNet, b) 2D UNet and c) SFCM methods. Significant differences between image acquisitions were assessed by Kruskal-Wallis tests are given for each metric. 131

Figure 5.6 Comparison of segmentation performance on the combined external validation datasets for each of the methods using the DSC (left), Avg HD (centre) and XOR (right) metrics. Significances of differences between DL methods and SFCM as assessed by Friedman tests with Bonferroni correction for multiple comparisons are displayed for each metric. 132

Figure 5.7 Bland-Altman agreement analysis of lung volumes for 164 external validation set cases compared to volumes derived from manual segmentations for a) 3D UNet b) 2D UNet and c) SFCM methods. 132

Figure 5.8 Comparison of the combined external validation datasets stratified by inspiratory and expiratory scans using the DSC, Avg HD and XOR metrics for a) 3D UNet, b) 2D UNet and c) SFCM methods. P-values between inspiratory and expiratory scans are shown 133

Figure 6.1 ^{129}Xe-MRI signal (green) and background noise (red) delineations. 146

Figure 6.2 ^{1}H-MRI signal (green) and background noise (red) delineations. 146

Figure 6.3 Illustration showing the motivation to generate lung cavity estimations in the spatial domain of ^{129}Xe-MRI due to misalignments in image acquisitions between modalities. Example cases demonstrating misalignments between ^{129}Xe- and ^{1}H-MRI. Misalignments are indicated by green arrows. 147

Figure 6.4 From left to right: ventilation-only, structural-only, and dual-input deep learning workflows. 149

Figure 6.5 Example coronal slices showing the ^{1}H-MRI and corresponding similar-breath hyperpolarised gas MRI scan with the expert LCE and the LCE generated from the three DL methods for four cases within the testing set. Expert hyperpolarised gas MRI segmentations are provided for each case to aid visualisation of alignments. DSC values are provided for each case. 153

Figure 6.6 a) Quantitative results for the testing set (n=58) using the DSC, Avg HD (mm) and XOR metrics for the ventilation-only, structural-only, and dual-input DL methods. Median (range) values are given with the best values shown in bold. b) Comparison of LCE performance for each of the three DL methods using the DSC (left), Avg HD (centre) and XOR (right) metrics. Significance of differences between DL methods as assessed by Friedman tests with Bonferroni correction for multiple comparisons are displayed for each metric. 154

Figure 6.7 Bland-Altman analysis (left) and Pearson correlation (right) of lung volumes for 58 testing set cases comparing the manual LCEs to the dual-input DL-generated LCEs. 155

Figure 6.8 Example coronal slices of four cases with ventilation defects showing fused manual LCEs (white) and hyperpolarised gas MRI ventilated lung segmentations (pink) compared to those generated using the dual-input DL method and previously described hyperpolarised gas MRI ventilated lung segmentation method. Manual and DL-generated VDPs are given for each case. 155

Figure 6.9 a) Estimation plot of manual- and DL-generated VDPs (left) with significance of differences; b) Bland-Altman analysis (right) of VDPs for 58 testing set cases comparing the manual LCE to the dual-input DL-method. 156

Figure 6.10 a) Absolute differences between manual and DL VDPs stratified by presence (or absence) of image artifacts in the ^{1}H-MRI (left) and ^{129}Xe-MRI (right) scans. Mann-Whitney U tests were conducted, and p values indicated. b) Scatterplot of absolute differences between manual and DL VDPs and SNRs for the ^{1}H-MRI (left) and ^{129}Xe-MRI (right) scans. Spearman's ρ values are provided. 157

Figure 6.11 Example coronal slices of fused ^{1}H-MRI, ^{129}Xe-MRI and LCEs from three cases in the testing set which, through Bland-Altman analysis, fall outside the limits of agreement. Manual and DL-generated VDPs are given. Case 1 contains a motion artefact on the ^{129}Xe-MRI. Case 2 contains a zipper artefact on the ^{1}H-MRI. Case 3 exhibits a large degree of noise throughout the ^{129}Xe-MRI. Artifacts are indicated with green arrows. 158

Figure 7.1 Example coronal slices for three patients with lung cancer, cystic fibrosis, or asthma of inspiratory and expiratory CT scans (a) before and (b) after deformable registration to the spatial domain of (c) hyperpolarised gas MRI 172

Figure 7.2 Breakdown of cross-validation strategy used for training and testing across 47 patients. 175

Figure 7.3 Hybrid model- and DL-based synthetic ventilation workflow and accompanying automatic calculation of VLP. 177

Figure 7.4 Alignment of internal lung structures between inspiration and expiration CT. Bifurcations are marked with crosshairs as landmarks. 178

Figure 7.5 Example coronal slices from the CT^{HU} model and the four DL frameworks for three cases compared to ^{3}He MRI. Spearman's ρ values between each method and ^{3}He MRI are provided. Red arrows demonstrate examples of accurately replicated defects........179

Figure 7.6 (Top) Spearman's ρ values for synthetic ventilation scans derived from the CT^{HU} model and DL configurations. A paired *t*-test compared CT^{HU} with the hybrid DL configuration. One-way ANOVA tests compared Spearman's ρ values for DL configurations. (Bottom) MSE values for synthetic ventilation scans derived from DL configurations. One-way ANOVA tests compared MSE values for DL configurations. Only significant p-values are provided. 181

Figure 7.7 Fused ventilation (jet colormap showing minimum to maximum ventilation) and structural scans (grayscale) from four patients derived from either ^{3}He MRI and warped expiratory CT (top) or synthetic ventilation generated using the proposed hybrid model/DL approach and warped expiratory CT (bottom). Red arrows indicate defects replicated in synthetic ventilation scans. VLP values are given........182

Figure 7.8 Comparison of VLPs derived from hyperpolarised gas MRI versus a) the hybrid model/DL configuration and b) the CT^{HU} ventilation model using Bland-Altman analysis. 183

Figure 8.1 Registration workflow for modelling ^{1}H-MRI ventilation. 196

Figure 8.2 PhysVENeT architecture and training strategy. 199

Figure 8.3 Example coronal slices of TLC and RV ^{1}H-MRI, ^{129}Xe-MRI, DL-based synthetic ventilation maps and the ^{1}H-MRI SV model for five participants in the dataset. Voxel-wise Spearman's *rs* and SSIM values are given for each DL approach and the ^{1}H-MRI ventilation model. Green arrows indicate defects which are present in hyperpolarised gas MRI and replicated in synthetic ventilation scans........202

Figure 8.4 Comparison of performance for DL methods and ^{1}H-MRI SV model using the voxel-wise Spearman's *rs* (left) and SSIM (right) metrics. P-values are given for statistically significant comparisons..203

Figure 8.5 Comparison of performance stratified by participant pathology using Spearman's rs (left) and SSIM (right) metrics for the proposed PhysVENeT framework 204

Figure 8.6 Comparison of performance on external validation data using the five trained models generated by the PhysVENeT during cross-validation in terms of Spearman's *rs* (left) and SSIM (right). Significant p-values are given........207

List of Tables

Table 3.1 Glossary of key technical terms related to deep learning and image analysis. 47

Table 3.2 Summary of common pre-processing techniques used for lung image analysis tasks, including values prevalent in the literature. Modalities included are those for which the pre-processing techniques have been used in the reviewed studies. This is not an exhaustive list of pre-processing techniques used. 50

Table 3.3 Summary of reviewed studies on deep learning for lung image segmentation. Arranged alphabetically by 'Anatomical site', then by 'Modality'. 57

Table 3.4 Summary of reviewed studies using deep learning for lung image registration. 70

Table 3.5 Summary of reviewed studies using deep learning for lung image reconstruction............76

Table 3.6 Summary of reviewed studies using deep learning for lung image synthesis. 77

Table 4.1 Summary of demographics, clinical characteristics and image dataset information stratified by disease. 88

Table 4.2 Comparison of segmentation performance for the five DL training methods for all scans in the testing set. Medians (ranges) are given; the best result for each metric is in bold. 104

Table 4.3 Summary of 2D spoiled gradient-echo results. 105

Table 5.1 Summary of patient data. 116

Table 5.2 ^{1}H-MRI acquisition details. 118

Table 5.3 Breakdown of training and testing strategy with external validation 121

Table 5.4 Quantitative results for the testing set (n=50), external validation 1 (n=110) and external validation 2 (n=54) using the DSC, Avg HD (mm) and XOR metrics for the SFCM, 2D UNet and 3D UNet methods. Median (range) values are provided for each acquisition protocol, the combined testing set, and the external validation sets. 128

Table 6.1 Summary of patient demographic data. 143

Table 7.1 CT and hyperpolarised gas MRI acquisition details. 170

Table 7.2 Evaluation of overlap between inspiratory and expiratory CT after two-step registration process. Median DSC and range are given for the three studies comprising the data used in this work......178

Table 7.3 Descriptive statistics for the CT^{HU} model and DL methods after combining the testing set performance via 6-fold cross-validation. Mean±SD Spearman's ρ for the DL methods and the CT^{HU} model are shown. Additionally, mean±SD MSE are given for the DL methods. The best ρ and MSE values are shown in bold. 180

Table 7.4 Median (range) DSC of the hybrid DL configuration and CT^{HU} ventilation model for ventilated and defect lung regions. The best DSC values are shown in bold. 183

Table 8.1 Summary of patient demographic data. 193

Table 8.2 Synthetic ventilation results from the ^{1}H-MRI SV model and the three DL approaches compared to hyperpolarised gas MRI ventilation using the Spearman's rs and SSIM metrics. Median (range) is given. Metrics are given for each fold individually and the combined values across all folds. 205

Table 8.3 Synthetic ventilation results on the external validation dataset (n=20) from the three DL approaches compared to ^{129}Xe-MRI ventilation using the Spearman's rs and SSIM metrics. Median (range) is given. Metrics are given for ventilation surrogates generated by each of the five folds during cross-validation and the average values across all folds. 206

Abbreviations

^{129}Xe	Xenon-129
^{1}H	Proton
2D	Two Dimensional
3D	Three Dimensional
^{3}He	Helium-3
ANN	Artificial Neural Network
CF	Cystic Fibrosis
CNN	Convolutional Neural Network
COPD	Chronic Obstructive Pulmonary Disease
CT	Computed Tomography
CTVI	Computed Tomography Ventilation Imaging
DL	Deep Learning
DSC	Dice Similarity Coefficient
FRC	Functional Residual Capacity
GAN	Generative Adversarial Network
HD	Hausdorff Distance
HD95	Hausdorff Distance 95th Percentile
HP	Hyperpolarised
HU	Hounsfield Units
LCE	Lung Cavity Estimation
ML	Machine Learning
MRI	Magnetic Resonance Imaging
MSE	Mean Square Error
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
RV	Residual Volume
SFCM	Spatial Fuzzy C-Means
SSIM	Structural Similarity Index Measure
TLC	Total Lung Capacity
VDP	Ventilation Defect Percentage
XOR	Relative Error Metric

Chapter 1
Thesis overview

1.1 Thesis aims and objectives

Recent developments in deep learning (DL) have dramatically influenced the medical imaging field. Medical image analysis applications have been at the forefront of DL research efforts for many diseases and anatomical sites, including the lungs. Artificial neural networks have been leveraged for image analysis applications such as image segmentation and synthesis. The POLARIS Lung Imaging Centre at The University of Sheffield provides a hyperpolarised gas MRI clinical referral service, the first of its kind in the world, which requires specialised equipment and significant resources for image acquisition and analysis. For example, currently used algorithms for the segmentation of structural and functional lung images are only semi-automatic and, thus, require substantial time to manually correct generated outputs. Therefore, to increase clinical throughput, there is a pressing need to eliminate, or reduce, the time taken to perform these lung image analysis tasks. In addition, surrogates of regional lung function have been proposed which are derived from structural imaging without exogenous contrast; however, these modelling approaches show large variability in performance. Consequently, the ability to generate surrogates of regional lung function from structural imaging would drastically increase wider clinical adoption. Accordingly, the central aim of this thesis is to address the hypothesis that:

> *Artificial neural networks can outperform conventional machine learning and classical image analysis methods for the segmentation and synthesis of structural and functional lung imaging.*

To test this hypothesis, a series of inter-related investigations that have the following proposed objectives were performed:

1. Perform a systematic review of the DL literature related to structural and functional lung imaging to identify gaps in the literature, providing direction for the remainder of the thesis.

2. Build, validate and evaluate several 3D convolutional neural networks (CNNs) for ventilated lung segmentation using multi-nuclear hyperpolarised gas MRI and compare these approaches to currently used methods.

3. Develop a generalisable CNN-based method that is capable of segmenting proton MRI (^{1}H-MRI) robust to image resolution, acquisition sequence and lung pathology and compare this approach to current methods used for ^{1}H-MRI segmentation. In addition, experiments will investigate differences between 2D and 3D CNNs.

4. Use DL to generate lung cavity estimations (LCE) that represent the structural lung parenchyma in the spatial domain of hyperpolarised gas MRI ventilation scans.

5. Develop a novel hybrid framework, integrating model- and DL-based methods for the synthesis of hyperpolarised gas MRI ventilation from non-contrast, multi-inflation CT.

6. Propose a novel DL technique to generate 3D surrogates of hyperpolarised gas MRI ventilation images from non-contrast, multi-inflation ^{1}H-MRI scans across a range of diseases.

1.2 Thesis organisation

A background on lung disease, lung imaging and the theoretical underpinnings of DL, are described in **Chapter 2**. Furthermore, the technical knowledge required for the remainder of the thesis, including CNNs, loss functions, hyperparameters and experimental methodologies are described in further detail therein.

In **Chapter 3**, a systematic literature review focusing on DL in pulmonary image analysis was conducted. This review focuses specifically on segmentation, registration, reconstruction and synthesis applications of DL in lung imaging across a range of imaging modalities.

Chapter 4 details the development and comprehensive evaluation of several 3D CNNs for ventilated lung segmentation using multi-nuclear hyperpolarised gas MRI. A series of parameterisation experiments were conducted to determine the network architecture, loss function and the impact of pre-processing on segmentations generated by the network.

Building upon methods used in Chapter 4, **Chapter 5** focuses on the development of a generalisable CNN for lung segmentation in ^{1}H-MRI; the multi-sequence dataset used contains scans from multiple centres acquired at different field strengths and resolutions from a range of diseases.

Chapter 6 builds upon the methods developed in the previous two chapters to generate a novel approach for segmenting LCEs. Clinical biomarkers of lung function, such as the ventilation defect percentage (VDP), require the segmentation of both structural ^{1}H-MRI and hyperpolarised gas MRI ventilation scans. To this end, a dual-channel CNN that integrates functional and structural imaging to generate LCEs is proposed.

Unlike previous chapters, the following two chapters focus on image synthesis applications in functional lung imaging. **Chapter 7** demonstrates the development of a novel hybrid model- and DL-based framework for functional lung image synthesis using hyperpolarised gas MRI. The method generates 3D ventilation surrogates from non-contrast, multi-inflation CT with the aim of generating synthetic hyperpolarised gas MRI ventilation scans without the requirement for specialised equipment or exogenous contrast. In addition, we leverage the model in Chapter 4 to provide ventilated lung segmentations for the automatic calculation of VDP.

Similar ideas developed for multi-inflation CT synthesis in Chapter 7 are adapted and applied to multi-inflation ^{1}H-MRI synthesis in **Chapter 8.** The proposed framework uses a multi-channel CNN to generate hyperpolarised gas MRI ventilation surrogate images from 3D multi-inflation ^{1}H-MRI without the requirement for specialised equipment, exogenous contrast or exposure to ionising radiation. This investigation utilises a larger dataset than previous synthesis applications with the inclusion of external validation data, thus increasing the generalisability of the proposed approach.

The novel contributions, potential clinical applications and future research directions of this work are discussed in **Chapter 9**.

Chapter 2
Background and theory

2.1 Lung function and disease

The primary function of the lungs is gas exchange, whereby oxygen is inhaled, and carbon dioxide exhaled, during breathing; the pulmonary airways transport air to the alveoli where the gas exchange surface of the alveoli transfers fresh air across the blood-gas barrier delivering it to the pulmonary capillaries. This fresh gas then enters the pulmonary vascular system via two specific physiological functions, namely, ventilation and perfusion. Lung ventilation refers to the lungs' ability to deliver fresh gas to the alveoli. Lung perfusion refers to the lungs' ability to transfer gas across the blood-gas barrier, delivering oxygenated blood throughout the body. The physiology of the lungs, including the lung microstructure, facilitates lung ventilation and perfusion and is, therefore, key to the healthy and efficient functioning of the respiratory system. In healthy individuals, the lungs are ideally suited for their primary function; however, for individuals whose lungs are impaired, ventilation and perfusion are diminished. Individual alveoli have variable degrees of regional ventilation and perfusion, both in healthy individuals due to gravitational effects and in patients with respiratory diseases due to impairments induced by these diseases. Lung diseases can be broadly categorised as restrictive or obstructive. Restrictive lung diseases result in difficulty inhaling fresh gas, whereas obstructive lung diseases result in difficulty exhaling gases. Idiopathic pulmonary fibrosis and other interstitial lung diseases are examples of restrictive lung diseases which cause shortness of breath and lung inflammation. Obstructive lung diseases are significantly more common than restrictive diseases with substantial prevalence globally; 65 million people suffer from chronic obstructive pulmonary disease (COPD) and 339 million from asthma worldwide (GBD15 *et al.*, 2016; Vos *et al.*, 2017). Asthma results in constricted and inflamed airways, leading to an obstruction of airflow; it is believed that asthma has various causes, including genetic factors. COPD is a progressive lung disease most commonly caused by smoking, resulting in persistent coughing, shortness of breath and a tightening in the chest. COPD may lead to emphysema which is categorised by structural changes in the lungs, such as the destruction of alveoli sacs. Furthermore, inherited genetic conditions such as cystic fibrosis (CF) cause obstructive lung disease via increased mucus plugging which blocks the secretion of digestive enzymes. Cystic fibrosis is relatively uncommon; however, it results in a significant reduction in lifespan and is, therefore, an area of continued research. Unlike the aforementioned lung diseases, lung cancer is categorised as neither restrictive nor obstructive. There are 1.8 million new lung cancer cases diagnosed annually and 1.6 million deaths worldwide, making it the most common and deadliest cancer on the planet (Torre *et al.*, 2015). Approximately 40-70% of lung cancer patients have comorbidities, predominantly COPD (Congleton and Muers, 1995); in addition, lung cancer treatments such as radiotherapy can cause restrictive lung pneumonitis and fibrosis, known as radiation-induced lung disease (RILD) (Hanania *et al.*, 2019). Additionally, respiratory infections can lead to a number of restrictive lung diseases; the populous will now be familiar with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), commonly known as Covid-19, a viral respiratory infection which can cause restrictive lung diseases such as pneumonia. Lung disease is commonly diagnosed through pulmonary function tests (PFTs), whereby an individual's lung function is assessed through spirometry tests and compared to established values from the literature derived from healthy participants. PFTs have been shown to be insensitive to minor changes and the early onset of disease (Kirby *et al.*, 2011; Marshall *et al.*, 2017). Therefore, imaging of the lungs may be important in the diagnosis and treatment of various pulmonary pathologies.

2.2 Imaging of the lungs

Imaging of the lungs is a critical component in the diagnosis, treatment planning, monitoring and assessment of respiratory diseases. Lung imaging can broadly be divided into structural and functional imaging modalities, both of which provide important insights into respiratory diseases. Structural lung imaging allows for the detailed visualisation of the lungs, including the pulmonary parenchyma and vessels, among other clinically useful features. Functional lung imaging, in contrast, provides insight into the function of the lungs, including ventilation, perfusion or gas exchange within the lung parenchyma. The following sections will focus on various imaging modalities that are referenced throughout this thesis, detailing how they are acquired, the insights that can be gleaned from them, and their benefits and drawbacks for clinical lung imaging.

2.2.1 Structural lung imaging

Computed tomography (CT): is the most widely used anatomical imaging modality and is an integral part of clinical care for most patients with lung pathologies. CT scans are generated using X-rays deployed at a multitude of angles around the patient which are then combined and reconstructed to produce highly detailed 3D images of the lungs with resolutions of approximately 1-2mm^3 when high-resolution CT protocols are utilised (Corcoran *et al.*, 1992). This resolution allows for detailed imaging of internal lung structures such as vessels and fissures. Each voxel in a CT scan has a given intensity measured in Hounsfield Units (HUs) which is physiologically determined and consistent between images, facilitating comparison between CT scans acquired at different time points. CT employs ionising radiation which can be harmful to patients; thus, the use of CT is limited in paediatric or repeat scanning applications. CT images are acquired either at multiple breath-holds at different respiratory inflations (inspiration and expiration) or during tidal breathing at various phases (4DCT). Figure 2.1 shows an example CT scan of the lungs for a central 2D slice from a patient with lung cancer.

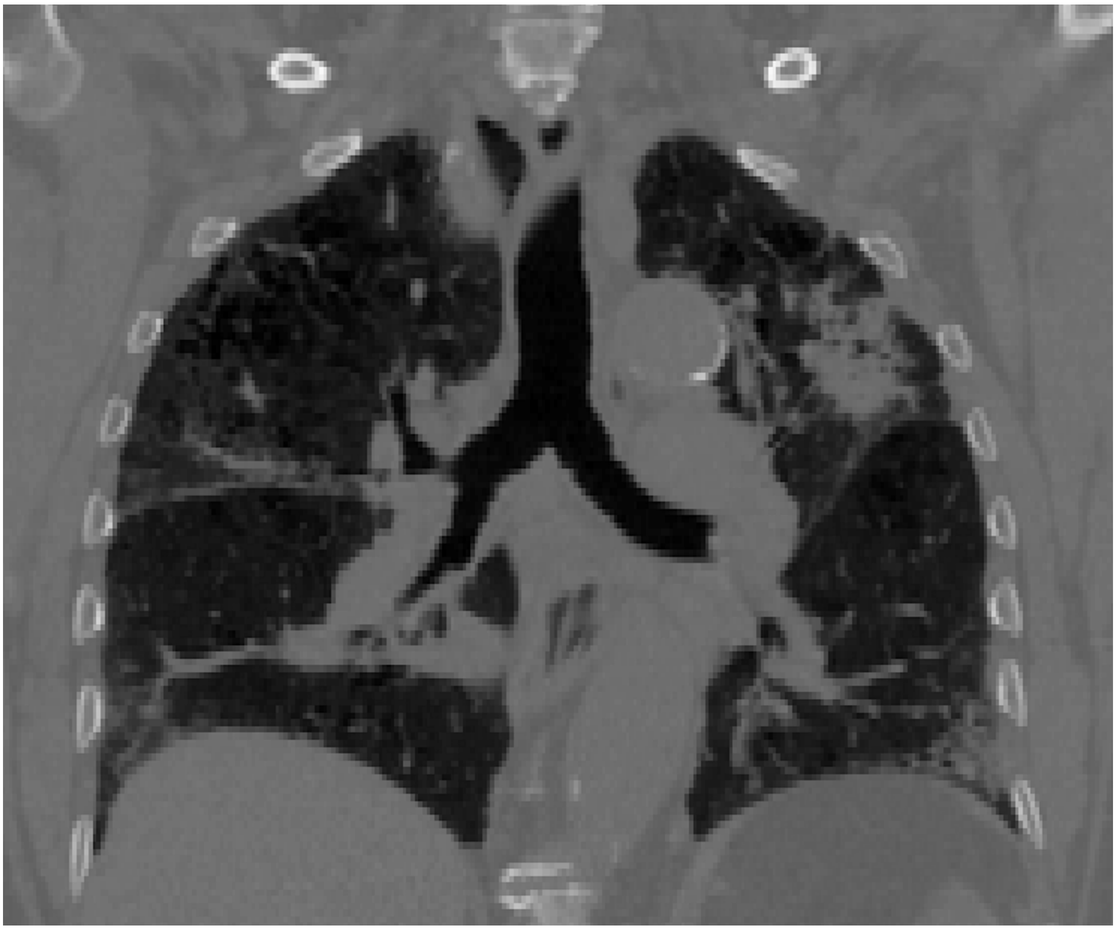


Figure 2.1 Example coronal slice of a non-contrast CT scan from a lung cancer patient.

Proton-magnetic resonance imaging (^{1}H-MRI): in the lungs is traditionally challenging due to the low proton density of the lungs which results in a lack of susceptibility differences between tissue and air, where the proton density is approximately 10 times less than other tissues (Wild *et al.*, 2012). MR signal is proportional to proton density and, thus, the signal in the lungs can be limited. Susceptibility artefacts at lung-air interfaces cause inhomogeneities in the magnetic field, resulting in rapid T_2* decay in gradient-echo imaging. This means that there is a requirement for very short echo times (TEs) in lung MRI compared to imaging of other tissues. Lung MRI can generally be acquired at either 1.5T or 3T depending on clinical requirements, where scans acquired at 3T have a higher signal but much shorter T_2*, further lowering the required TE (Wild *et al.*, 2012). In addition, lung MRI is challenging as the scan time must be minimised to reduce potential motion artefacts during breathing; consequently, MRI is regularly acquired at breath-hold inflations with a scan time on the order of 15 seconds. Unlike CT, MRI employs non-ionising radiation and thus it can be utilised for paediatric and longitudinal scanning. There are four main sequences used for lung MRI which produce scans with different contrasts and resolutions, namely, spoiled gradient-echo (SPGR), balanced steady-state free precession (bSSFP), single-shot fast spin-echo (FSE) and ultrashort echo time (UTE) (Wild *et al.*, 2012). Figure 2.2 depicts an example SPGR ^{1}H-MRI scan acquired at 1.5T; Figure 2.3 depicts an example SPGR ^{1}H-MRI scan acquired at 3T; Figure 2.4 depicts an example UTE ^{1}H-MRI scan acquired at 1.5T. These examples have been chosen as they include MRI scans with various sequences and readout parameters that are used throughout this thesis.

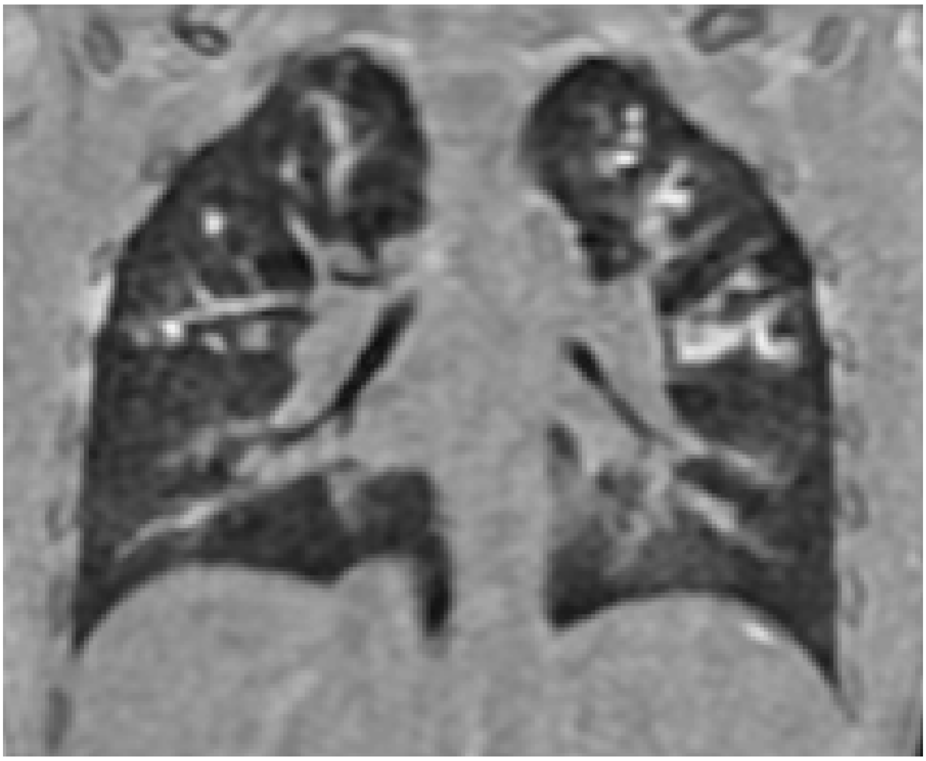


Figure 2.2 3D coronal slice of a spoiled gradient-echo (SPGR) lung MRI scan for a patient with cystic fibrosis acquired using a 1.5T GE HDx scanner with an 8-element cardiac coil at an isotropic resolution of 3mm^3. A TR/TE of 1.8/0.7 milliseconds with a flip angle of 3° and bandwidth of ±166.6kHz was used.

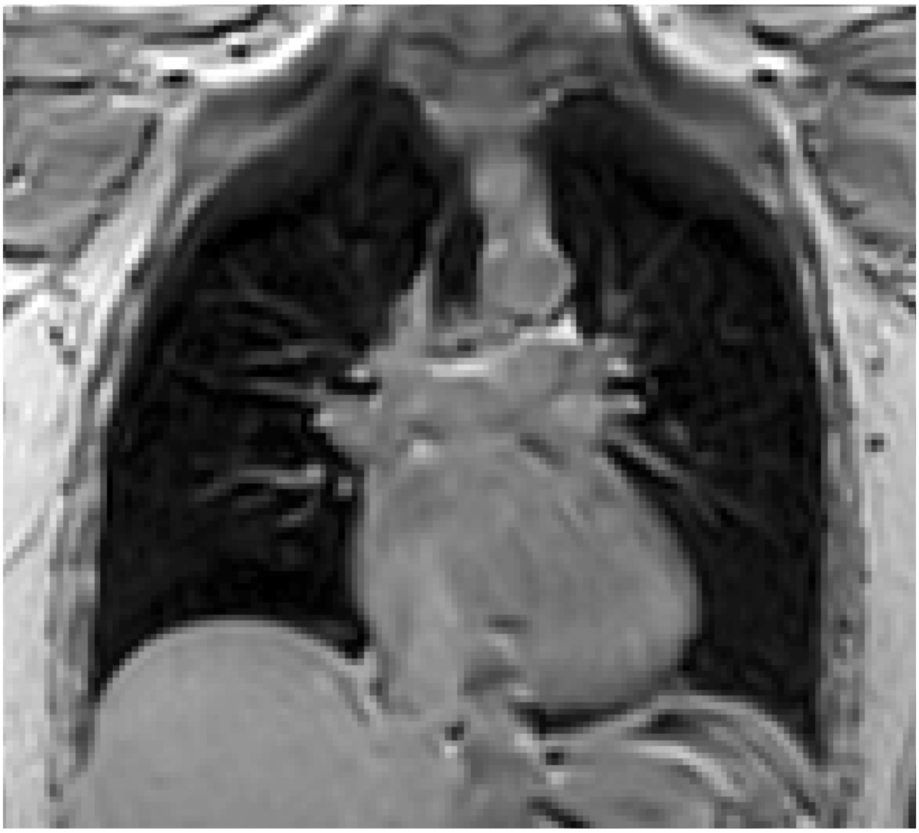


Figure 2.3 3D coronal slice of a spoiled gradient-echo (SPGR) lung MRI scan for a patient with Covid-19 acquired using a 3T Philips Ingenia scanner with a body coil at a resolution of 2x2x5mm^3. A TR/TE of 1.9/0.6 milliseconds with a flip angle of 3° and bandwidth of ±321.4kHz was used.

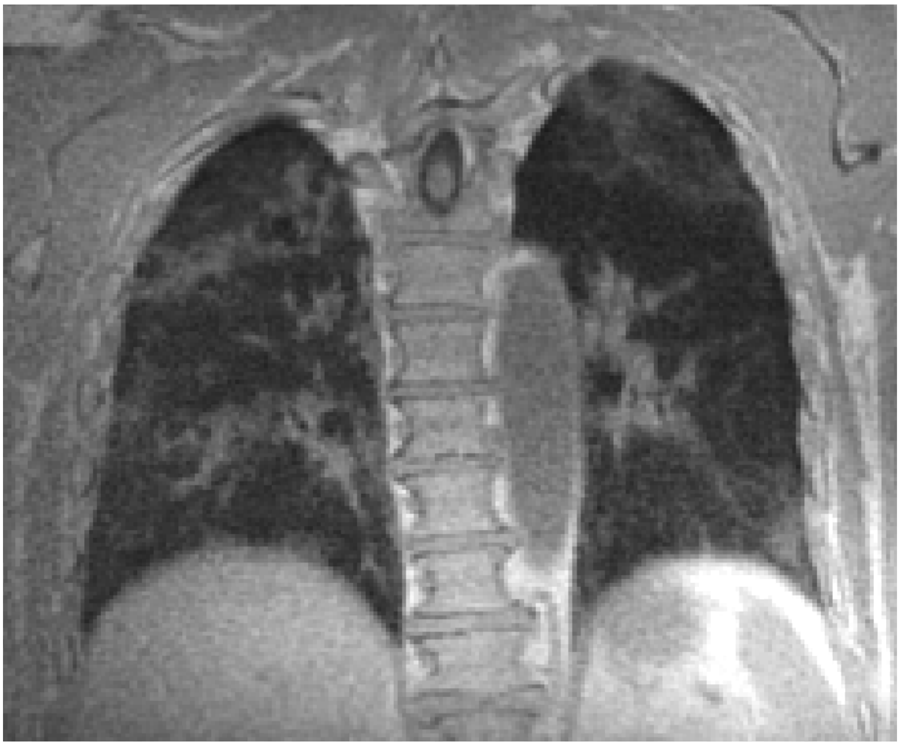


Figure 2.4 3D coronal slice of an ultrashort echo time (UTE) lung MRI scan of a patient with interstitial lung disease acquired using a 1.5T GE HDx scanner with an 8-element cardiac coil at an isotropic resolution of 1.5mm^3. A TR/TE of 2.8/0.078 milliseconds with a flip angle of 4° and bandwidth of ±125kHz was used.

2.2.2 Contrast-based functional lung imaging

Single-photon emission computed tomography (SPECT): is a nuclear medicine imaging modality that uses a radioactive tracer given prior to imaging that emits photons at various angles which are detected by a gamma camera and subsequently reconstructed using the 3D radioactivity distribution. Proton and tissue interactions cause scatter and attenuation leading to blurring; in the lungs, around 10-15% of all photons detected are scattered (Petersson *et al.*, 2007). Attenuation correction is generally employed via an external radiation source allowing the creation of an attenuation correction map used to improve image quality. Spatial resolution is dependent on the energy of the radionucleotide, the type of collimators and the distance between the source and the gamma camera, producing resolutions on the order of 10-20mm^3 (Petersson *et al.*, 2007). Due to the presence of a radioactive tracer, SPECT employs ionising radiation, limiting its application in paediatric patients. Acquisition of SPECT is relatively long with an acquisition time of 10-30 minutes. For SPECT ventilation, an aerosol radioactive tracer is used; commonly these include tracers such as Xenon-123 (^{123}Xe) and Technegas (^{99m}Tc) (Jögi *et al.*, 2010). Technegas can create aerosol deposition artefacts, leading to clumping, particularly in the airways, characterised by high signal in these regions. Figure 2.5 depicts an example coronal slice from a SPECT ventilation scan.

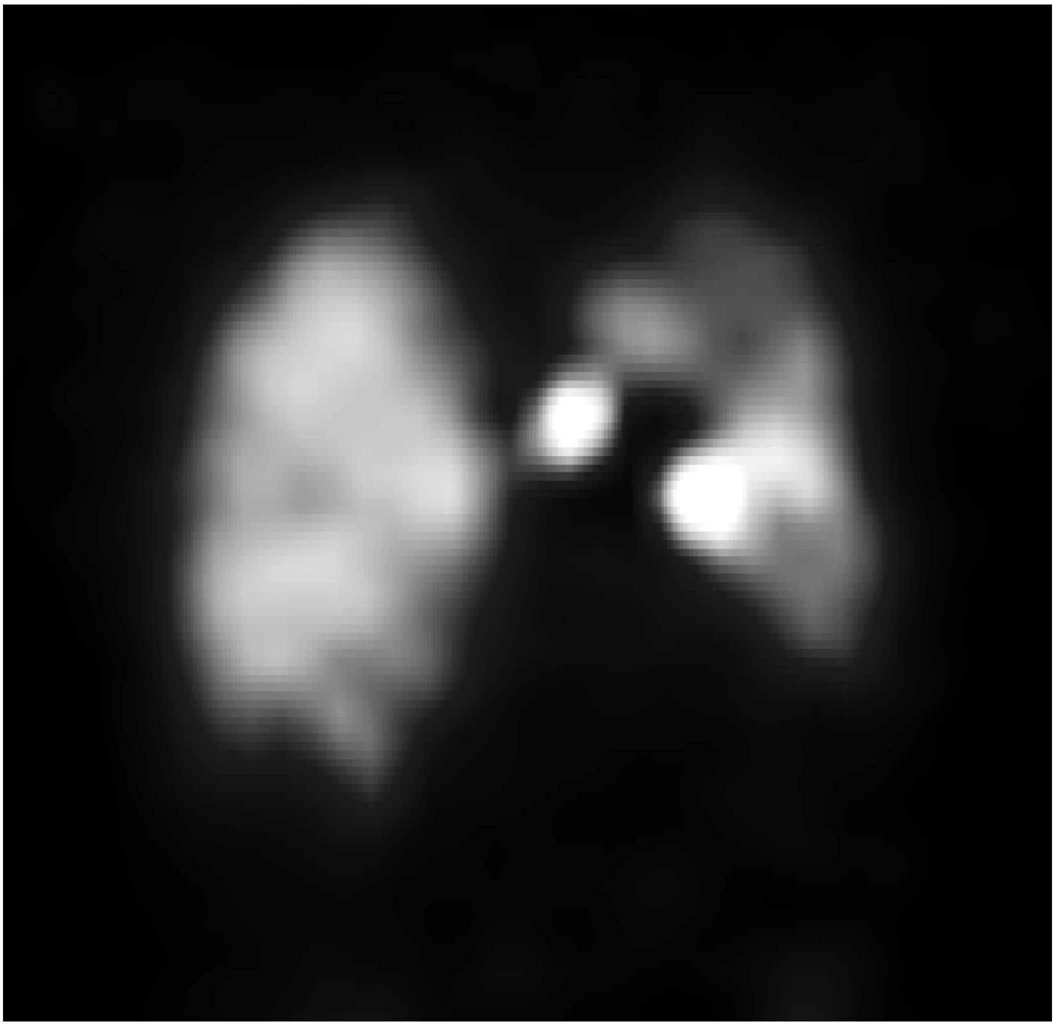


Figure 2.5 ^{99m}Tc-diethylenetriaminepentaacetic acid (^{99m}Tc DTPA) free-breathing SPECT ventilation scan from a lung cancer patient.

Positron emission tomography (PET): is an alternative nuclear medicine imaging modality to SPECT that can directly image metabolism, ventilation or perfusion depending on the radioactive tracer used. A positron is emitted from the tracer which produces a pair of photons when it collides with tissues; these photons can be detected by crystals in the PET scanner to produce an image. ^{68}Ga-aerosol (Galligas) is commonly used as a radioactive tracer for PET ventilation imaging (Le Roux *et al.*, 2019). PET imaging also utilises ionising radiation, reducing its application for repeat or longitudinal scanning. PET scans have a similar acquisition time as SPECT with images acquired over multiple breaths; this can lead to some defects resolving over time due to delayed ventilation filling effects. However, PET has a higher spatial resolution than SPECT on the order of approximately 4-6mm^3. Figure 2.6 depicts an example Galligas PET ventilation scan.

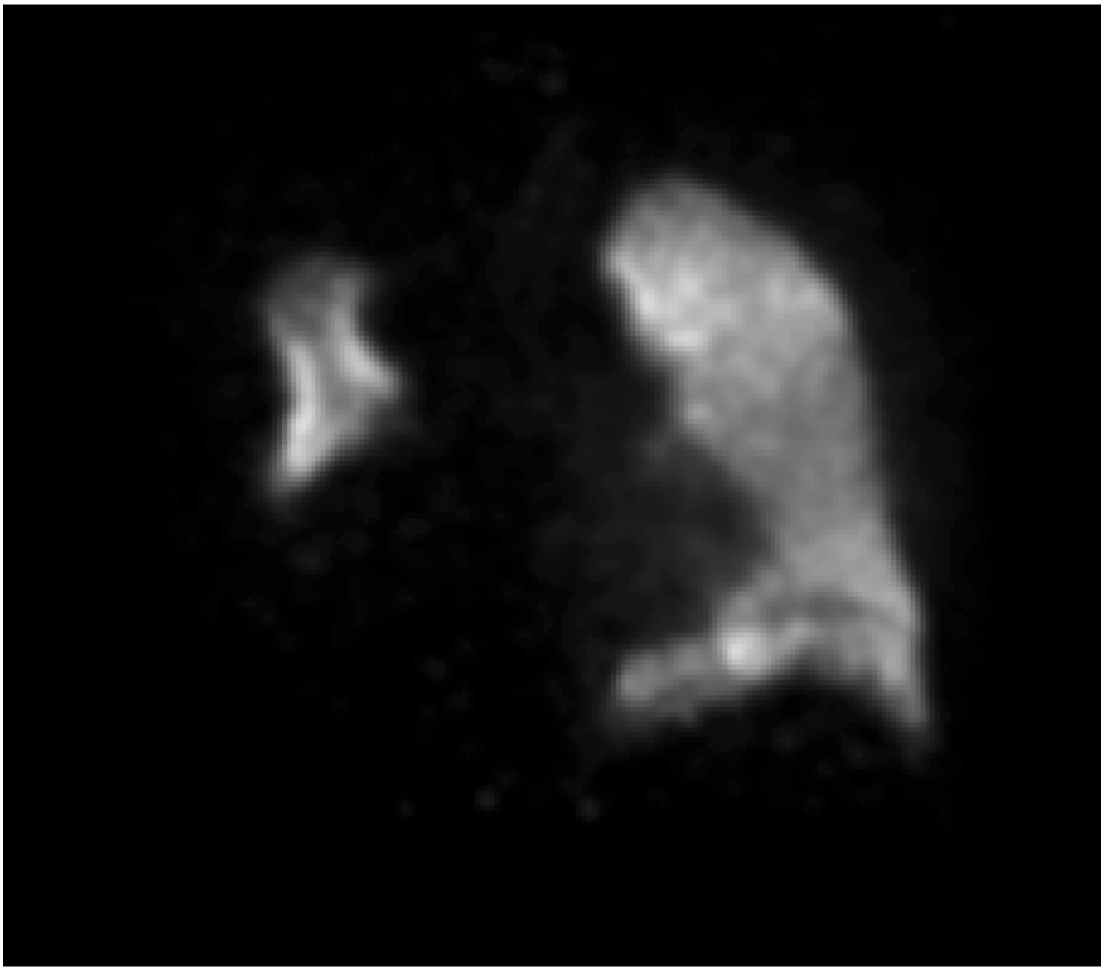


Figure 2.6 ^{68}Ga-aerosol (Galligas) free-breathing PET ventilation scan from a lung cancer patient.

Hyperpolarised gas MRI: scans use hyperpolarised noble gases inhaled immediately before image acquisition to visualise various aspects of lung function. Gases are intrinsically low spin density, leading to low polarisation; this creates a weak MR signal, limiting the application of gases in clinical imaging. However, the spin density of gases can be increased by around four to five times using spin-exchange optical pumping via a high-powered laser, leading to hyperpolarisation (Norquay *et al.*, 2018). This polarisation is non-permanent, decaying according to T_1 relaxation in approximately 10 seconds (Stewart *et al.*, 2021). In addition, a radiofrequency transmit-receive coil is required to image the nuclei of interest. Unlike nuclear medicine imaging modalities such as SPECT and PET, hyperpolarised gas MRI does not require ionising radiation. Non-ionising contrast agents, including Helium-3 (^{3}He) and Xenon-129 (^{129}Xe), can be used for hyperpolarised gas MRI with both nuclei exhibiting slightly different properties. ^{3}He was originally preferred due to its higher gyromagnetic ratio, leading to an intrinsically stronger MR signal; however, a worldwide paucity of ^{3}He has led to a large increase in cost, resulting in increased use of ^{129}Xe for hyperpolarised gas MRI. Furthermore, spatial resolution differs between ^{3}He and ^{129}Xe MRI where ^{3}He has a higher spatial resolution due to its increased gyromagnetic ratio and subsequently higher SNR (Stewart *et al.*, 2018). The lower SNR of ^{129}Xe MRI means a lower bandwidth is also required which produces longer scan times. ^{3}He and ^{129}Xe also differ in their diffusivity where ^{3}He has a higher diffusion coefficient making it less likely to penetrate partial obstructions, leading to increased heterogeneity in ^{129}Xe MRI scans.

Hyperpolarised gas MRI can be used to image lung ventilation, microstructure, or gas exchange. Ventilation is measured directly via the distribution of inhaled gas throughout the lung. Lung microstructure is assessed via hyperpolarised gas diffusion-weighted MRI; ^{3}He and ^{129}Xe are approximately five times more diffusive than water in tissue where they can diffuse in the acinar airspace through Brownian motion (Yablonskiy *et al.*, 2017). Thus, if there is damage to the acinar microstructure then the diffusion will be restricted. ^{129}Xe is soluble in the blood with distinct resonances in the tissue, blood plasma, red blood cells and the airspace (Ebner *et al.*, 2017). Using these resonances, dissolved-phase ^{129}Xe MRI can provide insights into the lung's gas exchange abilities.

Hyperpolarised gas MRI ventilation has been validated in several diseases. In COPD patients, hyperpolarised gas MRI shows significant ventilation abnormalities and is highly sensitive to airway obstruction when compared to spirometry measures (Kirby *et al.*, 2013; Kirby *et al.*, 2010). In addition, ventilation defects are observed in asthma patients who have undergone hyperpolarised gas MRI and correlate with spirometry measures (de Lange *et al.*, 2006; Zha *et al.*, 2018). In CF patients, agreement between defect location in structural MRI and ^{129}Xe MRI has been observed (Thomen *et al.*, 2020). Furthermore, hyperpolarised gas MR imaging has demonstrated greater sensitivity than conventional spirometry in the detection of mild CF disease (Aurora *et al.*, 2004). ^{3}He and ^{129}Xe MRI offer a non-invasive, non-ionising functional imaging modality that is sensitive to disease progression, facilitating longitudinal monitoring, particularly in paediatric patients. The University of Sheffield is currently the only centre worldwide which offers clinical hyperpolarised gas MRI, gaining MHRA approval for the acquisition of ^{129}Xe and ^{3}He MRI in patients referred for clinical imaging (Stewart *et al.*, 2015). As stated, hyperpolarised gas MRI can provide insight into lung ventilation, microstructure and gas exchange; this thesis will focus on the use of hyperpolarised gas MRI for the direct measurement of ventilation. Thus, example images shown in Figure 2.7 and Figure 2.8 depict ^{129}Xe and ^{3}He hyperpolarised gas MRI ventilation scans, respectively.

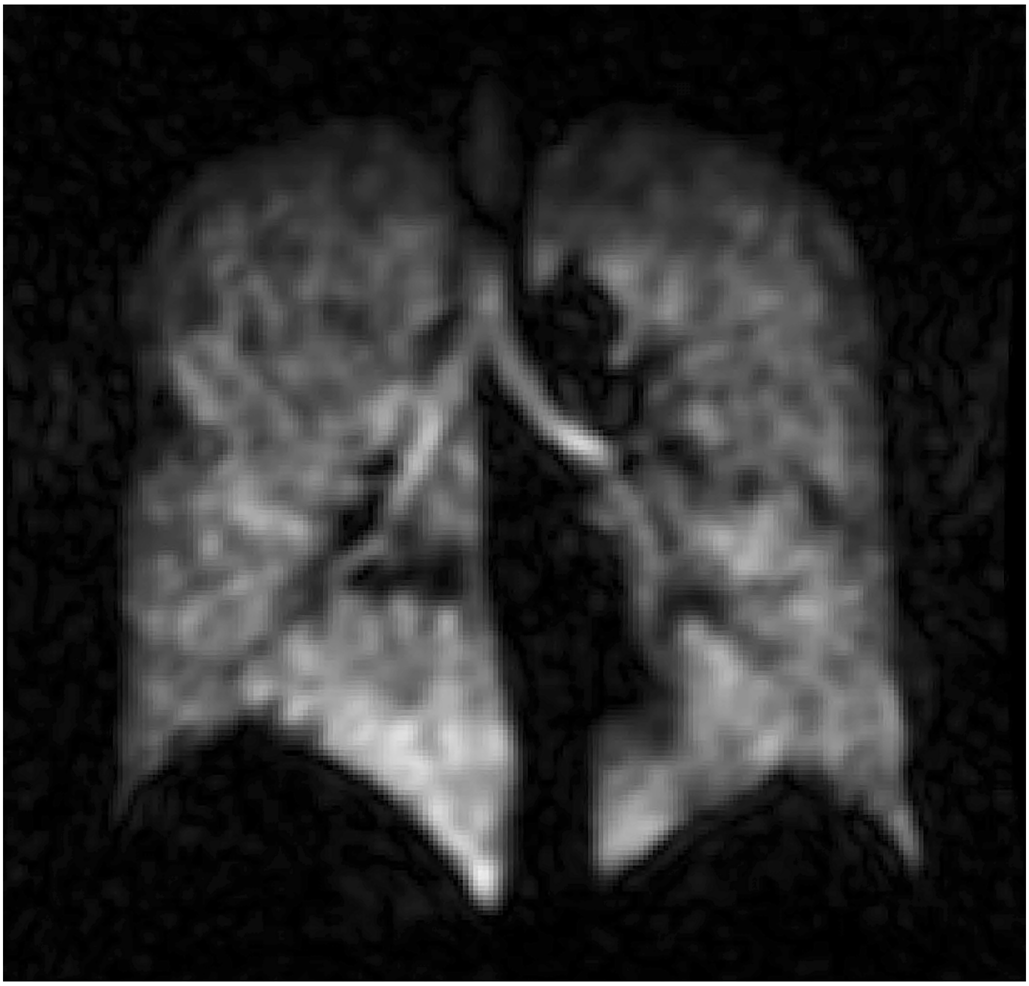


Figure 2.7 3D coronal balanced steady-state free precession (bSSFP) hyperpolarised ^{129}Xe lung MRI for a lung cancer patient acquired using a 1.5T GE HDx scanner with a flexible quadrature radiofrequency coil for transmission and reception of MR signals at the Larmor frequency of ^{129}Xe at a resolution of 4x4x10mm^{3}. A TR/TE of 6.7/2.2 milliseconds with a flip angle of 9° and bandwidth of ±8kHz was used. ^{129}Xe was polarised on site to approximately 25% by using an in-house developed rubidium spin-exchange polariser.

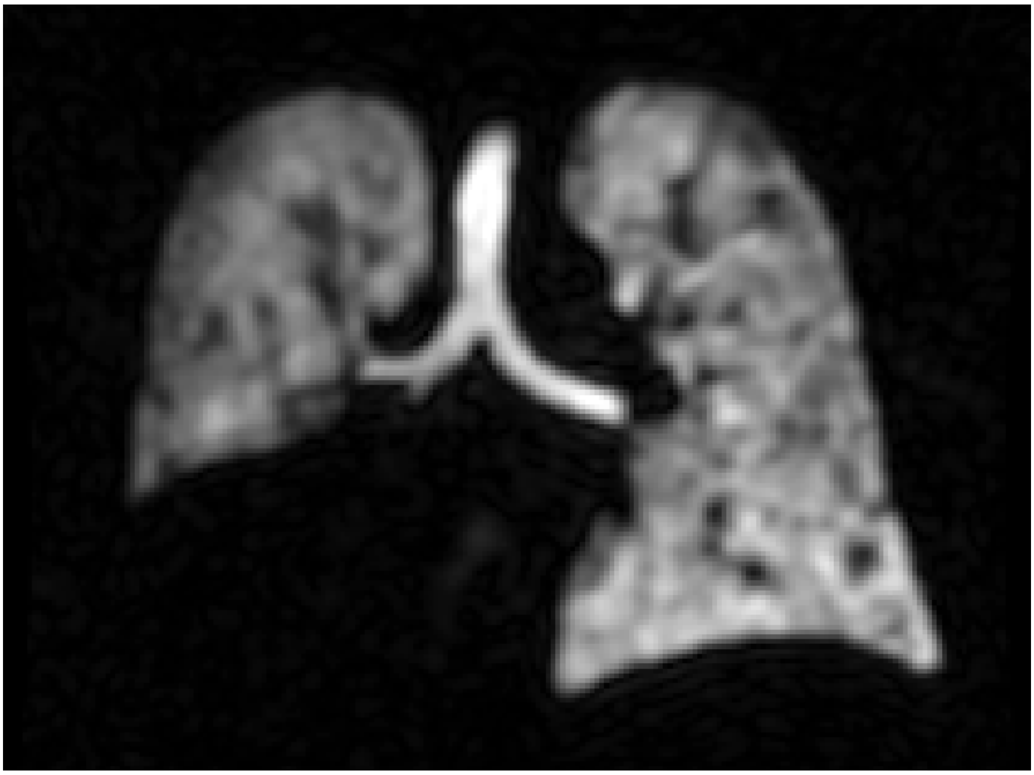


Figure 2.8 3D coronal balanced steady-state free precession (bSSFP) hyperpolarised ^{3}He lung MRI for a lung cancer patient acquired using a 1.5T GE HDx scanner with a flexible quadrature radiofrequency coil for transmission and reception of MR signals at the Larmor frequency of ^{3}He at a resolution of 4x4x5mm^3. A TR/TE of 1.9/0.6 milliseconds with a flip angle of 10° and bandwidth of ±166.6kHz was used. ^{3}He was polarised on site to approximately 25% by using an in-house developed rubidium spin-exchange polariser.

2.2.3 Image registration-based non-contrast functional lung imaging

Various techniques have been proposed that utilise image registration to model lung ventilation from multi-inflation structural imaging. Image registration is traditionally formulated as an optimisation problem whereby a cost function is employed to produce an optimal spatial transform from one (moving) image to the spatial domain of a second (fixed) image, facilitating qualitative and quantitative comparison between images; this can occur intra- or inter-modality. Image registration is broadly divided into four components, namely, transformation, interpolation, the similarity metric and optimisation. Transformations can be rigid, affine or deformable. Rigid transformations include rotations and translations, giving a total of six degrees of freedom in 3D. Affine transformations include rigid deformations with the addition of scaling and shearing, giving a total of 12 degrees of freedom in 3D. Deformable transformations employ curved non-linear deformations that are often employed for lung registration tasks due to the deformable movements exhibited by pulmonary structures. This includes diffeomorphic deformations which can model 3D deformations based on the computation of a differentiable vector field. Often a combination of these transforms is used in sequence to generate accurate registrations. Image interpolation is required to calculate voxel intensities in the new coordinate system after transformation. The interpolation algorithm used will differ depending on the application and image being registered; for example, if a binary segmentation is being registered, traditionally, a nearest neighbour algorithm would be employed. A voxel similarity measure such as the sum of squared differences or normalised cross-correlation is selected which is optimised by minimising a cost function using an algorithm such as gradient decent to generate accurate registrations between two images. In addition, a regularisation term is applied to ensure that only physically plausible transformations are included. Registration is used in the generation of non-contrast functional lung images, where multi-inflation structural images are registered to the same spatial domain, allowing the calculation of various ventilation metrics. Subsequently, the registered structural scans can be transformed to the same spatial domain as a corresponding contrast-based functional lung imaging modality such as hyperpolarised gas MRI, facilitating comparison between non-contrast and contrast-based functional lung images.

CT ventilation imaging (CTVI): has been proposed as a non-contrast functional lung imaging modality derived from multi-inflation structural CT imaging. The method assumes that the air fraction of a parenchymal voxel in a CT scan (F_{air}) is equal to:

$$\mathrm{F_{air}} = -\frac{\mathrm{HU}}{1000} \tag{2.1}$$

Specific ventilation (SV) is defined as the ratio of fresh gas volume delivered to the alveoli following inspiration divided by the volume at expiration as follows:

$$\mathrm{SV} = \frac{\Delta \mathrm{V}}{\mathrm{V_{air}^{exp}}} \tag{2.2}$$

where $\Delta\mathrm{V}$ refers to the change in volume between inspiratory and expiratory voxels and $\mathrm{V_{air}^{exp}}$ refers to the volume of air at expiratory inflation. A pulmonary parenchymal voxel comprises air and tissue; assuming that any change in volume is due to changes in ventilation i.e., that there is no change in tissue, then a surrogate of SV can be derived as follows:

$$\mathrm{SV} \approx \frac{\mathrm{F}_{\mathrm{air}}^{\mathrm{insp}} - \mathrm{F}_{\mathrm{air}}^{\mathrm{exp}}}{\mathrm{F}_{\mathrm{air}}^{\mathrm{exp}}\,(1 - \mathrm{F}_{\mathrm{air}}^{\mathrm{insp}})} \tag{2.3}$$

where $\mathrm{F}_{\mathrm{air}}^{\mathrm{insp}}$ and $\mathrm{F}_{\mathrm{air}}^{\mathrm{exp}}$ denote the fraction of air at inspiratory and expiratory volumes, respectively. Using this formulation, the HU CTVI metric ($\mathrm{CT}^{\mathrm{HU}}$) can be calculated. The $\mathrm{CT}^{\mathrm{HU}}$ metric is computed using voxel-wise intensity differences in HU values based on the formulation by Guerrero *et* al. (2005) shown below:

$$\mathrm{CT}^{\mathrm{HU}} = 1000\,\frac{\mathrm{HU}_{\mathrm{insp}} - \mathrm{HU}_{\mathrm{exp}}}{\mathrm{HU}_{\mathrm{exp}}\,(1000 + \mathrm{HU}_{\mathrm{insp}})} \tag{2.4}$$

where $\mathrm{HU}_{\mathrm{insp}}$ represents the HU of voxels in the warped inspiratory scan that spatially correspond to voxels in the expiratory scan and $\mathrm{HU}_{\mathrm{exp}}$ represents the HU of expiratory voxels. $\mathrm{CT}^{\mathrm{HU}}$ aims to measure the change in the fractional content of air, in a voxel-wise manner, between expiratory and inspiratory phases (Simon *et al.*, 2012). Other CTVI metrics have been proposed such as the determinant of the Jacobian matrix ($\mathrm{CT}^{\mathrm{JAC}}$) which aims to model local volume change via deformation vector fields. These metrics have demonstrated moderate correlation with hyperpolarised gas MRI across various respiratory diseases (Tahir *et al.*, 2018). Recently, more advanced, proprietary CTVI metrics such as the mass conserving volume change have been developed to account for potential limitations in previously developed CTVI metrics (Castillo *et al.*, 2019). Figure 2.9 depicts an example $\mathrm{CT}^{\mathrm{HU}}$ ventilation scan for a patient with lung cancer.

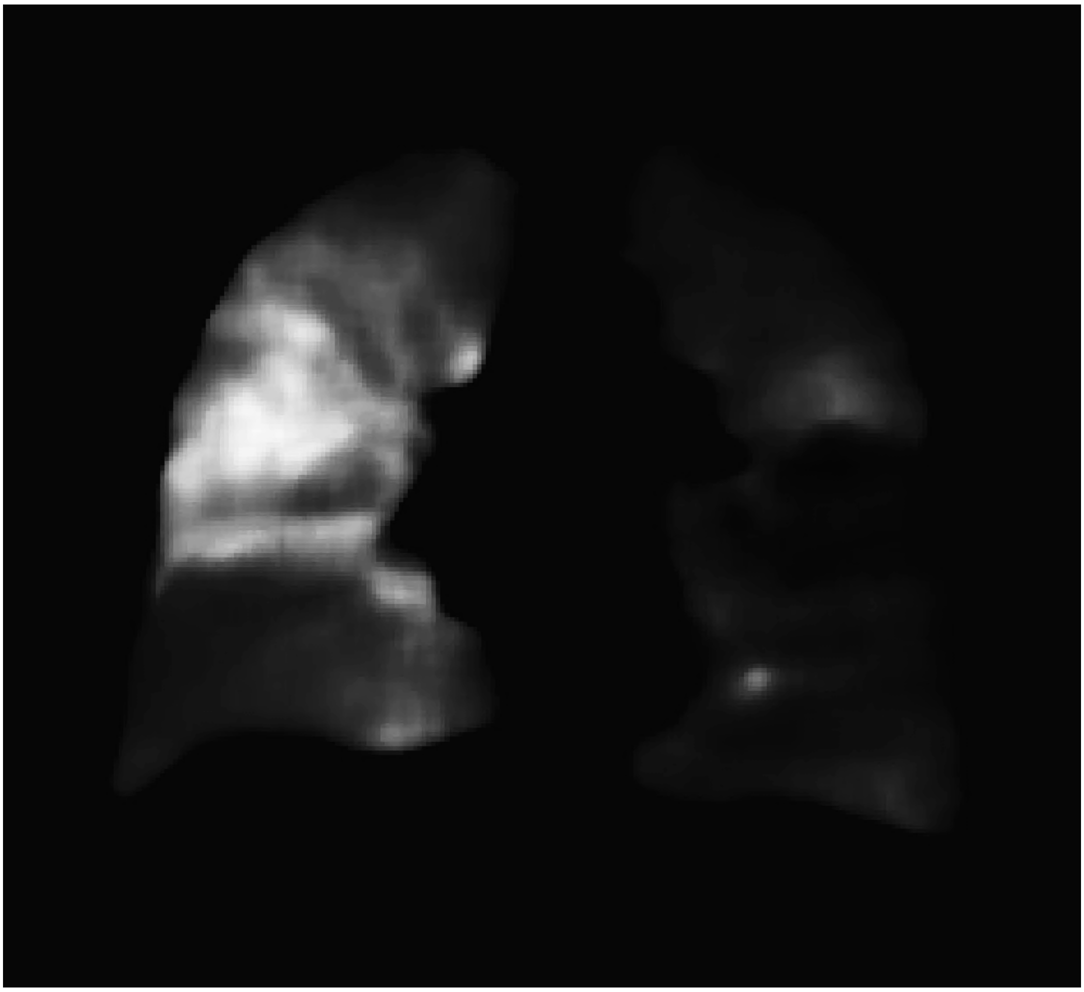


Figure 2.9 Coronal slice of a CT^{HU} ventilation scan at expiratory geometry generated from inspiratory and expiratory breath-hold CT scans for a patient with lung cancer.

^{1}H-MRI ventilation: scans are generated in a similar manor as CTVI scans where lung ventilation can be calculated from multi-inflation structural ^{1}H-MRI (Zapke *et al.*, 2006). ^{1}H-MRI ventilation models assume that differences in signal intensities of co-registered voxels in multi-inflation ^{1}H-MRI reflect naturally occurring density variations in the lungs during breathing (Kjørstad *et al.*, 2017). As with the $\mathrm{CT^{HU}}$ ventilation metric, ^{1}H-MRI ventilation models aim to compute the SV with similar assumptions via deformably registered expiratory and inspiratory ^{1}H-MRI scans (Capaldi *et al.*, 2018) as follows:

$$\mathrm{SV} = \frac{\Delta \mathrm{V}}{\mathrm{V_{air}^{exp}}} \approx \frac{\mathrm{F_{air}^{insp}} - \mathrm{F_{air}^{exp}}}{\mathrm{F_{air}^{exp}}} \tag{2.5}$$

where $\mathrm{F_{air}^{insp}}$ and $\mathrm{F_{air}^{exp}}$ denote the fraction of air at inspiration and expiration, respectively. Due to the arbitrary units of structural ^{1}H-MRI voxels, the MRI signal intensity (SI) cannot be directly calculated in the same way as CT and, therefore, it is assumed that SI is approximately inversely proportional to F$_{\text{air}}$ (Zapke *et al.*, 2006).

$$\mathrm{SI}\ \tilde{\propto}\ \frac{1}{\mathrm{F_{air}}} \tag{2.6}$$

Substituting Equation (2.6 into Equation (2.5 allows the SV to be approximated as follows:

$$\mathrm{SV}\ \approx \left(\frac{\mathrm{SI_{exp}} - \mathrm{SI_{insp}}}{\mathrm{SI_{insp}}}\right) \tag{2.7}$$

where SI_{exp} and SI_{insp} are voxel-wise signal intensities at expiratory and inspiratory inflations, respectively. ^{1}H-MRI ventilation models produce non-contrast functional lung images, without ionising radiation, that limit the need for specialised equipment required for contrast-based functional lung imaging, such as hyperpolarised gas MRI. Figure 2.10 depicts a ^{1}H-MRI ventilation surrogate for a patient with COPD.

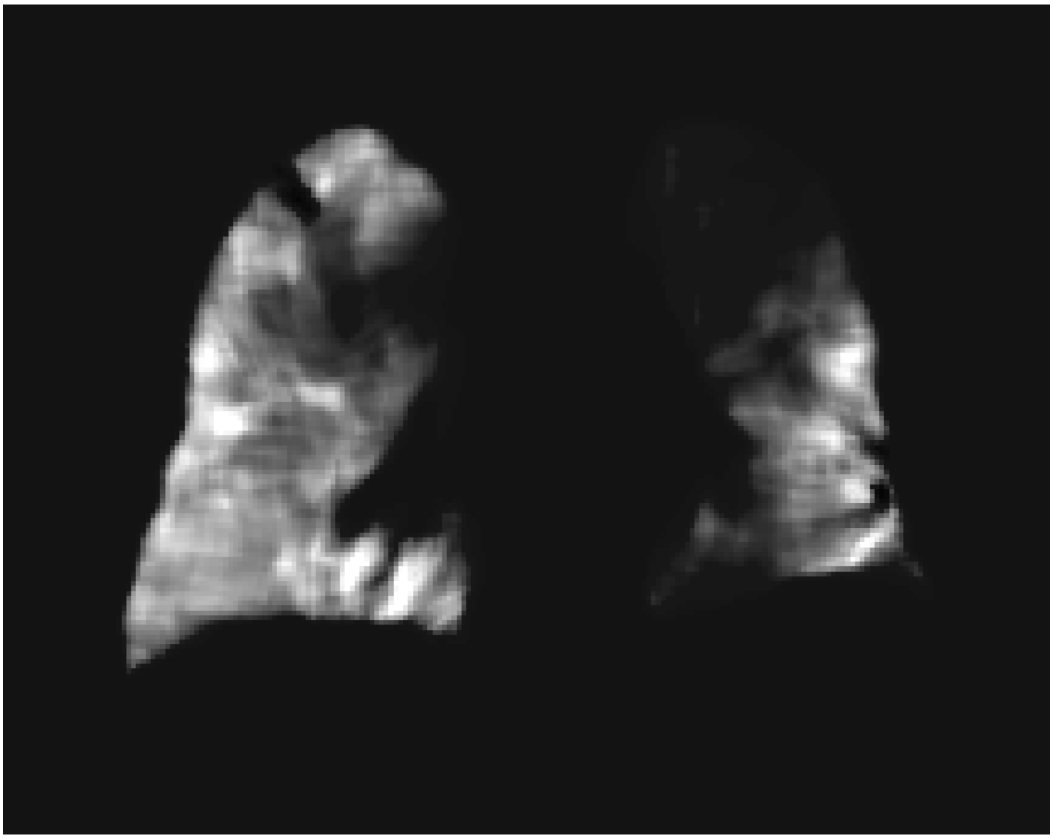


Figure 2.10 Coronal slice of ^{1}H-MRI ventilation scan at expiratory geometry generated from inspiratory and expiratory breath-hold isotropic SPGR ^{1}H-MRI scans.

2.2.4 Image segmentation-based regional functional lung imaging ventilation biomarkers

Several biomarkers have been proposed to quantify ventilation and ventilation heterogeneity in functional lung images. These biomarkers are computed over a region of the image; therefore, image segmentation is required to delineate these regions of interest. Image segmentation is the process of defining a region, such as the lung parenchyma or the ventilated lung, either by manually delineating it or through a semi-automated algorithm which is subsequently edited by a trained individual. Various algorithms exist to semi-automatically segment functional lung images, such as hyperpolarised gas MRI.

A k-means clustering algorithm was previously modified for hyperpolarised gas MRI ventilation segmentation (Kirby *et al.*, 2012a). This method attempts to find k data points, given the integer k, in an n-dimensional space R^n given m data points. These k data points are known as centres/centroids and the aim is to minimise the distance from each data point (m) to its centre/centroid (Kanungo *et al.*, 2002). The method attempts to delineate the image data into a number of clusters that can best represent a radiologist's analysis of the ventilation image with clusters defined from defects to hyperintense signal (Kirby *et al.*, 2012a). The first stage of this method requires image normalisation into the range of 0-255, following which the cluster initial centres are set at 25% intervals between these values. Commonly, two-stage clustering is used, whereby initially four clusters are selected (the lowest of which contains both signal void and hypointense signal) followed by a second clustering applied to the lowest cluster from the first stage to define background, ventilation defect and hypointense signal regions.

The spatial fuzzy c-means clustering (SFCM) algorithm has been used to segment ventilation and structural MR image pairs (Biancardi *et al.*, 2018). Images are initially bilaterally filtered to remove noise and maintain edges (Tomasi and Manduchi, 1998b). The standard FCM algorithm assigns N pixels to C clusters via fuzzy memberships with the assumption that pixels in close proximity are highly correlated and hence have similarly high membership to the same cluster (Bezdek *et al.*, 1984). This spatial information will modify the membership value only if, for example, the voxel is noisy and would have been incorrectly classified. The SFCM method makes use of nearby voxels during the iteration process by considering the membership of voxels within a predefined window and will weigh the central voxel depending on the provided weighting variables (Chuang *et al.*, 2006). The number of clusters and thresholds for inclusion can be altered manually to generate the most accurate segmentations. SFCM has an advantage over k-means clustering as it is applied to both hyperpolarised gas MRI scans and structural ^{1}H-MRI scans in a pair-wise fashion to take advantage of the combined information arising from the co-location of the image pair to segment both the ventilated and structural regions simultaneously (Biancardi *et al.*, 2018).

Segmentations of the lung cavity or the ventilated portion of the lungs can be used to calculate several important regional lung ventilation biomarkers from hyperpolarised gas MRI and ^{1}H-MRI scans. The ventilation defect percentage (VDP), defined as the percentage of low/zero intensity voxels in the scan, or the $\mathrm{VDP's}$ inverse, the ventilated volume percentage ($\%\mathrm{vV}$), are common metrics used as biomarkers of regional ventilation. VDP and $\%\mathrm{vV}$ are calculated by comparing structural and ventilated lung segmentations to generate a percentage value of ventilated lung volume as follows:

$$\mathrm{VDP} = \left(1 - \frac{\text{ventilated lung volume}}{\text{total lung volume}}\right) \times\ 100 \tag{2.8}$$

$$\%\mathrm{vV} = \left(\frac{\text{ventilated lung volume}}{\text{total lung volume}}\right) \times\ 100 \tag{2.9}$$

Similarly, the ventilated volume can be calculated from the ventilated lung segmentation; however, this metric is not normalised by the total lung volume which can lead to biases depending on the size of the lungs. An additional biomarker which can be calculated from these segmentations is the number of defects present in a hyperpolarised gas MRI scan, which may be important in specific diseases, such as CF (Smith *et al.*, 2018; Stewart *et al.*, 2018). A further regional ventilation biomarker which determines ventilation heterogeneity, known as the coefficient of variation (CoV), can be calculated from image intensity values as follows (Stewart *et al.*, 2021):

$$\mathrm{CoV} = \frac{\text{standard deviation}}{\text{mean}} \tag{2.10}$$

Recently, binning-based biomarkers have been proposed which classify pixels into defect, low, normal and high ventilation bins depending on various thresholds. This binning approach has been used to quantify ventilation in hyperpolarised gas MRI scans (He *et al.*, 2016).

2.3 Deep learning theory

Deep learning (DL) is a subfield of machine learning that commonly employs artificial neural networks (ANNs) with multiple deep or hidden layers in an end-to-end learning approach where features are learnt implicitly by the network. In contrast to traditional machine learning algorithms, such as random forests, K-means clustering and support vector machines, hand-crafted feature extraction is not required. Traditional machine learning utilises explicitly defined features, which are manually engineered via domain-specific knowledge, and can include features such as shape priors and intensity histograms. Due to the hierarchical nature of ANNs, feature selection at varying levels of abstraction is integrated into the function mapping process between an input and an output within the ANN. The features 'selected' by the network are those that produce a better mapping between the input and output domains. The varying levels of abstraction achieved by hidden layers in ANNs often allow for more complex features than those defined through hand-crafted feature selection and, consequently, frequently lead to improved modelling when an end-to-end approach is employed. Differences between traditional machine learning and DL approaches are displayed for an image segmentation task in Figure 2.11. In addition, traditional machine learning algorithms often require access to all training examples when classifying a new datapoint; for example, in clustering algorithms, the algorithm must know the position of all other data points within the dataset to assign a new data point to a specific cluster, whereas in DL algorithms, this information is stored within the learned weights and biases of a trained ANN.

As with traditional machine learning, DL can be either supervised or unsupervised; supervised learning uses labelled training data to map complex functions between input and output domains. Conversely, unsupervised learning does not use labelled data; instead, inherent patterns are discovered directly from unlabelled data. The investigations detailed within this thesis will primarily use supervised DL approaches and, therefore, the following sections will focus on these algorithms; this includes an introduction to ANNs, convolutional neural networks (CNNs), DL training strategies and techniques to validate DL models.

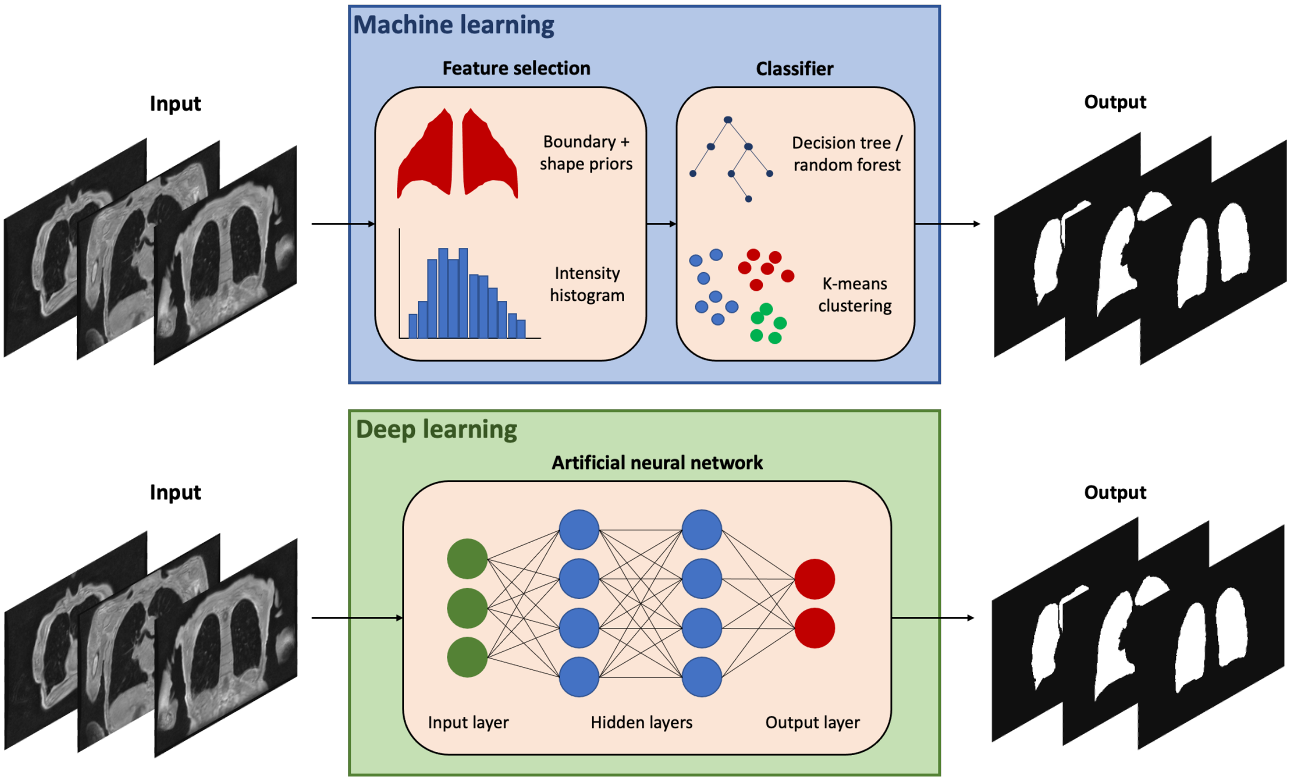


Figure 2.11 Comparison of traditional machine learning and DL algorithms for a lung segmentation task, contrasting the hand-crafted feature selection and the end-to-end approach of machine learning and DL, respectively.

2.3.1 Artificial neural networks

The concept of modelling biological neurons that mimic the brain's computational capacity was first explored in the 1950s with the implementation of the perceptron (Rosenblatt, 1958). ANNs, or multi-layer perceptrons, are algorithms that map a complex function between an output and an input domain. The development of fully-connected ANNs with multiple hidden layers allows for an increased number of parameters, giving the network more freedom to tune these parameters for the task of learning complex functions (Werbos and John, 1974). The fundamental unit of an ANN is a neuron which stores a value; in fully-connected networks, each neuron in a layer is connected to every other neuron in the proceeding and subsequent layers. The strength of connections between neurons is determined by a weight (w) and a bias (b) term. A neuron computes the weighted sum of its activation plus a bias, representing a minimum threshold activation for each neuron, and applies a nonlinear activation function (σ) to produce an output $\alpha^{(\mathrm{L})}$. Neurons are organised into layers in a hierarchical fashion whereby the output of layer $\mathrm{L}-1$ corresponds to the input activation of a neuron in layer L. Weights in a network can be represented by a matrix and the activations and biases at each layer as vectors where $\mathbf{W}$ represents the weight matrix and $\boldsymbol{\alpha}^{(\mathrm{L}-1)}$ and $\mathbf{b}$ represent the corresponding vectors of activations and biases, respectively.

$$\boldsymbol{\alpha}^{(\mathrm{L})} = \sigma\left(\mathbf{W}\boldsymbol{\alpha}^{(\mathrm{L-1})} + \mathbf{b}\right)$$

(2.11)

Activations are propagated through a network to produce an output in what is known as the forward pass. For an ANN with an input layer containing pixel activations of x, a hidden layer ($\alpha^{(\mathrm{L-1})}$) and an output layer ($\alpha^{(\mathrm{L})}$), the difference between the network's output and the expected outcome is the cost (C), which can be calculated via a loss function ($\mathcal{L}$) for a single neuron using a forward pass as follows:

$$\begin{gathered}
\mathrm{z}^{(\mathrm{L-1})} = \mathrm{w}^{(\mathrm{L-1})}\mathrm{x} + \mathrm{b}^{(\mathrm{L-1})} \\
\alpha^{(\mathrm{L-1})} = \sigma\left(\mathrm{z}^{(\mathrm{L-1})}\right) \\
\mathrm{z}^{(\mathrm{L})} = \mathrm{w}^{(\mathrm{L})}\alpha^{(\mathrm{L-1})} + \mathrm{b}^{(\mathrm{L})} \\
\alpha^{(\mathrm{L})} = \sigma\left(\mathrm{z}^{(\mathrm{L})}\right) = \hat{\mathrm{y}} \\
\mathrm{C} = \mathcal{L}(\hat{\mathrm{y}}, \mathrm{y})
\end{gathered}$$

(2.12)

where y refers to the expected output and $\hat{\mathrm{y}}$ refers to the output of the final layer of the network. A non-linear activation function is required as this allows the network to learn complex non-linear functions. The activation function used is a hyper-parameter which can be tuned to improve performance. Activation functions can vary at different layers of the network; for example, in classification problems, the final output layer commonly utilises a SoftMax activation function that scales output weights into probability distributions. Common activation functions include the sigmoid, the rectified linear unit (ReLU) and the Leaky ReLU functions (He *et al.*, 2015; Agarap, 2018). The Leaky ReLU activation function is a special case of the more general partial ReLU (pReLU); in both cases, an α value is given for negative values, generating a differentiable slope for negative values. Maas (2013) initially proposed an α of 0.01 for use in neural networks defined as the Leaky ReLU; however, any constant value can be employed, producing the more general pReLU. Figure 2.12 depicts common activation functions used in ANNs with example arbitrary values.

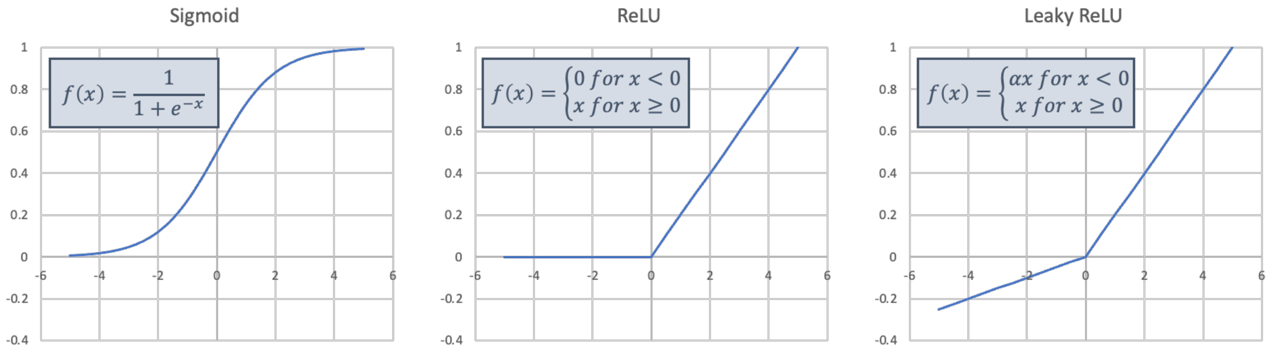


Figure 2.12 Common activation functions, including the Sigmoid, ReLU and Leaky ReLU functions. The example Leaky ReLU employs an α of 0.025.

At the heart of ANNs is optimisation; an algorithm learns by optimising weights and biases for a generalisable solution to a complex function. To achieve this, various weights and biases need to be updated over multiple iterations, leading to convergence between the output and the expected outcome. This process of continually updating weights and biases to optimise a loss function is known as training. The algorithm for updating weights and biases based on training examples in a dataset is known as backpropagation (Werbos and John, 1974; Plaut *et al.*, 1986). Backpropagation utilises the chain rule and is applied iteratively at each layer of the network. The development of backpropagation algorithms was the catalyst that allowed for the implementation of ANNs with several hidden layers, dramatically reducing the number of calculations required and, therefore, reducing the computational requirements associated with training ANNs. For a network with one neuron in each layer, the partial derivative of the cost with respect to the weights and biases of a layer, that is the impact on the cost from small changes in the weights and biases, is denoted as C_0 and calculated as follows:

$$\frac{\partial \mathrm{C}_0}{\partial \mathrm{w}^\mathrm{L}} = \frac{\partial \mathrm{z}^\mathrm{L}}{\partial \mathrm{w}^\mathrm{L}} \frac{\partial \mathrm{a}^\mathrm{L}}{\partial \mathrm{z}^\mathrm{L}} \frac{\partial \mathrm{C}_0}{\partial \mathrm{a}^\mathrm{L}}$$

$$\frac{\partial \mathrm{C}_0}{\partial \mathrm{b}^\mathrm{L}} = \frac{\partial \mathrm{z}^\mathrm{L}}{\partial \mathrm{b}^\mathrm{L}} \frac{\partial \mathrm{a}^\mathrm{L}}{\partial \mathrm{z}^\mathrm{L}} \frac{\partial \mathrm{C}_0}{\partial \mathrm{b}^\mathrm{L}}$$

(2.13)

C_0 represents the cost for a single training case; the cost C for all training examples k is the sum of the costs for each case with respect to the weights and biases of the output layer L and is computed as:

$$\frac{\partial C}{\partial w^L} = \frac{1}{n}\sum_{k=0}^{n-1}\frac{\partial C_k}{\partial w^L} = \theta_w$$

$$\frac{\partial C}{\partial b^L} = \frac{1}{n}\sum_{k=0}^{n-1}\frac{\partial C_k}{\partial b^L} = \theta_b$$

(2.14)

ANNs often contain multiple neurons in each layer, so the partial differentiation of weights and biases with respect to the cost is commonly given as a vector denoting the gradient of the cost at each layer with respect to all the neurons in the layer. The loss function penalises errors compared to some expected answer and is used to calculate the cost; it is an important hyper-parameter in network training as it represents the mechanism by which the network determines differences between its output and the expected outcome. Therefore, it is minimised to produce an effective ANN. Loss functions vary depending on the application; for example, a mean square error (MSE) loss function may be used in regression tasks, whereas in segmentation tasks, a binary cross-entropy loss function is more suitable. The development of novel loss functions for specific applications is a key area of DL research and has been used in medical imaging applications to improve segmentation performance with the development of the generalised Dice loss (Sudre et al. 2017). It is important that the loss function has a smooth distribution that contains no step changes; this allows for its minimisation through gradient descent methods.

The partial differential of the cost with respect to the weights and biases represents the gradient, or slope, of the cost function at the current position, where θ_w and θ_b represent the weight gradient and bias gradient, respectively. By taking the negative gradient, the downhill direction can be determined, representing a minimisation of the loss function. This is known as gradient descent and provides the network with a mechanism of adjusting its parameters to reduce the difference between its output and the expected value. The process of updating the weights and biases for a single neuron network is as follows:

$$w_{NEW} = w_{OLD} - \eta(\theta_w)$$

$$b_{NEW} = b_{OLD} - \eta(\theta_b)$$

(2.15)

where w_{NEW} and b_{NEW} are the updated weights and biases, and w_{OLD} and b_{OLD} are the previous weights and biases, respectively, and where η is the learning rate. These changes are propagated iteratively at each layer of the network, representing the backward pass, otherwise referred to as network training where weights and biases are continually updated to produce a network capable of generating expected outputs on previously unseen data. In practice, gradient descent is unfeasible due to the significant computational requirement of simultaneously calculating gradients for all training examples. Instead, variations on gradient descent, such as stochastic gradient descent (SGD) or mini-batch gradient descent, are used to reduce the computational requirement (Robbins and Monro, 1951; Feyzmahdavian *et al.*, 2015). SGD calculates the gradient for each example in the training dataset, leading to inefficient convergence if there are large variations between training examples. In contrast, mini-batch gradient descent often convergences faster than both conventional gradient descent and SGD. It employs a subset of the training examples, known as a batch, and calculates the gradients for each randomly shuffled batch, leading to a minimisation of the loss function; mini-batch gradient descent is common in DL applications due to the large number of training cases. For all variants of gradient descent, a learning rate η is required. The learning rate determines how large the convergence step should be at each iteration. For traditional gradient descent, the learning rate is often held constant as convergence tracks a straight path towards either a global minimum if the loss function is convex, or a local minimum if it is not; however, for SGD and mini-batch gradient descent, the learning rate can be variable, reducing as convergence approaches a local minimum. This is due to both algorithms fluctuating about a local minimum instead of directly converging on the desired minimum. Recently, a type of gradient descent known as Adam has been employed to achieve fast convergence on non-convex loss functions (Kingma and Ba, 2015). Adam can be described as leveraging the advantages of common SGD variants, namely, adaptive gradient algorithm (AdaGrad) and Root mean square propagation (RMSProp) (Duchi *et al.*, 2011; Dauphin *et al.*, 2015). It utilises per-parameter learning rates, whereby each parameter of the network has an independent learning rate based on the first- and second-order moments of the gradient; it has shown efficacy in several DL applications (Gerard *et al.*, 2018; Garcia-Uceda Juarez *et al.*, 2019; Li *et al.*, 2017).

2.3.2 Convolutional neural networks

Fully connected ANNs are computationally expensive and inefficient when processing large images; 3D images in the medical field tend to have millions of voxels, exacerbating this problem. In addition, ANNs are unable to account for spatially correlated information and features present in images and videos. To address these challenges, specifically adapted

ANNs, known as CNNs, have been widely adopted in applications that utilise large images in the dataset. CNNs are partially inspired by biological mechanisms in the eyes, whereby specific neurons respond to changes in small areas of the visual field, known as the receptive field (Hubel and Wiesel 1968). CNNs famously gained prominence in 2012 when AlexNet, a type of CNN, produced ground-breaking classification performance in the ImageNet Large Scale Visual Recognition Challenge, outperforming the second-place algorithm by over 10%. Using multiple graphical processing units (GPU), AlexNet distinguished between over 1000 classes, creating a new benchmark for image classification tasks in the process. The highly influential paper has been cited over 85,000 times and was successful in pushing CNNs and GPUs to the forefront of image-based DL tasks (Krizhevsky et al. 2012).

As previously stated, unlike traditional ANNs, CNNs are not fully-connected and, instead, use a convolving filter which performs a convolution operation on an input image. For an image i of size $X \times Y \times Z \times C$, where X, Y and Z are spatial dimensions of the image and C is the number of channels, a convolution operation is applied to the image with a filter of size $f \times f \times f$ and a stride length of s. In a CNN, network layers are arranged in a hierarchical fashion, whereby the output of one-layer acts as the input to the subsequent layer. In a convolutional layer with input i, the four-dimensional (4D) tensor is convolved with a filter to produce a 3D tensor which is then concatenated in the channel dimension with a defined number of filters so that for the subsequent layer L, a 4D tensor of $i_L = X_L \times Y_L \times Z_L \times C_L$ is used as an input. The number of filters is equivalent to the number of channels in i_L. Filters can be thought of as a matrix of values with a given size, where the number of filters is a hyper-parameter which can differ depending on the layer of the network. Whilst the number of channels is dependent on the number of learned filters, the spatial components of i_L can be manipulated based on the size of the filters and the stride. Stride refers to the number of pixels that a filter traverses during the convolution operation. Depending on the type of padding (p) used, for example zero-padding, the spatial dimensions of the input image can remain constant if a stride length of 1 is used. In general terms, the size of the output feature map O_{Map} can be calculated using the following equation:

$$O_{Map} = \frac{i + 2p - f}{s} + 1 \tag{2.16}$$

It is common to apply a pooling layer directly after a convolutional layer to reduce the spatial dimensions of the input image i. If a pooling layer with a filter of size of $2 \times 2 \times 2$ and a stride length of 2 was used, the spatial dimensions of the image would be reduced by a factor of two so that $i_L = \frac{X_L}{2} \times \frac{Y_L}{2} \times \frac{Z_L}{2} \times C_L$. This process of downsampling images with a filter of $2 \times 2 \times 2$ and a stride length of 2 is referred to as max pooling.

Filters are traditionally used in machine learning for problems where the filters are hand-crafted to detect specific, relevant features. In contrast, CNNs allow the values of filters to be trainable parameters, analogous to the weights in an ANN. As with traditional ANNs, a non-linear activation is required to learn complex features. CNNs often use similar activation functions to ANNs, including the sigmoid, ReLU, pReLU and hyperbolic tangent (TanH) functions. The filters are then 'learnt' through the familiar method of forward propagation and backpropagation with a minimisation of the loss function occurring through the continual updating of weights contained within each filter. For a classification task, the filters in the last convolutional layer will be connected to a fully-connected layer, traditionally with a SoftMax activation, that acts as the output layer, where each filter acts as a neuron for each class. Filters are analogous to feature detectors that correspond to the detection of curved edges for example. However, the key strength of CNNs for imaging applications is that, due to the hierarchical nature of CNNs, filters in deeper layers become increasingly abstract. If dimensionality is reduced during pooling so that $O_{Map} < i$, then the convolution operation in the next layer will be convolved over a compressed version of the image, producing increasingly abstract feature maps. For example, filters in the first hidden layer may detect vertical edges, whereas, in the final hidden layer, each filter may be capable of detecting the trachea. This allows for significantly more complex function mapping between input and output domains than hand-crafted features, where domain-specific knowledge is required to determine which features are relevant. Activation, or feature, maps can be generated at each convolutional layer using the learned filters, providing a representation of which features the CNN is using to produce its output (Lee *et al.*, 2011). The combination of convolutional, pooling and activation layers can be referred to as a 'block'. By linearly combining several blocks, the spatial dimensionality can be further reduced, leading to highly abstract and specialised feature detectors. Batch normalisation layers are often employed after convolutional and pooling layers in each network block. Batch normalisation resets the distribution of activations in the previous layer, reducing the covariate shift between network layers, leading to more efficient network training with a reduction in data loss between network layers; normalisation is achieved by subtracting the mean activation from each batch and then dividing by its standard deviation. Thus, it is ensured that there are no activations which are too large or too small which can cause non-differentiable gradients, where the partial differential of the gradient used in the backpropagation algorithm tends to zero. Consequently, the value of weights in the network remains unchanged; therefore, the network can no longer effectively 'learn' through continually updating network parameters. Batch normalisation also acts as a form of regularisation, minimising overfitting in training where some examples produce much larger activations than other examples, thereby reducing the networks reliance on these examples which produce large activation values.

When several convolutional blocks containing pooling layers are combined, the features it can detect are highly abstract; however, the dimensions of the output layer are significantly different from those of the input image. For applications such as image segmentation or image synthesis, the predicted segmentation is required to be the same size as the input image. In some CNNs, upsampling is conducted using interpolation functions, such as linear interpolation, to enforce the size of the CNN output. However, these interpolation functions contain no learnable parameters and are therefore applied independently to each image without change. Many CNNs have utilised the transposed convolution operation to allow fully learnable upsampling. Increasing the spatial dimensions of the output feature map O_{Map}, where i is the input image dimensions to this layer, can be achieved similarly to how the convolution operation can be used to reduce image dimensions; here, the transposed convolution operation is used to increase the size of O_{Map} as follows:

$$O_{Map} = (i - 1)s + f - 2p \tag{2.17}$$

Several other hyper-parameters can be employed, such as dilation and output-padding, which modify the size of O_{Map} to produce the desired spatial dimension. For segmentation tasks, through a series of convolutional and transposed convolutional layers, it can be ensured that $O_{Map} = i$, where i is the original input image dimensions. These transposed convolutional layers are often conducted with the same filter size and stride as the original downsampling convolutions. This creates two 'sides' of the network, where one is the mirror image of the other, one being responsible for downsampling through convolutions and the other for upsampling through transposed convolutions. Various CNN architectures have been developed for image segmentation which utilise this network configuration. A CNN architecture refers to the specific configuration of convolutional, pooling, interpolation and transposed convolutional layers; these architectures are usually proposed in original research articles for use in specific applications, either to reduce computational expenses or to generate improved predictions when compared to previously proposed architectures. Advancements in CNN architectures have produced complex network configurations with several convolutional and transposed convolutional blocks for the downsampling and upsampling of images, respectively. Variations in architectures usually occur in the output layer; in classification tasks, the output layer is required to predict between n number of classes, giving n number of neurons in the output layer. For example, the VGGNet (Simonyan and Zisserman, 2014) is a CNN used for image classification tasks that has been utilised in medical imaging applications (Alebiosu and Muhammad, 2019). In this instance, upsampling is not required as the output need not be the same size as the input image.

2.3.3 Fully convolutional neural networks

CNN configurations which utilise convolution and transposed convolutions to produce output maps with the same spatial resolution as the input image are referred to as fully convolutional neural networks (fCNNs). For image segmentation or image synthesis tasks, upsampling is required; fCNNs have become the dominant network configuration for these tasks. Common fCNN architectures include the UNet and VNet architectures which employ residual connections and transposed convolutions for upsampling. Variations of the UNet have been developed which can receive 2D or 3D input images; the original UNet was implemented as a 2D fCNN, meaning that each slice of an image is processed independently (Ronneberger *et al.*, 2015). The UNet architecture contains 19 total layers consisting of convolutional, transposed convolutional and max pooling layers. Convolutional layers employ 3×3 convolutions followed by 2×2 max pooling to reduce image resolution with transposed convolutions of 2×2 in the upsampling path. The network can receive patches of various sizes; however, more pooling is required for larger patch sizes as well as reduced localisation accuracy when patches are too large. The UNet, is commonly used in many segmentation applications; more specifically, several investigators have utilised the UNet for pulmonary image segmentation tasks (Zhu et al. 2019; Eppenhof & Pluim 2019; Ren et al. 2019). In contrast to the UNet, the VNet architecture employs no pooling layers; instead, it contains only convolutional layers with a kernel size of 2×2 and a stride of 2 is used for downsampling. The use of convolutional layers instead of pooling layers generates learnable parameters with a reduced memory footprint as inputs do not need to be stored to conduct backpropagation. The VNet uses $5 \times 5 \times 5$ convolutions in both the upsampling and downsampling paths. The VNet architecture was designed to allow non-isotropic patch sizes and was specifically developed for fully-volumetric image segmentation in the medical imaging domain (Milletari *et al.*, 2016). Both the UNet and the VNet employ residual connections to forward information from the downsampling path to the upsampling path. Residual connections, sometimes called skip connections forward the output of convolutional blocks during downsampling directly to the equivalent layer in the upsampling path; this provides high-resolution features to transposed convolutional operations, improving network training and limiting the possibility of exploding/vanishing gradients. By providing alternative paths for information to be conveyed to the upsampling path, residual connections act as a kind of ensemble network where features can be forwarded and combined at several different resolutions. This often results in a more accurate output than networks which do not employ residual connections. In addition, the VNet uses fine-grained feature forwarding, allowing features to be shared between downsampling and upsampling sides of the network (Milletari *et al.*, 2016). Extensions of the UNet have been proposed, such as the nn-UNet (Isensee *et al.*, 2018). The nn-UNet is a 3D implementation of the UNet which is adapted for use in the medical imaging domain and can be deployed on a large range of input image sizes and regions of interest. To adequately process large input images (in terms of numbers of voxels), such as those of the lungs, the nn-UNet uses a 3D UNet cascade-based architecture to provide patch-based sampling at different image resolutions (Isensee *et al.*, 2018).

Various other mechanisms have been developed to improve the performance of fCNN architectures, including attention-mechanisms (Oktay *et al.*, 2018) and densely connected layers (Gibson *et al.*, 2018a), each providing improved performance for specific tasks during network training and inference. Hesamian *et al.* (2019) have reviewed in detail several common CNN architectures used for medical image segmentation. The development of novel CNN architectures is an ongoing area of research.

2.3.4 Patch-based sampling

The input of a CNN is an image. This can be the whole image; however, in most cases, due to memory constraints, a small patch of the image is used as an input. In addition to reducing the memory requirements of the CNN, patch-based analysis also provides a regularisation function where network weights must perform consistently across a range of locations within the image, which may exhibit high-resolution features, as well as across various training images. Several types of patch-based sampling can be employed; these include overlapping or non-overlapping configurations, uniform or grid sampling methods, and the use of prior regions of interest to define sampling frequency. Overlapping and non-overlapping patch-based sampling methods are largely self-explanatory; if the spatial window size of the patch is equally divisible into the size of the total image, non-overlapping sampling can be deployed. Overlapping patch-based sampling is used when this is not the case; overlapping patches can be randomly located or arranged in a grid-like fashion. If random overlapping patches are used, the set of all feasible spatial locations are first computed, so that all patches are within the border of the image, and then random locations are drawn from this set during each pass through the CNN. If grid sampling is used, the full set of patches is computed and then arranged so that the minimal amount of overlap is present; this is effectively the same as a sliding window with a stride length equal to the size of the patch. Figure 2.13 depicts a hyperpolarised gas MRI scan with an in-plane resolution of 256×256 with examples of non-overlapping and random overlapping patch-based sampling methods.

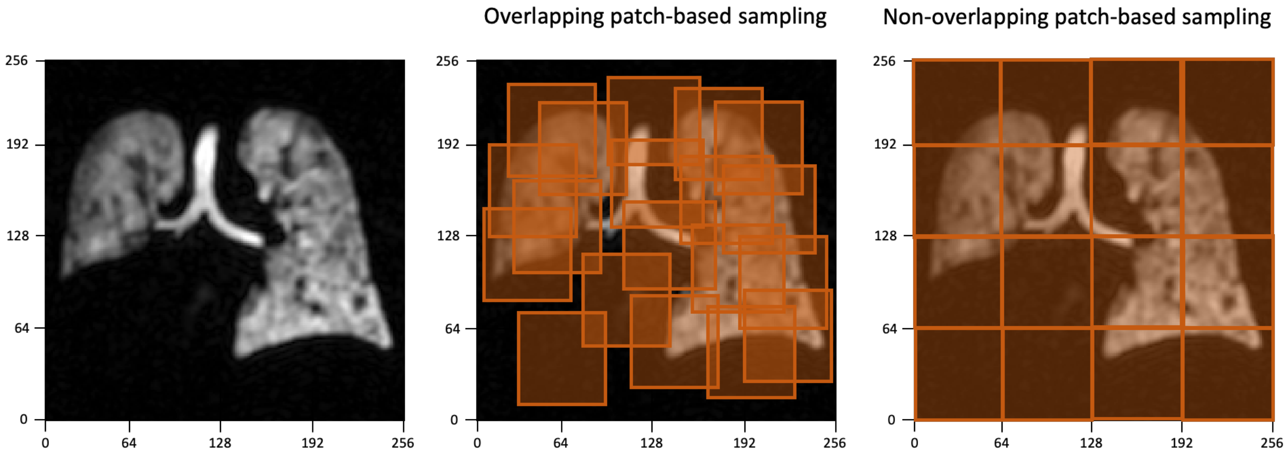

Overlapping patch-based sampling
Non-overlapping patch-based sampling
256
192
128
64
0
0
64
128
192
256

Figure 2.13 Hyperpolarised gas MRI scan with example overlapping and non-overlapping patch-based sampling methods for a patch size of 64 × 64 pixels.

If the CNN receives a 3D input, then the patch becomes a cubic volume; however, the same principles can be used to generate non-overlapping or overlapping volumetric patches. The type of patch-based sampling used can depend on the specifics of the application. For example, in hyperpolarised gas MRI segmentation, random overlapping patches can be used as minimal information is contained at the edges of the image that is relevant to the segmentation. The network's output for each patch is subsequently reassembled by an aggregator using relevant location information, either into their location in a 2D slice or at a volume-level in the 3D scan. Various network architectures have constraints on the patch size that can be used. For example, the nn-UNet requires isotropic input image patch sizes; therefore, if the original input image is anisotropic, the patches may contain large amounts of redundant information. In contrast, the VNet architecture can accommodate anisotropic patch sizes and, consequently, for hyperpolarised gas MRI scans which have a very anisotropic resolution, a patch size of $96 \times 96 \times 24$ can be utilised.

2.3.5 Memory and computational constraints

DL is notoriously computationally expensive. DL training often requires purpose built computational resources with a large memory footprint. At test time, inference can be conducted using a central processing unit (CPU) with limited processing power; however, during network training, a graphical processing unit (GPU) is required. GPUs are specialised processors which contain dedicated memory designed for performing floating point operations and rendering complex graphics. Neural networks often contain millions of parameters where matrix multiplications are computed at each layer of the network, representing a significant amount of memory. GPUs contain more cores than CPUs and have a higher memory bandwidth making them more suitable for DL calculations; the number of cores also allows for parallelisation of computational processes, further improving GPUs’ efficiency for network training. Simple matrix multiplications can be performed in parallel by each core reducing the time taken to train a CNN and increasing the size of the dataset which can be utilised. Whilst GPUs have several features which are advantageous for training a CNN, each core often contains only a limited amount of memory. Therefore, external dynamic rapid access memory (DRAM) is required to store large medical images. The amount of DRAM available will affect the patch size and batch size that can be employed. fCNNs such as the VNet employ various techniques to reduce the memory footprint of the network, including replacing pooling operations with convolutional layers, thereby reducing the computational requirements associated with processing large 3D medical images. Countless researchers are currently developing new and innovative ways to improve computational efficiency in the training of large neural networks; however, the field of DL optimization is complex and beyond the scope of this thesis. Nevertheless, computational constraints will play an important role in the investigations contained herein, whether that be by limiting the batch size which can be processed or the input patch size of the CNN.

2.4 Deep learning training strategies

When training a neural network, the performance on the testing set should be similar to the performance on the training test, indicating that the model can generalise to unseen inputs. Differences in performance between training and testing data are due to underfitting or overfitting. Underfitting occurs when the ANN is unable to encapsulate the relationship between the input and output domains, leading to reduced performance on both the training and testing sets; an underfit ANN is too simplistic and cannot account for important features in the training data. Conversely, overfitting occurs when the ANN learns both features and noise in the training data such that they are no longer applicable to new, unseen inputs. This results in good performance on training data but poor performance on testing data. Both overfitting and underfitting limit a network's capacity to generalise to new inputs and, therefore, the model's usefulness is reduced. The presence of overfitting and underfitting is also referred to as the bias vs variance trade-off, whereby a network with a high bias experiences underfitting and a network with high variance exhibits overfitting. The ideal, or optimal, model is one with low bias and low variance; however, there is often a trade-off between minimising these sources of error. A visual representation of underfitting and overfitting is displayed in Figure 2.14.

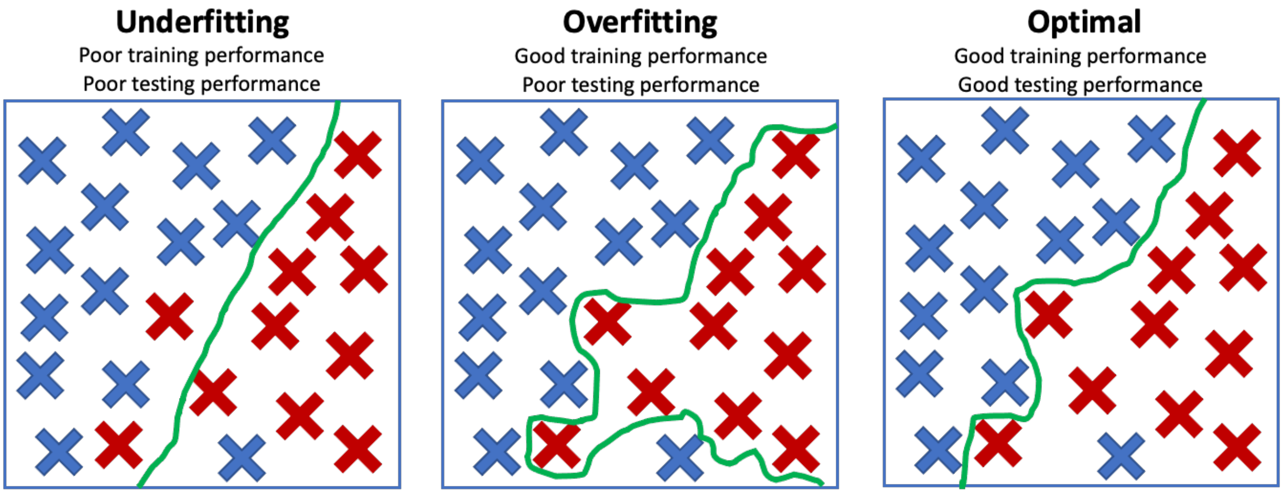


Figure 2.14 Diagram of underfitting, overfitting and optimal performance of a classification algorithm.

To ensure that a network is optimally trained, a validation dataset is often used alongside training and testing sets; this allows researchers to measure performance without biasing ANNs to a specific testing set. If the networks loss deviates significantly between the training and validation sets, there is a high likelihood that the network is experiencing overfitting. In the majority of cases, the most effective method for improving model performance is to increase the amount and variation of training data. In pulmonary imaging applications, generating more data is often unfeasible; thus, other approaches can be employed to mitigate underfitting and overfitting; these include transfer learning, data augmentation, regularisation and dropout.

2.4.1 Transfer learning

Transfer learning is the process of using pre-trained models developed for one task and re-purposing them for use in a second, somewhat related task. It is primarily used to mitigate underfitting where there is a lack of available training data and, consequently, is used regularly in medical imaging to improve performance when a lack of training data is available (Tajbakhsh et al. 2016). Transfer learning can take multiple forms depending on the specific requirements; in certain cases, weights from a pre-trained ANN are used as a starting point instead of random initialisations, and in other cases, layers are added to the network where weights are fixed for the pre-trained layers and weights for the additional layers are learnt during network training. CNNs pre-trained using the ImageNet dataset are commonly used for transfer learning in the medical imaging domain (Morid *et al.*, 2021).

2.4.2 Data augmentation

Data augmentation is the process of generating new data by manipulating original training data. Augmentation can be used to mitigate both underfitting and overfitting; augmenting training data can alleviate the limitations associated with the small datasets often encountered in medical imaging tasks. Furthermore, augmentation can be used to mitigate overfitting by producing variations in the training data which lead to additional noise, reducing the potential for a network to capture all variations generated by the augmentation accurately. Data augmentation methods can include horizontal/vertical flipping, random rotations, elastic deformations and scaling. In most cases, data augmentation increases the number of training examples; however, another method for data augmentation is to not directly increase the number of training samples, but to apply random augmentations each time a training example is passed to the network. Therefore, the number of training examples remains the same as before data augmentation; this method is often used to increase the number of epochs during training and primarily guards against overfitting. A primary consideration when applying data augmentation techniques is to ensure that the augmentations are representative of images encountered in the specific use case of the network. For example, in lung imaging applications, applying horizontal flipping is unlikely to generate improved performance as there are significant differences between the left and right lungs, in terms of both the presence of the heart and the number of lobes.

2.4.3 Regularisation

Regularisation is commonly used during ANN training to minimise overfitting. Regularisation works by adding additional terms to the loss function that update the value of weights within the network to increase sparseness and reduce model complexity. Three main regularisation parameters are employed in DL experiments, namely, L1 regularisation, L2 regularisation and dropout.

L1 regularisation employs lasso regression that adds an absolute value of the magnitude to the loss function that becomes undifferentiable at 0. This incentivises weights close to 0 to be 0 and removes connections within the network and hence removes features with low activations. Incentivising weights to be 0 increases the sparsity of matrices, often leading to less overfitting due to the pruning of less important features which may represent variation in the training data only. L1 regularisation has the effect of reducing complexity in the model and acts as an in-built feature selection tool that is robust to outliers.

L2 regularisation employs ridge regression that adds a squared value of the magnitude to the loss function. L2 regularisation does not increase sparsity; however, it achieves the same goal of reducing overfitting. This is achieved by both discouraging large weights in the matrix and encouraging smaller weights to be closer to 0. This removes overreliance on some features and minimises the effect of less important features which are still useful in complex problems; however, it is not robust to outliers.

Dropout is a form of regularisation that randomly removes connections between neurons within the network; this introduces an element of randomness to the network and reduces the model's complexity. It is often combined with L1 and L2 regularisation to limit overfitting. Dropout can be applied at each layer of the network during network training, dropout is primarily used after dense layers rather than after convolutional layers. These dropouts are not permanent as a proportion of neurons are temporarily removed at each epoch of network training to facilitate improved learning; however, during inference dropout is removed and therefore if a trained network is provided for use dropout will not be present in this trained model once training has been completed. The level of dropout used will depend on the level of overfitting exhibited by the network. It is generally advised that dropout should be below 50% of the number of neurons to maximise learning performance.

2.5 Deep learning validation strategies

To determine if a model is overfit or underfit and if the training strategies previously described have generated a model that is able to effectively generalise, various validation strategies can be employed. The most basic is dividing the dataset into training and testing sets. As previously described, it is also commonplace to also define a validation dataset to assess underfitting or overfitting during network training; however, this approach is limited due to the small subsection of data used for validation. To avoid this problem, cross-validation can be employed; cross-validation results in a series of models where each is trained and tested on subsections of the dataset. Several techniques, including K-fold, stratified and Monte-Carlo, cross validation are discussed. In addition, another method for validating ANNs is to use an external validation dataset. This provides an understanding of the network's generalisability to new cases and its resilience to domain shift.

2.5.1 K-fold cross validation

K-fold cross validation involves splitting the dataset into K subsections where K models are trained; each of the trained models is tested on one of these subsections with the remaining subsections used for network training. This is repeated until K models have been trained, so that each training image will have been represented in the testing set once. The error is subsequently averaged over all folds. In traditional K-fold cross-validation, the images contained within each subset, or fold, are defined randomly. If $K = N$, where N represents the number of images in the dataset, this is referred to as leave-one-out cross-validation. Figure 2.15 contains a visual representation of a K-fold cross validation where $K = 5$.

2.5.2 Stratified cross-validation

Stratified cross-validation also splits the dataset into K parts where K models are trained. However, instead of randomly defining which images are contained within each fold, the folds are constructed to ensure each fold is representative of the dataset as a whole. For example, in lung imaging, a dataset may contain numerous pulmonary pathologies with varying levels of representation; in stratified cross-validation, each of these folds would contain pulmonary pathologies with the same representation as the whole dataset. Stratified cross-validation is primarily used when the dataset has large class imbalances.

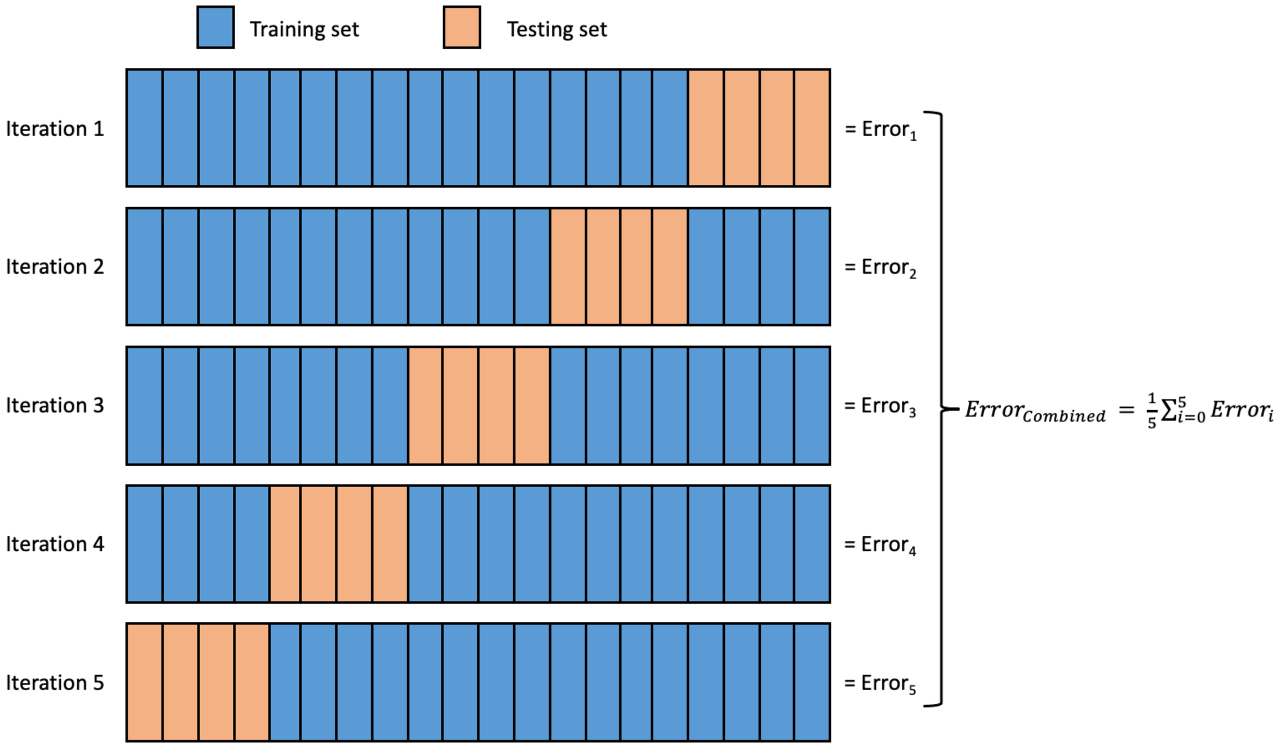


Figure 2.15 Visual representation of K-fold cross-validation where K=5.

2.5.3 Monte-Carlo cross validation

Monte-Carlo cross validation splits the data into training and testing sets; however, this is repeated with random subsections of the data and a variety of pre-defined random splits; for example, in the first cross-validation iteration, an 80/20% split is used, and in the second iteration, a 75/25% split is used. Unlike in K-fold cross-validation, the number of iterations is not limited to K number of subsections because in each cross-validation iteration, a random subsection of the data is used; therefore, each training image is represented in the testing cohort any number of times. Hence, there are an infinite number of possible iterations. However, since an image can be included in the testing set multiple times, it is possible that the results can be biased to specific cases. Consequently, Monte-Carlo cross-validation is used when the data is largely balanced. Figure 2.16 contains a visual representation of a Monte-Carlo cross-validation with splits of 80/20% and 75/25% for N number of iterations.

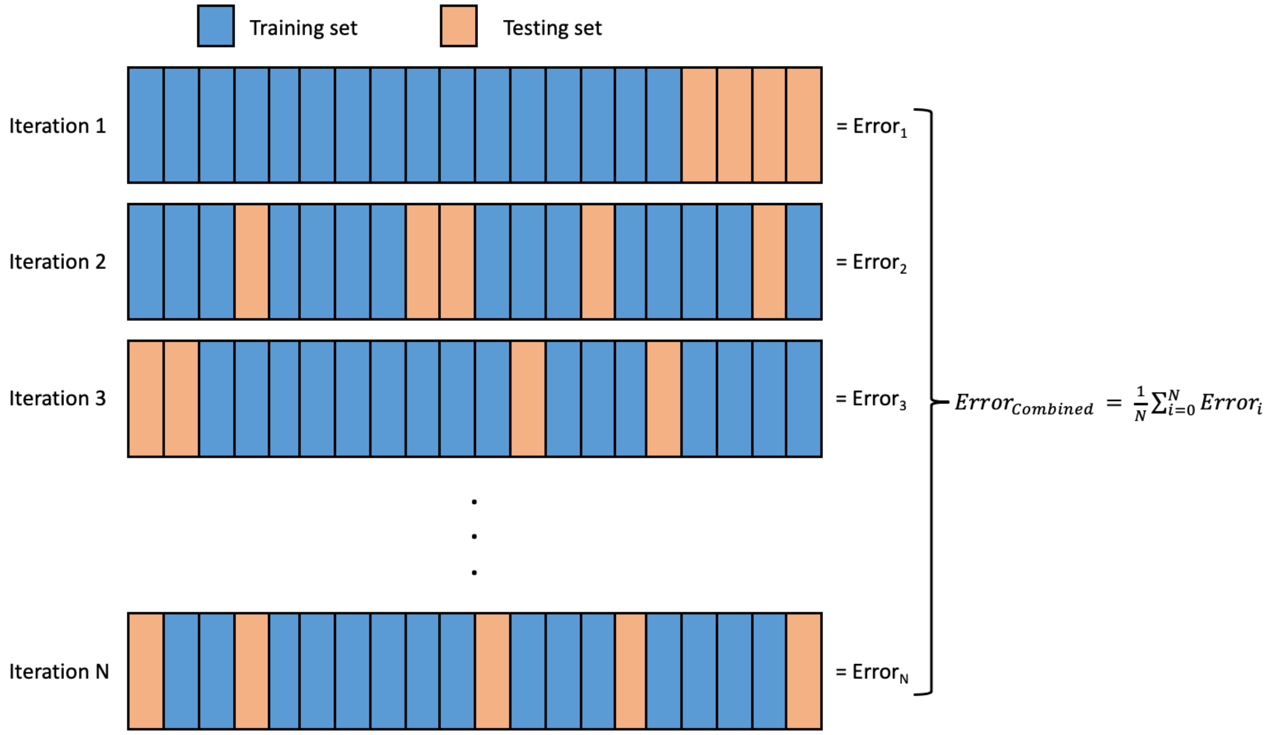


Figure 2.16 Visual representation of Monte-Carlo cross-validation for N number of iterations with data splits of 80/20% and 75/25%.

2.5.4 External validation

External validation is regularly used alongside cross validation techniques; it is often considered the most effective form of validation. External validation datasets use data from a different centre, scanner or disease to validate a model trained on data that does not contain these features, potentially leading to domain shift. Domain shift occurs when the data distribution significantly differs between the internal and external validation datasets, leading to reduced performance. If the network performs similarly on cross-validation and external validation data, it is a clear sign of generalisability to new cases and, thus, a lack of overfitting. The presence of a testing cohort not contained in the dataset used for cross-validation gives the opportunity to compare the performance of each cross-validation model on the same set of images; large variations in performance between cross-validation models on the external validation dataset are indicative of potential overfitting.

2.6 Related works

The following section will detail several papers which are related to the work presented in this thesis. A brief outline of each paper's methods, results and limitations are outlined with a view to understand how the work presented in this thesis builds upon these related investigations and represents significant improvements over previously developed approaches.

Tustison *et* al. used a 2D UNet for hyperpolarised gas MRI ventilated lung segmentation on a dataset of 113 scans, developing a novel template-based method to augment the lung imaging data alongside several pre-processing techniques, including N4 bias correction and adaptive denoising (Tustison *et al.*, 2019). The 2D CNN achieved a mean ± SD Dice similarity coefficient (DSC) of 0.94 ± 0.03 on ventilated lung regions from 40 testing set cases. In addition, Tustison *et* al. evaluated a 3D UNet CNN for isotropic ^{1}H-MRI lung cavity segmentation; a dataset of 268 scans was utilised with 62 ^{1}H-MRI scans used for testing. The proposed approach achieved a mean ± SD DSC of 0.94 ± 0.02 (Tustison *et al.*, 2019). The approach used relatively small datasets for both hyperpolarised gas and ^{1}H-MRI segmentation with a limited number of lung pathologies present in the dataset. For hyperpolarised gas MRI ventilated lung segmentation, although the dataset contained both ^{129}Xe and ^{3}He hyperpolarised gas MRI scans, no analysis was undertaken to distinguish between the two nuclei. Furthermore, the approach utilises a 2D CNN which cannot account for features which occur across several 2D slices. Singular hyperpolarised gas MRI and ^{1}H-MRI image acquisition protocols were used to acquire the dataset; due to the small number of acquisition protocols and the limited amount of lung pathologies, the proposed segmentation approaches demonstrated limited generalisability (Tustison *et al.*, 2019).

Zha *et* al. used a 2D UNet to segment the lung cavity on UTE ^{1}H-MRI scans utilising a dataset of 45 ^{1}H-MRI scans from healthy, asthma and CF participants (Zha *et al.*, 2019). 5-fold cross validation was used to maximise the size of the testing set. Various preprocessing strategies were employed including denoising, bias field correction and masking images so that only participants' bodies were present. The proposed approach achieved a mean ± SD DSC of 0.97 ± 0.02 for the right lung and 0.96 ± 0.01 for the left lung. The generalisability of this method was not demonstrated due to the small dataset which contained only 45 ^{1}H-MRI scans from a limited number of diseases and a single image acquisition sequence (Zha *et al.*, 2019). Furthermore, extensive preprocessing was utilised, such as generating body masks, which adds an additional step to the proposed workflow, potentially increasing the processing time.

Ren *et* al. (2022) have shown the capability of deriving synthetic perfusion maps from CT using SPECT perfusion as ground truth. A dataset comprising 33 lung cancer patients and 137 non-lung cancer patients utilising stratified 3-fold cross-validation to increase the number of testing examples was employed. A pre-trained 3D UNet CNN was utilised, where images were preprocessed by clipping intensity values between -1000 and -300 HU. The proposed approach achieved a voxel-wise Spearman's correlation of 0.64 averaged across all lobes and a DSC value of 0.81 for both high-functional and low-functional lung regions (Ren *et al.*, 2022). The dataset contained only one pulmonary pathology, namely, lung cancer, reducing the generalisability of the proposed approach if applied to patients with other lung diseases. Furthermore, the resolution of SPECT is limited (~10-20mm^3) which leads to less detailed outputs when compared to other functional lung imaging modalities, such as hyperpolarised gas MRI.

Liu *et* al. (2020) proposed a CNN-based approach to synthesise Technegas SPECT ventilation images from non-contrast 4DCT. A dataset of 50 participants which contained lung and oesophageal cancer patients was utilised for this investigation, whereby 10-fold cross-validation was employed for CNN training (Liu *et al.*, 2020). A 2D UNet CNN was trained, producing 2D synthetic ventilation scans by combining 2D ventilation surrogate outputs; normalisation was employed prior to network training. The authors indicate that, after median filtering, the proposed approach achieved mean Spearman's correlations of 0.73 and 0.71 for 10-phase and 2-phase 4DCT, respectively. Mean DSC across all folds of 0.83 for high-functional lung regions, 0.61 for medium-functional lung regions and 0.73 for low-functional lung regions were reported (Liu *et al.*, 2020). The dataset contained only 50 participants with thoracic malignancies, limiting the ability to deploy the CNN in participants with other lung pathologies. As previously stated, the resolution of SPECT is somewhat limited which is further compounded by the use of median filtering as a preprocessing step, further reducing the resolution of synthetic ventilation scans generated by the proposed CNN.

Capaldi *et al.* (2020) used a 2D UNet CNN with a MAE loss function to generate ventilation maps of a single 2D coronal section from free-breathing ^{1}H-MRI (Capaldi *et al.*, 2020). Image normalisation was employed before scans were provided to the network. The dataset contained 114 participants with various pulmonary pathologies; 6-fold cross-validation was employed to maximise the size of the testing set. The proposed approach achieved a Pearson correlation of 0.87 when synthetic ventilation scans were correlated with ^{3}He hyperpolarised gas MRI (Capaldi *et al.*, 2020). Furthermore, the authors segmented both the ventilated portion and defect portion of synthetic ventilation scans. These segmentations were compared to expert segmentations derived from hyperpolarised gas MRI scans, achieving a mean DSC of 0.90 and 0.37 for ventilated and defect lung regions, respectively (Capaldi *et al.*, 2020). The developed methodology utilised a 2D CNN and, therefore, generated only 2D intensity maps on specific coronal sections. Consequently, this method cannot contextualise the volumetric nature and spatial clustering of ventilation defects (Donovan and Kritter, 2015). This can lead to discontinuities between slices which reduces the plausibility of ventilation defect patterns in DL-based ventilation surrogates. In addition, the dataset used is relatively small with all data being acquired using the same acquisition protocol.

Chapter 3
Deep learning in structural and functional lung image analysis: a review

The recent resurgence of deep learning (DL) has dramatically influenced the medical imaging field. Medical image analysis applications have been at the forefront of DL research efforts applied to multiple diseases and organs, including those of the lungs. The aims of this review are twofold: (i) to briefly overview DL theory as it relates to lung image analysis; (ii) to systematically review the DL research literature relating to the lung image analysis applications of segmentation, reconstruction, registration and synthesis. The review was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. 479 studies were initially identified from the literature search with 82 studies meeting the eligibility criteria. Segmentation was the most common lung image analysis DL application (65.9% of papers reviewed). DL has shown impressive results when applied to segmentation of the whole lung and other pulmonary structures. DL has also shown great potential for applications in image registration, reconstruction and synthesis. However, the majority of published studies have been limited to structural lung imaging with only 12.9% of reviewed studies employing functional lung imaging modalities, thus highlighting significant opportunities for further research in this field. Although the field of DL in lung image analysis is rapidly expanding, concerns over inconsistent validation and evaluation strategies, inter-site generalisability, transparency of methodological detail and interpretability need to be addressed before widespread adoption in clinical lung imaging workflows.

3.1 Preface

The majority of the material in this chapter was originally published as an invited review article in the *British Journal of Radiology*:

> **Astley J.R,** Wild J.M. and Tahir B.A. (2020). Deep learning in structural and functional lung image analysis. *The British Journal of Radiology*. 20201107. 10.1259/bjr.20201107.

This article was published under an Open Access Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format (https://creativecommons.org/licenses/by/4.0/). Slight modifications have been made to the published version.

3.1.1 Author contributions

J.R.A conducted the literature search and was responsible for drafting the manuscript. J.M.W and B.A.T consulted on the definition of search terms and provided feedback or comments on versions of the draft manuscript.

3.2 Introduction

Respiratory diseases constitute significant global health challenges; five respiratory diseases are among the most common causes of death. 65 million people suffer from chronic obstructive pulmonary disease (COPD) and 339 million from asthma (GBD15 *et al.*, 2016; Vos *et al.*, 2017). There are 1.8 million new lung cancer cases diagnosed annually and 1.6 million deaths worldwide, making it the most common and deadliest cancer on the planet (Torre *et al.*, 2015). Lung imaging is a critical component of respiratory disease diagnosis, treatment planning, monitoring and treatment assessment. Acquiring lung images, processing them and interpreting them clinically are crucial to achieving global reductions in lung-related deaths. Traditionally, the techniques employed to quantitatively analyse these images evolved from the disciplines of computational modelling and image processing; however, in recent years, deep learning (DL) has received significant attention from the lung imaging community.

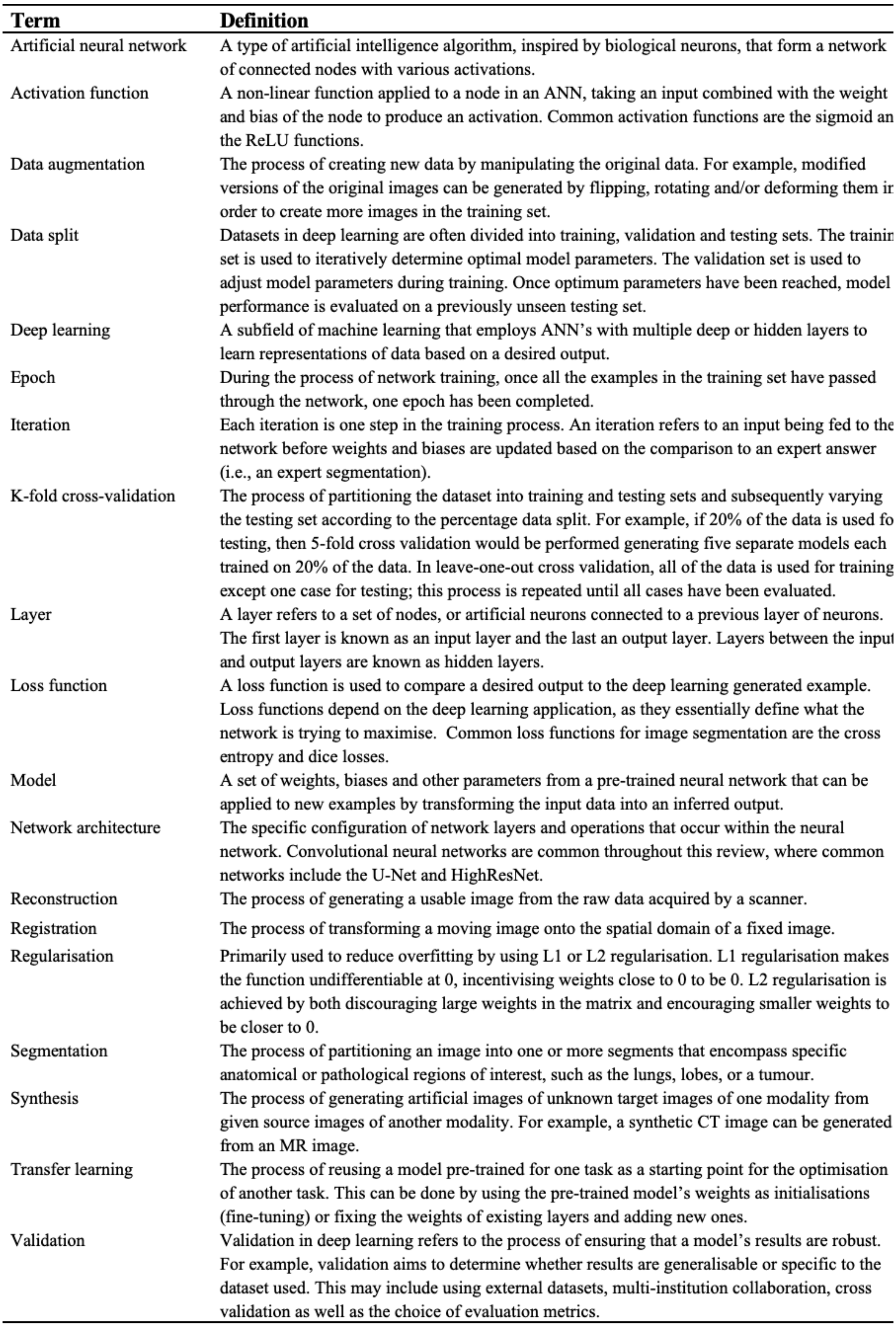
DL is a subfield of machine learning that employs artificial neural networks with multiple deep or hidden layers. Whilst the fundamental theory was posited several decades ago (Berners-Lee, 1968), DL gained international interest in 2012 when AlexNet, a type of neural network referred to as a convolutional neural network (CNN), won the ImageNet Large Scale Visual Recognition Challenge. That paper has been cited over 47,000 times and triggered a renaissance in DL research (Krizhevsky *et al.*, 2012). Subsequently, CNNs, and DL more generally, began to impact the medical imaging field profoundly. Development of fully convolutional networks such as V-Net demonstrated how deep-layered architectures could provide valuable functions in solving some of the field’s most critical applications, including common image analysis tasks (Milletari *et al.*, 2016; de Vos *et al.*, 2017). Increased computational power due to the reduced cost of graphical processing units (GPUs) and publicly available annotated imaging datasets have since led to rapid developments and applications (Yang *et al.*, 2018). This review assesses the current literature on DL's role in lung image analysis applications, discusses critical limitations for clinical adoption, and sets out a roadmap for future research.

3.3 Deep learning theory

3.3.1 Artificial neural networks

An artificial neural network (ANN), inspired by biological neurons, can be thought of as a series of connected nodes containing weights and biases which are combined using an activation function to produce an activation; the activation determines the strength of connections within the network. At the heart of DL is optimisation; an ANN learns by optimising weights and biases for a generalisable solution. This optimisation occurs in a two-step process of forward propagation and backpropagation. A basic diagram of an ANN with two hidden layers and generalised examples of forward propagation and backpropagation are shown in Figure 3.1. The use of hidden layers in the network allows more freedom for the weights and biases to be optimised. Forward propagation refers to the process of feeding an example to the network during training where the output of the neural network is compared to a desired output and a loss is calculated using a loss function. Backpropagation uses this loss to propagate changes in weights and biases throughout the network; thus, by continually providing new examples, known as iterations, the model is optimised to approximate the function between the input and output domains. Table 3.1 provides a glossary of the key technical terms used in this review.

Table 3.1 Glossary of key technical terms related to deep learning and image analysis.



Term	Definition
Artificial neural network	A type of artificial intelligence algorithm, inspired by biological neurons, that form a network of connected nodes with various activations.
Activation function	A non-linear function applied to a node in an ANN, taking an input combined with the weight and bias of the node to produce an activation. Common activation functions are the sigmoid an the ReLU functions.
Data augmentation	The process of creating new data by manipulating the original data. For example, modified versions of the original images can be generated by flipping, rotating and/or deforming them ir order to create more images in the training set.
Data split	Datasets in deep learning are often divided into training, validation and testing sets. The trainin set is used to iteratively determine optimal model parameters. The validation set is used to adjust model parameters during training. Once optimum parameters have been reached, model performance is evaluated on a previously unseen testing set.
Deep learning	A subfield of machine learning that employs ANN's with multiple deep or hidden layers to learn representations of data based on a desired output.
Epoch	During the process of network training, once all the examples in the training set have passed through the network, one epoch has been completed.
Iteration	Each iteration is one step in the training process. An iteration refers to an input being fed to the network before weights and biases are updated based on the comparison to an expert answer (i.e., an expert segmentation).
K-fold cross-validation	The process of partitioning the dataset into training and testing sets and subsequently varying the testing set according to the percentage data split. For example, if 20% of the data is used fo testing, then 5-fold cross validation would be performed generating five separate models each trained on 20% of the data. In leave-one-out cross validation, all of the data is used for training except one case for testing; this process is repeated until all cases have been evaluated.
Layer	A layer refers to a set of nodes, or artificial neurons connected to a previous layer of neurons. The first layer is known as an input layer and the last an output layer. Layers between the input and output layers are known as hidden layers.
Loss function	A loss function is used to compare a desired output to the deep learning generated example. Loss functions depend on the deep learning application, as they essentially define what the network is trying to maximise. Common loss functions for image segmentation are the cross entropy and dice losses.
Model	A set of weights, biases and other parameters from a pre-trained neural network that can be applied to new examples by transforming the input data into an inferred output.
Network architecture	The specific configuration of network layers and operations that occur within the neural network. Convolutional neural networks are common throughout this review, where common networks include the U-Net and HighResNet.
Reconstruction	The process of generating a usable image from the raw data acquired by a scanner.
Registration	The process of transforming a moving image onto the spatial domain of a fixed image.
Regularisation	Primarily used to reduce overfitting by using L1 or L2 regularisation. L1 regularisation makes the function undifferentiable at 0, incentivising weights close to 0 to be 0. L2 regularisation is achieved by both discouraging large weights in the matrix and encouraging smaller weights to be closer to 0.
Segmentation	The process of partitioning an image into one or more segments that encompass specific anatomical or pathological regions of interest, such as the lungs, lobes, or a tumour.
Synthesis	The process of generating artificial images of unknown target images of one modality from given source images of another modality. For example, a synthetic CT image can be generated from an MR image.
Transfer learning	The process of reusing a model pre-trained for one task as a starting point for the optimisation of another task. This can be done by using the pre-trained model's weights as initialisations (fine-tuning) or fixing the weights of existing layers and adding new ones.
Validation	Validation in deep learning refers to the process of ensuring that a model's results are robust. For example, validation aims to determine whether results are generalisable or specific to the dataset used. This may include using external datasets, multi-institution collaboration, cross validation as well as the choice of evaluation metrics.

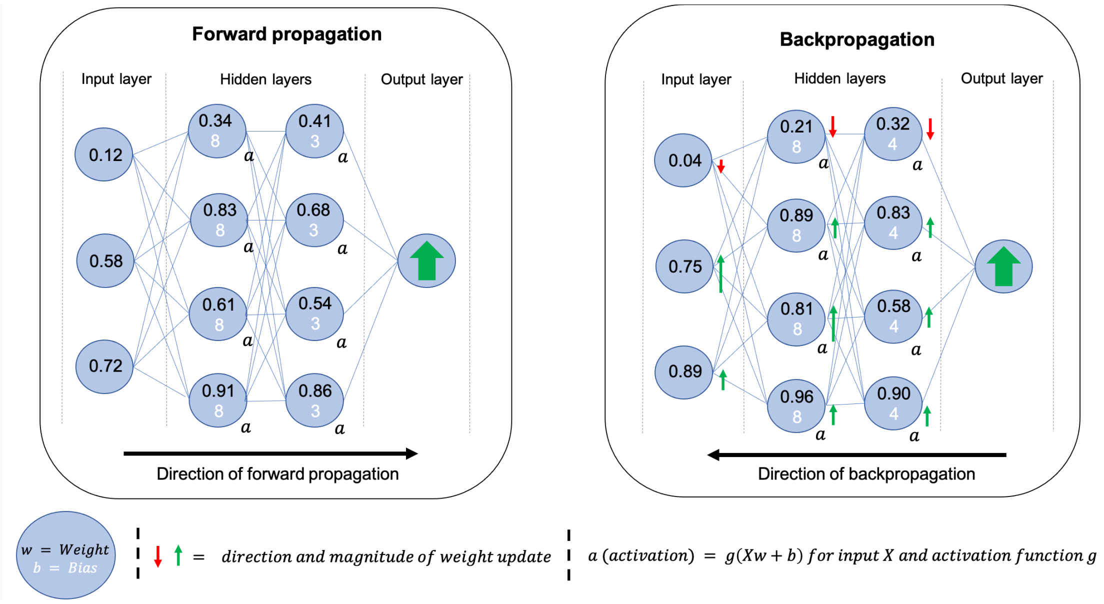


Figure 3.1 Simplified diagrams of the processes of forward propagation (left) and backpropagation (right) for a neural network with two hidden layers. The neural network is represented as a series of nodes, each of which contains a weight and bias. The weight and bias are combined using the activation function to produce an activation that impacts the strength of connections within the network. Once an input has been passed through the network, it is compared to a desired output, such as an expert segmentation of an anatomical region of interest, to produce a loss. This loss is used to propagate changes to weights and biases, hence, changing the strength of connections for the subsequent example.

The structure of a DL network is known as an architecture. In the medical imaging field, three key architectures, namely, CNNs, recurrent neural networks (RNNs) and generative adversarial networks (GANs) are particularly prevalent. These structures are outlined in Figure 3.2. Understanding specific architectures such as V-Nets and GANs requires an in-depth understanding of complex linear algebra and matrix manipulation and is beyond this review's scope; the interested reader is directed to several excellent papers on the subject (Milletari *et al.*, 2016; Kazeminia *et al.*, 2018; Goodfellow *et al.*, 2014).

3.3.2 Pre-processing

Before images are fed into a neural network, they are frequently processed, often by accentuating differences between foreground and background voxels, to enhance performance and/or reduce training time. DL theory suggests that in high-dimensional matrices, local minima are very unlikely; instead, saddle points are more common due to the improbable likelihood that every dimension produces a minimum at the same location. These techniques can decrease the likelihood that the algorithm reaches a shallow saddle point, thereby causing slower optimisation. This is achieved through regularisation techniques and limiting outlier intensities. Cropping is regularly used to restrict the processing to voxels within the patient (Jiang *et al.*, 2018), or coarse, manually-drawn bounding boxes (Negahdar *et al.*, 2018). Table 3.2 summarises commonly used pre-processing techniques in the DL lung image analysis literature. In CNNs, other techniques such as batch normalisation, have been shown to reduce training time, acting as secondary regularisation techniques to minimise outliers and improve performance (Ioffe and Szegedy, 2015; Huang *et al.*, 2018)

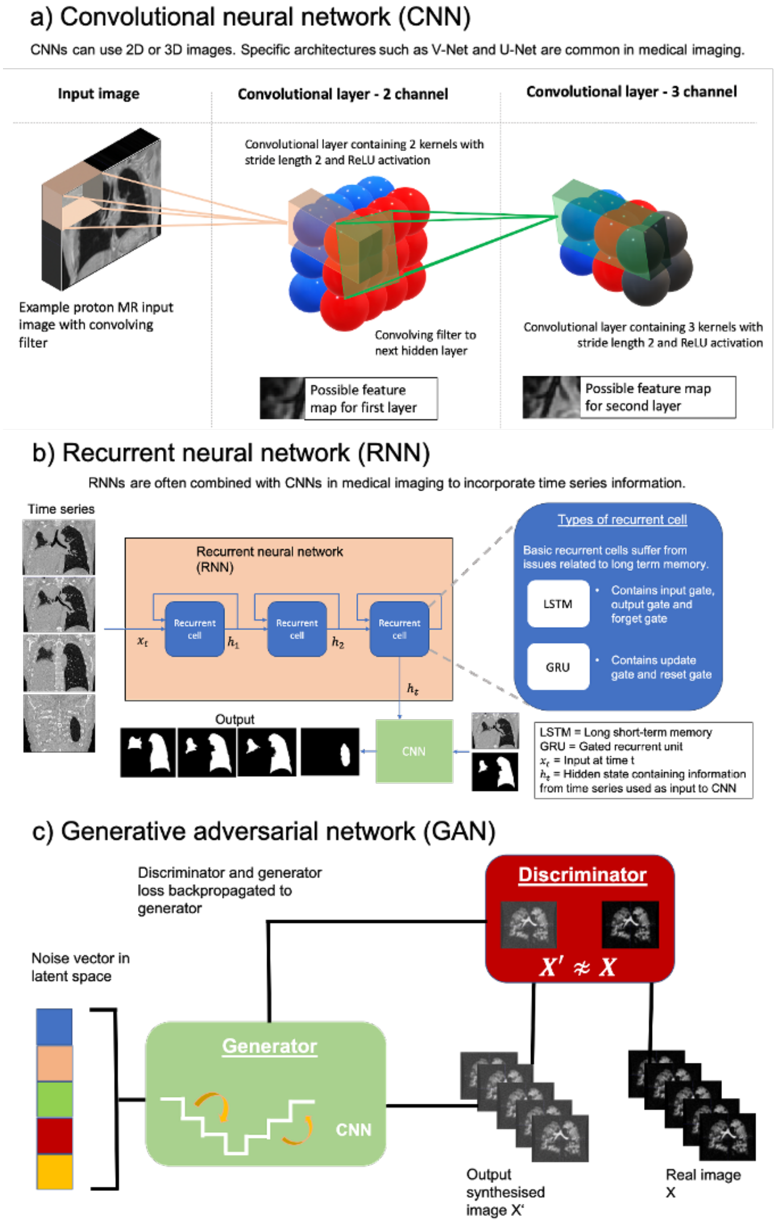

a) Convolutional neural network (CNN)
CNNs can use 2D or 3D images. Specific architectures such as V-Net and U-Net are common in medical imaging.
Input image
Convolutional layer - 2 channel
Convolutional layer - 3 channel
Convolutional layer containing 2 kernels with stride length 2 and ReLU activation
Example proton MR input image with convolving filter
Convolving filter to next hidden layer
Convolutional layer containing 3 kernels with stride length 2 and ReLU activation
Possible feature map for first layer
Possible feature map for second layer
b) Recurrent neural network (RNN)
RNNs are often combined with CNNs in medical imaging to incorporate time series information.
Time series
Recurrent neural network (RNN)
Recurrent cell
Recurrent cell
Recurrent cell
x_t
h_1
h_2
h_t
Output
CNN
Types of recurrent cell
Basic recurrent cells suffer from issues related to long term memory.
LSTM
Contains input gate, output gate and forget gate
GRU
Contains update gate and reset gate
LSTM = Long short-term memory
GRU = Gated recurrent unit
x_t = Input at time t
h_t = Hidden state containing information from time series used as input to CNN
c) Generative adversarial network (GAN)
Discriminator and generator loss backpropagated to generator
Discriminator
$X' \approx X$
Noise vector in latent space
Generator
CNN
Output synthesised image X'
Real image X

Figure 3.2 Illustration of three common types of deep learning architectures used in medical imaging: a) convolutional neural network (CNN), b) recurrent neural network (RNN) and c) generative adversarial network (GAN). In the lung image analysis examples given, the CNN and RNN are used for image segmentation while the GAN is used for image synthesis.

Table 3.2 Summary of common pre-processing techniques used for lung image analysis tasks, including values prevalent in the literature. Modalities included are those for which the pre-processing techniques have been used in the reviewed studies. This is not an exhaustive list of pre-processing techniques used.

Pre-processing technique	*Description*	*Modality*	*Literature Values*	*References*
Thresholding	The process of constraining the pixel values of an image to be between predefined values.	CT, MRI	CT intensity: [-1000,-700 HU, -400,700 HU] MRI intensity: [0,667]	(Wang *et al.*, 2018b), (Sousa *et al.*, 2019), (Javaid *et al.*, 2018), (Hofmanninger *et al.*, 2020), (Jiang *et al.*, 2019), (Tahmasebi *et al.*, 2018), (Zhong *et al.*, 2019b), (Zhou *et al.*, 2019), (Park *et al.*, 2020), (Gerard *et al.*, 2019), (Yun *et al.*, 2019), (Eppenhof and Pluim, 2019), (Fu *et al.*, 2020), (Jiang *et al.*, 2020), (de Vos *et al.*, 2019), (Stergios *et al.*, 2018), (Ren *et al.*, 2019)
Normalisation and whitening	The process of transforming the distribution of image pixels to some distribution which is standardised across images.	CT, MRI, X-ray	Normalisation: [0,1] Mean/variance ≈ 0	(Wang *et al.*, 2018b), (Liu *et al.*, 2019), (Javaid *et al.*, 2018), (Hofmanninger *et al.*, 2020), (Akila Agnes *et al.*, 2018), (Novikov *et al.*, 2018), (Gaál *et al.*, 2020), (Jiang *et al.*, 2019), (Tahmasebi *et al.*, 2018), (Zhou *et al.*, 2019), (Hatamizadeh *et al.*, 2019), (Sandkühler *et al.*, 2019), (Rajchl *et al.*, 2017), (Sentker *et al.*, 2018), (Fechter and Baltas, 2020), (Jiang *et al.*, 2020), (de Vos *et al.*, 2019), (Galib *et al.*, 2020), (Ferrante *et al.*, 2018), (Stergios *et al.*, 2018), (Beaudry *et al.*, 2019), (Duan *et al.*, 2019), (Liu *et al.*, 2020), (Ren *et al.*, 2019), (Olberg *et al.*, 2018)

Denoising	The process of removing noise from images in order to improve their quality.	CT, MRI	Gaussian, adaptive patch-based	(Xu and Liu, 2017), (Zha *et al.*, 2019), (Tustison *et al.*, 2019)
Bias correction	A technique to correct for the low-frequency bias field that corrupts MR images.	HP gas MRI, MRI	N3/N4 bias correction	(Tustison *et al.*, 2019), (Zha *et al.*, 2019), (Rajchl *et al.*, 2017)
Cropping	Cropping refers to the process of removing unwanted outer pixels or voxels of an image prior to being inputted to the network. This includes cropping by manually-defined regions of interest or external body masks. Cropping is commonly used to reduce computational cost and/or eliminate the influence of background voxels.	CT, MRI, X-ray, PET	Cropping to body mask, specific organ or manually-defined region.	(Negahdar *et al.*, 2018), (Soans and Shackleford, 2018), (Zhu *et al.*, 2019), (Hofmanninger *et al.*, 2020), (Zha *et al.*, 2019), (Hooda *et al.*, 2018), (Mittal *et al.*, 2018), (Jiang *et al.*, 2018), (Zhao *et al.*, 2018), (Zhou *et al.*, 2019), (Moriya *et al.*, 2018), (Kalinovsky *et al.*, 2017), (Sandkühler *et al.*, 2019), (Anthimopoulos *et al.*, 2019), (Gao *et al.*, 2016), (Rajchl *et al.*, 2017), (Wang *et al.*, 2019), (Garcia-Uceda Juarez *et al.*, 2019), (Juarez *et al.*, 2018), (Eppenhof and Pluim, 2019), (Sentker *et al.*, 2018), (Fechter and Baltas, 2020), (Blendowski and Heinrich, 2019), (Zhong *et al.*, 2019a), (Liu *et al.*, 2020), (Olberg *et al.*, 2018)

3.4 Validation

Validation is used to evaluate the performance of trained DL networks and assess their generalisability to non-experimental settings. The goal is to develop a validation strategy that best represents the situation in which the algorithm is to be deployed.

3.4.1 Evaluation metrics

It is imperative to evaluate the performance of DL algorithms accurately. Evaluation metrics are used to compare DL-based outputs with ground-truth segmentations or images. A large selection of evaluation metrics are used to assess the quality of segmentations, registered scans and synthesised images. Overlap-based evaluation metrics are used to assess the accuracy of segmentations by comparing the overlap of voxels in comparative segmentations; these include the Dice similarity coefficient (DSC), Jaccard similarity coefficient (JSC) and the relative error (XOR) metrics. Distance-based evaluation metrics are used to assess the accuracy of segmentations by comparing the distance between all voxels, or boundary voxels, in comparative segmentations; these include the Hausdorff distance (HD), average boundary Hausdorff distance (Avg HD) and the Hausdorff 95th-percentile (HD95) metrics. Error-based metrics aim to quantify the bidirectional error between images with continuous intensity values; this can include registered or synthesised images. The mean square error (MSE), root mean square error (RMSE) and the mean absolute error (MAE) are examples of error-based evaluation metrics. The target registration error (TRE) can also be used to assess the accuracy of registered images by comparing the location of landmarks which are defined based on physical locations within the image. Similarity-based metrics quantify the structural similarity between images with continuous intensity values. The structural similarity index measure (SSIM), multiscale-SSIM (MS-SSIM) and the normalised cross-correlation (NCC) are examples of common similarity-based metrics.

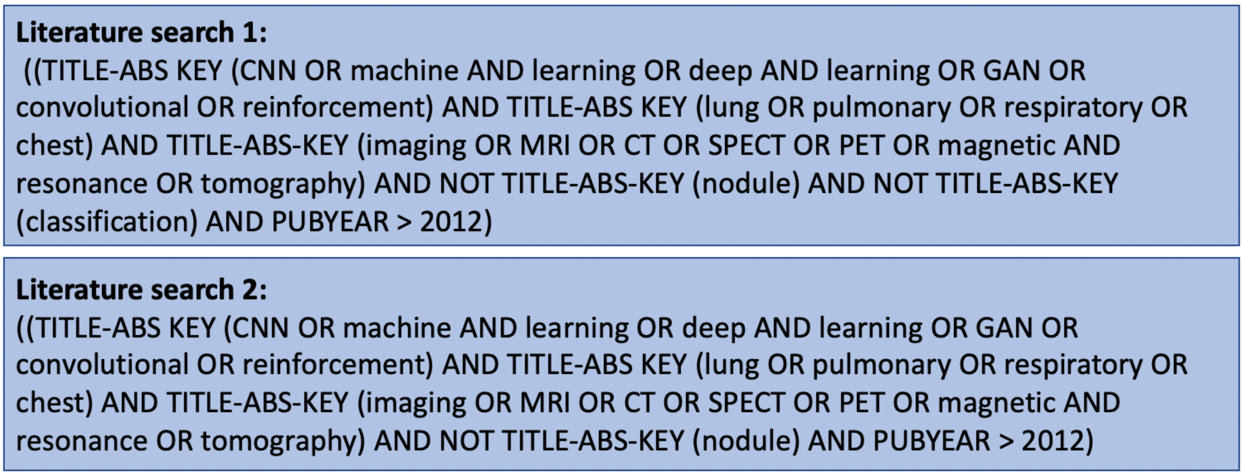
3.4.2 Validation techniques

Aside from the training set, an internal validation set is commonly used for tuning DL parameters to improve performance. A testing set is then used to provide an unbiased evaluation of performance on unseen data. In this review, validation sets used throughout the training phase are counted as training sets as the network has previously seen these images before testing. Therefore, the data split is the percentage of the total data used for training and internal validation versus that used for testing. Maintaining completely separate testing sets is somewhat uncommon in the literature and represents the ideal form of validation (Yun *et al.*, 2019; Gerard *et al.*, 2019; Dai *et al.*, 2018). Validating on external multicentre datasets that have not been used for training should be the gold standard in ensuring comparison between methods and generalisability (Bluemke *et al.*, 2019). However, this is uncommon as single-centre datasets, split into training and testing sets, are frequently used. To make the validation process more robust and generalisable, specific techniques are applied, such as k-fold cross-validation. In 4-fold cross-validation, the dataset is randomly partitioned into a 75/25% training/testing split; this process is repeated with four different 25% blocks. Another approach is leave-one-out cross-validation which uses all of the data for training except one case for testing and repeats until all cases have been evaluated.

3.5 Methods

The protocol for this literature review was performed using the preferred reporting items for systematic reviews and meta-analyses (PRISMA)-statement (Moher *et al.*, 2009). The literature search was conducted on 1 April 2020 using multiple databases (Web of Science, Scopus, PubMed) and aimed to identify studies written in English published between 1 January 2012, the same year that the seminal AlexNet paper was published (Krizhevsky *et al.*, 2012), and the date of the search. The search strategy is defined in **Error! Reference source not found.**. Further studies that met the selection criteria were identified by handsearching references and through the authors' input.

Several recent reviews have focussed primarily on DL-based lung classification and detection (Lobo and Guruprasad, 2018; Chassagnon *et al.*, 2020; Pehrson *et al.*, 2019); accordingly, this review was limited in scope to the lung image analysis applications of segmentation, registration, reconstruction and synthesis. Both published peer-reviewed scientific papers and conference proceedings were included due to recent developments in the field.



Literature search 1:
((TITLE-ABS KEY (CNN OR machine AND learning OR deep AND learning OR GAN OR convolutional OR reinforcement) AND TITLE-ABS KEY (lung OR pulmonary OR respiratory OR chest) AND TITLE-ABS-KEY (imaging OR MRI OR CT OR SPECT OR PET OR magnetic AND resonance OR tomography) AND NOT TITLE-ABS-KEY (nodule) AND NOT TITLE-ABS-KEY (classification) AND PUBYEAR > 2012)

Literature search 2:
((TITLE-ABS KEY (CNN OR machine AND learning OR deep AND learning OR GAN OR convolutional OR reinforcement) AND TITLE-ABS KEY (lung OR pulmonary OR respiratory OR chest) AND TITLE-ABS-KEY (imaging OR MRI OR CT OR SPECT OR PET OR magnetic AND resonance OR tomography) AND NOT TITLE-ABS-KEY (nodule) AND PUBYEAR > 2012)

Figure 3.3 The search strategy used on Scopus, Web of Science and PubMed to identify relevant studies for inclusion in the review. Further studies that met the selection criteria were identified by handsearching references and through the authors' input.

3.6 Results and discussion

3.6.1 Study selection

479 non-overlapping papers were retrieved. 355 papers were excluded due to not meeting the eligibility criteria. In particular, many papers focused on classification or used traditional machine learning techniques beyond this review's scope. Upon reviewing the remaining papers, 82 studies were included for analysis. The PRISMA flowchart is shown in Figure 3.4.

No studies that met the inclusion criteria were published before 2016 with the majority appearing since 2018. Image segmentation applications accounted for 65.9% of the studies reviewed. The remaining 34% are divided between synthesis, reconstruction and registration applications (Figure 3.5). Full details are shown in Figure 3.6**.** The majority of studies reviewed used structural imaging modalities (87.8%), with most using CT (63.5%). Functional lung imaging studies only constitute 12.1% of the reviewed studies and are spread across PET, SPECT and hyperpolarised gas MRI.

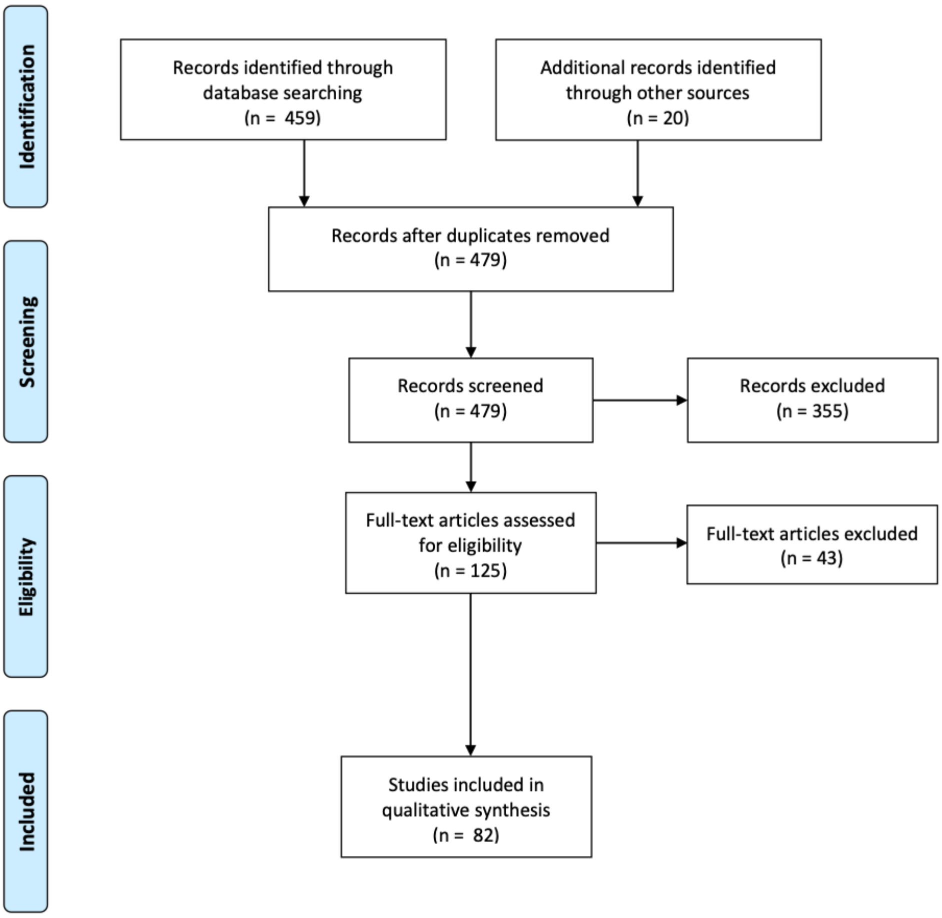

Identification
Screening
Eligibility
Included
Records identified through database searching (n = 459)
Additional records identified through other sources (n = 20)
Records after duplicates removed (n = 479)
Records screened (n = 479)
Records excluded (n = 355)
Full-text articles assessed for eligibility (n = 125)
Full-text articles excluded (n = 43)
Studies included in qualitative synthesis (n = 82)

Figure 3.4 PRISMA flowchart of studies identified, screened, assessed for eligibility and included in the literature review analysis.

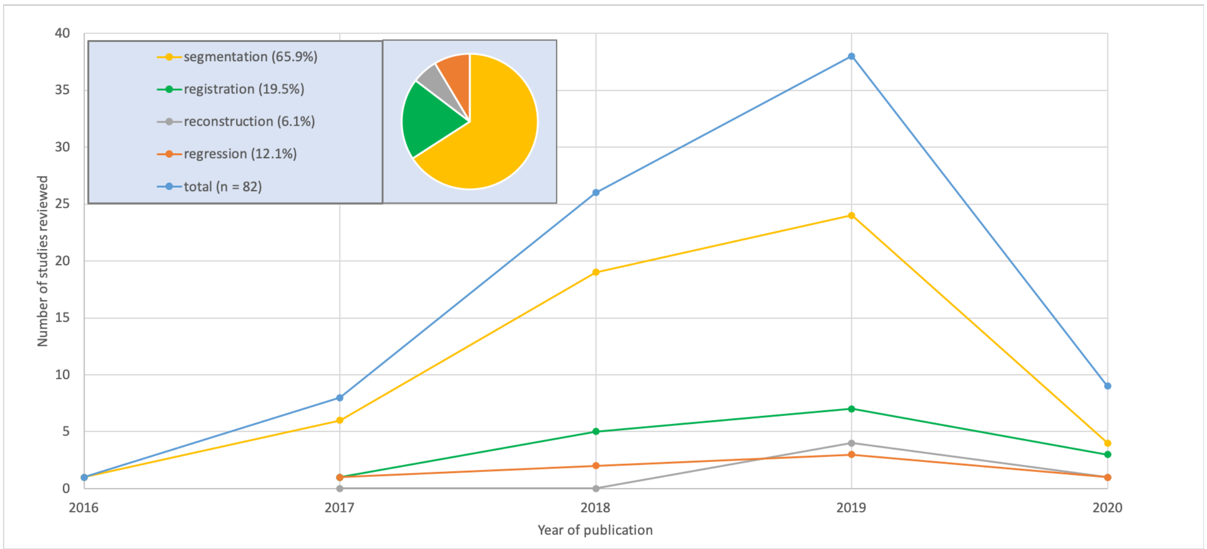


Figure 3.5 Graphical overview of the number of studies per year for the four image analysis applications considered in this review. 2020 values calculated up to 1 April 2020.

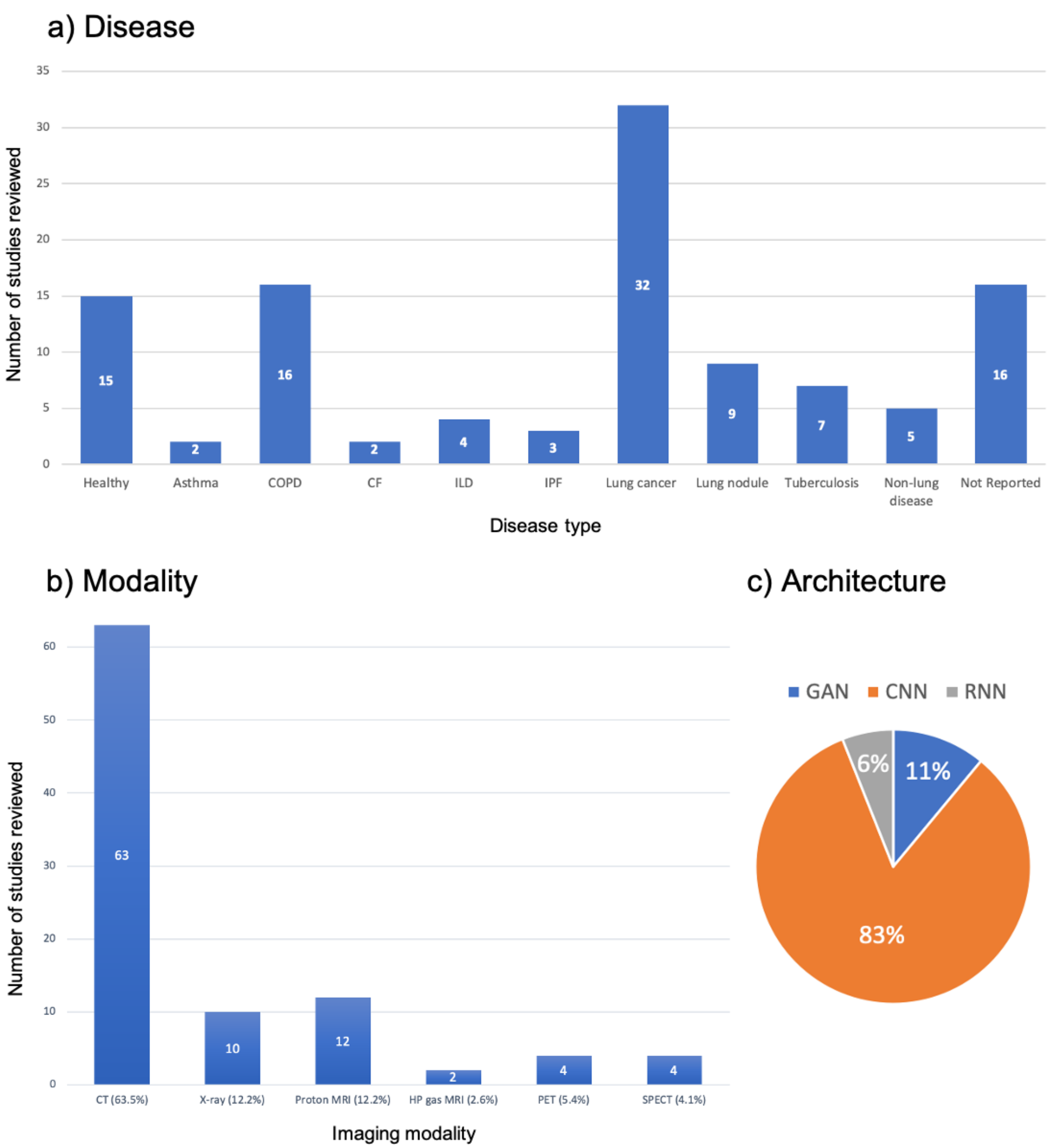


Figure 3.6 Graphical overview of deep learning lung image analysis studies reviewed by a) disease present in patient cohorts, b) imaging modality and c) architecture. Absolute numbers of papers are provided in a) and b).

3.6.2 Segmentation

Image segmentation is the process of partitioning an image into one or more segments that encompass anatomical or pathological specific regions of interest (ROIs), such as the lungs, lobes, or a tumour. Studies describing DL-based segmentation applications of pulmonary ROIs are summarised in Table 3.3.

Table 3.3 Summary of reviewed studies on deep learning for lung image segmentation. Arranged alphabetically by 'Anatomical site', then by 'Modality'.

Study	*Modality*	*ROI*	*Disease*	*Number of subjects*	*Dimensions*	*Architecture*	*Pre-processing*	*Percentage data split (training*/testing)*	*Performance*
(Wang et al., 2018b)	CT	Whole lung	COPD, IPF	575	2D	ResNet-101	Clipped -1000 to +1000 HU, Normalisation [0,1]	5-fold CV	DSC = 0.988 ± 0.012 Avg HD = 0.562 ± 0.52mm
(Dong et al., 2019)	CT	Whole lung	Lung cancer	35	3D	U-Net-GAN		LOOCV	DSC = 0.97 ± 0.01 HD95 = 2.29 ± 2.64mm Avg HD = 0.63 ± 0.63mm
(Liu et al., 2019)	CT	Whole lung	NR	100	2D	SegNet	Class grouping, Normalisation [-1000,800]	40/60	DSC = 0.98
(Lustberg et al., 2018)	CT	Whole lung	Lung cancer	470	NR	CNN		95/5	DSC = 0.99 ± 0.01 Median HD = 0.4 ±0.2cm
(Negahdar et al., 2018)	CT	Whole lung	Multiple	83	3D	V-Net	Bounding box for lung, cropped to bounding box	58/42	DSC(n=12) = 0.983±0.002 DSC(n=23) = 0.990±0.002

(Soans and Shackleford, 2018)	CT	Whole lung	Lung cancer	422	3D	CNN with spatial constraints	ROI extraction for organ localisation	71/29	ROC(Left) = 0.954 ROC(right) = 0.949
(Soliman et al., 2017)	CT	Whole lung	NR	95	3D	Deep-CNN	Post-processed hole filling	LOOCV	DSC = 0.984 ± 0.068 HD95 = 2.79 ± 1.32mm PVD = 3.94 ± 2.11%
(Sousa et al., 2019)	CT	Whole lung	Lung lesion	908	3D	Modified V-Net	Clipped [-1000, 400 HU]	98/2	Avg HD = 0.576mm DSC = 0.987
(Zhou et al., 2017)	CT	Whole lung	NR	106	2D/3D	FCN VGG16	Transfer learning from ImageNet ILSVRC-2014	95/5	JSC = 0.903 ± 0.037
(Zhu et al., 2019)	CT	Whole lung	Lung Cancer	66	3D	U-Net	Cropping to ROI	55/45	DSC = 0.95 ± 0.01 Avg HD = 1.93 ± 0.51mm HD95 = 7.96 ±2.57mm
(Gerard et al., 2018)	CT	Whole lung	COPD, IPF	1749	3D	Course-Fine ConvNet	Transfer learning from COPDGene and SPIROMICS, fine-tuned on animal model	92/8	JSC = 0.99 Avg HD = 0.29mm
(Javaid et al., 2018)	CT	Whole lung	Lung cancer	13	2D	Dilated U-Net	Only axial slices selected, clipped -	94/6	DSC = 0.99 ± 0.01 HD ≈ 4.5mm

							1000 to 3000 HU, Normalisation [0,1]		
(Xu and Liu, 2017)	CT	Whole lung	NR	20	2D	MFCNN	Gaussian denoising	50/50	DSC = 0.754
(Hu et al., 2020)	CT	Whole lung	NR	75	2D	Mask R-CNN + k-means		NR	DSC = 0.973 ± 0.032
(Hofmanninger et al., 2020)	CT	Whole lung	Multiple	266	2D	U-Net	Body mask, Clipped [-1024, 600 HU], Normalisation [0,1]	87/13	DSC = 0.98 ± 0.03 HD95 = 3.14 ± 7.4mm Avg HD = 0.62 ± 0.93mm
(Xu et al., 2019)	CT	Whole lung	Lung cancer, COPD	224	2D	1 layer CNN	Post-processed hole filling	8-fold CV	DSC = 0.967 ± 0.001 HD = 1.44 ± 0.04mm
(Tustison et al., 2019)	HP gas MRI	Functional lung	NR	113	2D	U-Net	Template-based data augmentation, N4 bias correction, denoising	65/35	DSC (HP gas) = 0.92
	Proton MRI	Whole lung	NR	268	3D	U-Net		77/23	DSC (Proton) = 0.94
(Akila Agnes et al., 2018)	LDCT	Whole lung	NR	220	2D	CDWN	Normalised [mean=0]	91/9	DSC = 0.95 ± 0.03 JSC = 0.91 ± 0.04
(Zha et al., 2019)	UTE proton MRI	Whole lung	Healthy, CF, asthma	45	2D	CED (U-Net+ autoencoder)	Denoising, bias field correction, body mask	5-fold CV	DSC (right) = 0.97 ± 0.015

									DSC (left) = 0.96 ± 0.012
(Hwang and Park, 2017)	X-ray	Whole lung	Healthy, lung nodules	247	2D	U-Net		2-fold CV	DSC = 0.980 ± 0.008 JSC = 0.961 ± 0.015 Avg HD = 0.675 ± 0.122mm Avg boundary HD = 1.237 ± 0.702mm
(Souza et al., 2019)	X-ray	Whole lung	Healthy, Tuberculosis	138	2D	ResNet-18 with FC layer	Scaled to same input size, post processing erosion, dilation, filtering	73/27	DSC = 0.936 JSC = 0.881
(Dai et al., 2018)	X-ray	Whole lung	Healthy, Tuberculosis , lung nodules	385	2D	SCAN (structure correcting adversarial network)	Scaled to same input size	85/15	IoU = 94.7% ± 0.4% DSC = 0.973 ± 0.02
(Wang, 2017)	X-ray	Whole lung	Healthy, lung nodules	247	2D	Multi-task U-Net	Scaled to same input size, post processing hole filling	NR	JSC = 0.959 ± 0.017 AD = 1.29 ± 0.80mm
(Novikov et al., 2018)	X-ray	Whole lung	Healthy, lung nodules	247	2D	InvertedNet + All-dropout	Normalised [mean=0, SD = 0]	3-fold CV	DSC = 0.974 JSC = 0.949

(Hooda et al., 2018)	X-ray	Whole lung	Healthy, Tuberculosis, lung nodules	385	2D	FCN-8 + dropout	Scaled to same input size, random cropping	75/25	DSC = 0.959
(Mittal et al., 2018)	X-ray	Whole lung	Healthy, Tuberculosis, lung nodules	385	2D	LF-SegNet	Scaled to same input size, random cropping	48/52	DSC = 0.951
(Gaál et al., 2020)	X-ray	Whole lung	Healthy, Tuberculosis, lung nodules	1047	2D	Adversarial attention U-Net	Scaled to same input size, CLAHE, Normalisation [-1,1]	24/76	DSC = 0.962 ± 0.04
(Chen et al., 2019)	CT	Lung tumour	Lung cancer	134	3D	HSN (2D + 3D CNN)		78/22	DSC = 0.888 ± 0.033
(Jiang et al., 2018)	CT, MRI	Lung tumour	Lung cancer	400 CT (377) MRI (23)	2D	Tumour aware semi-supervised Cycle-GAN	Scaled to same input size, Image synthesis from CT to MRI, body mask	98/2	DSC = 0.63 ± 0.24 HD95 = 11.65 ± 6.53
(Jiang et al., 2019)	CT, MRI	Lung tumour	Lung cancer	405 CT (377) MRI (28)	2D	Tumour aware pseudo MR and T2w MR U-Net	Scaled to same input size, Image synthesis from CT to MR, Clipped [-1000,500 HU] and [0,667], Normalised [-1, 1]	95/5	DSC = 0.75 ± 0.12 HD95 = 9.36 ± 6.00mm VR = 0.19 ± 0.15

(Tahmasebi et al., 2018)	MRI	Lung tumour	Lung cancer	6	2D	Adapted FCN	Rescaled 10-95% of intensities, Normalisation [0,1]	5-fold CV	DSC = 0.91 ± 0.03 HD = 2.88 ± 0.86 mm RMSE = 1.20 ± 0.34
(Zhong et al., 2019b)	FDG PET, CT	Lung tumour	Lung cancer	60 PET (60) CT (60)	3D	DFCN Co-Seg U-Net	Scaled to same input size, Clipped [-500,200 HU] and [0.01,20]	80/20	DSC (CT) = 0.861 ± 0.037 DSC (PET) = 0.828 ± 0.087
(Zhao et al., 2018)	PET, CT	Lung tumour	Lung cancer	84 PET (84) CT (84)	3D	V-Net + feature fusion	Cropped to ROI	57/43	DSC = 0.85 ± 0.08 VE = 0.15 ± 0.14
(Zhou et al., 2019)	CT	Lung tumour	NR	1350	3D	P-SiBA	Transfer learning from ImageNet ILSVRC-2014, Cropped to ROI, Rescaled by +1000 HU and dividing by 3000 and Normalisation [0,1]	NR	DSC = 0.809 ± 0.12 HD = 7.612 ± 5.03mm VS = 0.883 ± 0.13
(Moriya et al., 2018)	Micro CT	Lung tumour	Lung cancer	3	3D	JULE CNN + k-means	Body mask, patch extraction		NMI = 0.390
(Imran et al., 2020)	CT	Lobes	COPD, ILD	563	3D	Progressive dense V-Net		48/52	DSC (n=84) = 0.939 ± 0.02

									DSC (n=154) = 0.950 ± 0.007 DSC (n=55) = 0.934
(Park et al., 2020)	CT	Lobes	COPD	196	3D	U-Net	Clipped [-1024,-400 HU]	80/20	DSC = 0.956 ± 0.022 JSC = 0.917 ± 0.031 Avg HD = 1.315 ± 0.563mm HSD = 27.89 ± 7.50
(Wang et al., 2018b)	CT	Lobes	COPD, IPF	1280	3D	DenseNet	Clipped -1000 to +1000 HU, Normalisation [0,1]	5-fold CV	DSC = 0.959 ± 0.087 ASD = 0.873 ± 0.61 mm
(Hatamizadeh et al., 2019)	CT	Lung lesion	NR	87	3D	DALS CNN	Scaled to same input size, Normalisation [NR]	90/10	DSC = 0.869 ± 0.113 HD = 2.095 ± 0.623mm
(Kalinovsky et al., 2017)	CT	Lung lesion	Tuberculosis	338	2D	GoogLeNet CNN	Images cropped into 4 quadrants	80/20	IoU = 0.95 ROC = 0.775
(Gerard et al., 2019)	CT	Lung fissure	COPD, Lung cancer	5327	3D	Two Seg3DNets	Clipped [-1024,-200 HU], Linear rescaling	30/70	Avg HD = 1.25mm SDSD = 2.87
(Sandkühler et al., 2019)	MRI	Lung defect region	NR	35	2D	GAE-LAE RNN with LCI Loss	Z-normalisation [-4,4], Lung mask, Normalisation [0,1], Histogram stretching	80/20	Qualitative evaluation - 42% images rated ‘very good’, 19% rated ‘perfect’

(Vakalopoulou et al., 2018)	CT	ILD pattern	ILD	46	2D	AtlasNet		37/63	DSC = 0.677 HD = 3.981mm Avg HD = 1.274mm
(Anthimopoulos et al., 2019)	CT	ILD pattern	ILD	172	2D	FCN-CNN	Pre-computed lung mask	5-fold CV	Accuracy = 81.8%
(Park et al., 2019)	CT	ILD pattern	COP, UIP, NSIP	647	2D	U-Net		88/12	DSC = 0.988 ± 0.006 JSC = 0.978 ± 0.011 Avg HD = 0.27 ± 0.18mm HSD = 25.47 ± 13.63mm
(Gao et al., 2016)	CT	ILD pattern	ILD	17	2D	CNN based CRF unary classifier	Transfer learning from ImageNet, Pre-computed lung mask		Accuracy = 92.8%
(Suzuki et al., 2020)	CT	Diffuse lung disease	NR	372	3D	U-Net		5-fold CV	DSC = 0.780 ± 0.169
(Wang et al., 2018a)	MRI	Foetal lung	NR	18	2D	BIFSeg P-Net	Trained on different organs, Image specific fine-tuning	66/33	DSC = 0.854 ± 0.059
(Rajchl et al., 2017)	MRI	Foetal lung	Healthy, IUGR	55	3D	DeepCut CNN + CRF	Bounding box for ROI, Bias correction, Normalisation	5-fold CV	DSC = 0.749 ± 0.067

							[mean=0], Transfer learning from LeNet		
(Edmunds et al., 2019)	Cone-beam CT	Diaphragm	Lung cancer	10	2D	Mask R-CNN	Scaled to same input size	9-fold CV	Mean error = 4.4mm
(Wang et al., 2019)	CT	Airways	NR	38	3D	Spatial-CNN (U-Net)	Random cropping	92/8 3-fold MCCV	DSC = 0.887 ± 0.012 CO = 0.766 ± 0.06
(Garcia-Uceda Juarez et al., 2019)	CT	Airways	Lung cancer	32	3D	U-Net GNN	Bounding box for ROI	63/37	DSC = 0.885 Airway completeness = 74%
(Yun et al., 2019)	CT	Airways	COPD	89	2D	2.5D CNN	Clipped [-700,700 HU]	78/22	Mean Branch detected = 65.7%
(Juarez et al., 2018)	CT	Airways	Healthy, CF, CVID	24	3D	U-Net	Bounding box for ROI	75/25	DSC = 0.8

Abbreviations: Chronic obstructive pulmonary disorder (COPD), Convolutional neural network (CNN), Idiopathic pulmonary fibrosis (IPF), Hounsfield unit (HU), Average Hausdorff distance (Avg HD), Dice similarity coefficient (DSC), Generative adversarial network (GAN), Not reported (NR), Hausdorff distance (HD), Jaccard similarity coefficient (JSC), Average boundary Hausdorff distance (Avg boundary HD), HD 95th Percentile (HD95), Percent ventilated defect (PVD), Receiver operating characteristic (ROC), Intersection over union (IoU), Average distance (AD), Relative volume ratio (VR), Root mean square error (RMSE), Region of interest (ROI), Classification error (CE), Volume error (VE), Normalised mutual information (NMI), Volumetric similarity (VS), Standard deviation of surface distances (SDSD), Mean average precision (MAP), Centreline overlap (CO), Cross-validation (CV), Leave-one-out cross-validation (LOOCV), Convolutional deep wide network (CDWN), Contrast limited adaptive histogram equalisation (CLAHE), Cystic fibrosis (CF), Interstitial lung disease (ILD), Hausdorff surface distance (HSD), Monte carlo cross-validation (MCCV), Usual interstitial pneumonia (UIP), Nonspecific interstitial pneumonia (NSIP), Intrauterine growth restriction (IUGR), Common variable immunodeficiency disorders (CVID), Standard deviation (SD), Fluorine-18-fluorodeoxyglucose (FDG). *The training dataset includes internal validation data.

CT is the most common modality for clinical lung imaging due to superior spatial resolution, rapid scan times and widespread availability. This is reflected in the DL lung segmentation literature with the majority of studies to date focusing on CT. For whole-lung segmentation, 3D networks are often used, whereas in interstitial lung disease (ILD) pattern segmentation, only 2D networks have been applied to date. The application often dictates the use of 2D and 3D networks; segmentation of the whole lung leads to a volumetric 3D region in which features such as overall lung shape, or the position of the trachea can be encoded. In contrast, segmenting ILD patterns is often conducted on central 2D slices; hence, a 2D network may be more appropriate as, in this approach, no features are conserved between slices (Anthimopoulos *et al.*, 2019; Park *et al.*, 2019). Across the CT papers reviewed, both the median and mode training/testing data splits were 80/20%, with many using k-fold cross-validation with less than 50 patients. Even as an independent testing set, using only 5-10 patients for testing limits generalisability. Moreover, some studies cite the number of images or 2D slices rather than the number of subjects. If data from the same subject is included in both the testing and training phases, it is likely that the algorithm has already seen a similar slice from the same patient as the individual data points are spatially correlated and do not strictly represent independent data points.

The Dice similarity coefficient (DSC) overlap metric is the most common evaluation metric used. Most studies tackling whole-lung segmentation report DSC values above 0.90, with some achieving values above 0.98. For other pulmonary ROIs, the highest DSC values reported are often lower (e.g. DSC (airways) ≈ 0.85). However, overlap metrics such as the DSC can be insensitive to errors in large volumes as the percent error is low compared to the overall pixel count (Taha and Hanbury, 2015). Frequently, high DSC values are reported despite errors that require significant manual intervention before a segmentation is clinically useful. As the airways occupy smaller volumes, the DSC metric is more sensitive. In terms of Hausdorff-based distance metrics, whole-lung segmentation studies report HD95 values ≈10mm; however, (Dong *et al.*, 2019) report a HD95 as low as 2.249±1.082mm averaged across both lungs. The lack of a standardised evaluation metric can make direct comparisons between different methods challenging.

Image segmentation is challenging to evaluate. Currently, manual segmentations by expert observers are used as the gold standard; however, it is well-known that expert segmentations are susceptible to inter-observer variability (Mukesh *et al.*, 2012). Often, only one observer segments all the images in a training dataset; hence, if a different observer segments the testing images, the algorithm may not perform as expected. This poses problems for widespread generalisation if certain biases in segmentation are preserved as there is no clear 'true' expert segmentation; therefore, differences in DL segmentations and expert segmentations may not be solely the result of DL errors. Most expert segmentations are conducted using semi-automatic software and image editing tools; the tools given to the user can convey a propensity for features, such as smooth lung borders, which may, in fact, be inaccurate. In other anatomical sites such as the liver, a DSC of 0.95 was obtained by DL; the inter-observer variability for the DL approach was 0.69% compared to 2.75% for manual expert observers (Chlebus *et al.*, 2019). The low degree of inter-observer variability in DL segmentations may be a positive step towards consistent segmentations between institutions. Using multiple expert segmentations and averaging the error may reduce inter-observer variability effects; however, this is unlikely to be widely adopted due to the time required. In addition, medical imaging grand challenges can provide diverse data from multiple institutions with corresponding expert segmentations, limiting the extent of individual researcher bias.

There are limited studies to date regarding pulmonary MRI segmentation, attributable perhaps to less widespread clinical use of the modality and lack of large-scale annotated pulmonary MRI datasets. However, pulmonary MRI techniques, such as contrast-enhanced lung perfusion MRI and hyperpolarised gas ventilation MRI, can provide further insights into pulmonary pathologies currently not possible with alternative techniques (Woodhouse *et al.*, 2005). Quantitative biomarkers derived from hyperpolarised gas MRI, including the ventilated defect percentage, require accurate segmentation of ventilated and whole-lung volumes which can be very time consuming when performed manually. Example images of DL-based hyperpolarised gas MRI segmentations are provided in Figure 3.7. (Tustison *et al.*, 2019) used CNNs to provide fast, accurate segmentations for hyperpolarised gas and proton MRI (Tustison *et al.*, 2019). A 2D U-Net was used for hyperpolarised gas MRI segmentation whilst a 3D U-Net was used for proton MRI segmentation. They introduced a novel template-based data augmentation method to expand the limited lung imaging data. Hyperpolarised gas and proton MR images were segmented with DSC values of 0.94±0.03 and 0.94±0.02, respectively. Research evaluated a DL-based proton MRI segmentation network, which yielded an average DSC of 0.965 across both lungs, outperforming conventional region growing and k-means techniques (Zha *et al.*, 2019).

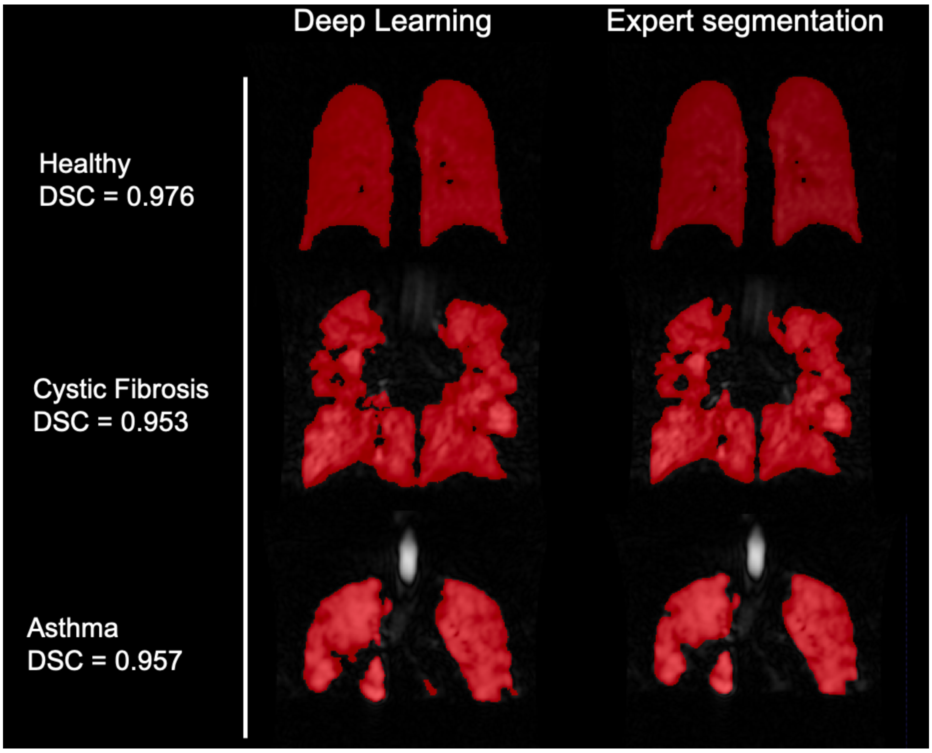


Figure 3.7 Example images from the authors' own work using deep learning for hyperpolarised gas MRI segmentation. The ^{129}Xe MR ventilation images are taken from three subjects in a testing set, a healthy volunteer, asthma patient and cystic fibrosis patient. The patient images selected are characterised by significant ventilation defects. These are compared to expert segmentations of the same image. Dice similarity coefficient (DSC) values are displayed for all images.

Although the majority of segmentation studies reviewed used CT and MRI, early studies focused on X-ray segmentation (Wang, 2017; Hwang and Park, 2017). This was due to the public availability of large-scale, annotated X-ray datasets, such as the Japanese Society of Radiological Technology (JSRT) (Shiraishi *et al.*, 2000) and Montgomery (Jaeger *et al.*, 2014) datasets, enabling researchers to experiment with large numbers of images not previously accessible. The majority of X-ray studies reviewed used these datasets, making comparisons between methods more applicable (Hooda *et al.*, 2018; Mittal *et al.*, 2018; Novikov *et al.*, 2018; Souza *et al.*, 2019; Wang, 2017; Dai *et al.*, 2018).

3.6.3 Registration

Image registration is the process of transforming a moving image onto the spatial domain of a fixed image. Registration is used in numerous applications within the lung imaging field, including adaptive radiotherapy (Moro *et al.*, 2013), computation of functional lung metrics such as the VDP (Hughes *et al.*, 2018) and generation of surrogates of regional lung function from multi-inflation CT (Tahir *et al.*, 2018) or ^{1}H MRI (Bauman *et al.*, 2009). However, most image registration algorithms assume that the moving and fixed images' topology are the same. This is not always the case in lung imaging as often functional images do not follow the same topology as structural images, especially in individuals with severe pathologies where functional lung images may show substantial heterogeneity (Tahir *et al.*, 2016). Studies concerning DL-based pulmonary registration are summarised in Table 3.4

(Eppenhof and Pluim, 2019) built upon previous work (Lafarge *et al.*, 2018) using publicly available datasets to directly map displacement vector fields from inspiratory and expiratory CT pairs using a 3D U-Net with extensive data augmentation. Synthetic transforms were used to directly train the network as the deformation fields are known. The approach achieved fast, accurate registrations, reducing mean TRE from 8.46mm to 2.17mm. The results are further validated using landmarks from multiple observers, indicating the level of inter-observer variability. Notwithstanding, only 24 images for testing and training were used, limiting the study's generalisability. In addition, synthetic transforms do not directly represent real transforms likely found in patients. Other approaches use a CNN to learn expressive local binary descriptors from landmarks before applying Markov random field registration (Blendowski and Heinrich, 2019). This is compared to a method using handcrafted local descriptors with high self-similarity, facilitating faster computation. The results suggest that a combination of both CNN-learned descriptors and handcrafted features produce the best registration results.

In a generic registration approach, a U-Net-like architecture with a differentiable spatial transformer that can register both X-ray and MR images was used (Ferrante *et al.*, 2018). The algorithm was evaluated using the contour mean distance (CMD). CMD was approximately 5mm on average across the testing data. Whilst this is a less accurate registration than other methods reviewed, it is more broadly applicable; the generic algorithm (in this case trained on X-ray and MR images) can learn features that are independent of modality. By fixing these weights and adding additional layers, transfer learning can then be applied to a specific modality; the additional data across modalities may lead to improved results (Tajbakhsh *et al.*, 2016).

Table 3.4 Summary of reviewed studies using deep learning for lung image registration.

Study	*Modality*	*Disease*	*Public dataset*	*Number of subjects*	*Dimensions*	*Architecture*	*Pre-processing*	*Percentage data split (training*/testing)*	*Performance*
(Lafarge et al., 2018)	4DCT	Lung cancer	DIR-LAB, CREATIS	17	3D	Modified VGG	Synthetic DVFs for data augmentation	42 (CREATIS) / 58 (DIR-LAB)	TRE = 4.02 ± 3.08
(Eppenhof and Pluim, 2019)	4DCT	Lung cancer	DIR-LAB, CREATIS	17	3D	Modified U-Net	Synthetic DVFs for data augmentation, Resised, Pre-computed body mask, mask < -250 HU	42 (CREATIS) / 58 (DIR-LAB)	TRE = 2.17 ± 1.89mm
(Ali and Rittscher, 2019)	4DCT	Lung cancer	DIR-LAB, CREATIS	17	2D	Conv2Wrap (Linear and Deformable ConvNet)		58 (DIR-LAB) / 42 (CREATIS)	DSC = 0.90 JSC = 0.84
(Sentker et al., 2018)	4DCT	Lung cancer	DIR-LAB, CREATIS	86	3D	GDL-FIRE^{4D} U-Net with VarReg	Normalisation [0,1], Cropped to same input size, pre-computed body mask	69 / 31 (DIR-LAB, CREATIS, In house)	TRE (DIR-LAB) = 2.50 ± 1.16mm TRE (CREATIS) = 1.74 ± 0.57mm
(Fechter and Baltas, 2020)	4DCT	Lung cancer	DIR-LAB, CREATIS, Sunnybrook	31	3D	U-Net one-shot learning	Pre-computed body mask, Normalisation [mean=0, SD=1]	LOOCV (DIR-LAB) 0 / 100 (CREATIS)	TRE (DIR-LAB) = 1.83 ± 2.35mm TRE (CREATIS) = 1.49 ± 1.59mm

(Fu et al., 2020)	4DCT	Lung cancer	DIR-LAB	20	3D	LungRegNet (CourseNet, FineNet)	Vessel enhancement, Clipped at -700 HU	5-fold CV, DIR-LAB testing	MAE (in house) = 52.1 ± 18.4 TRE (in house) = 1.00 ± 0.53 TRE (DIR-LAB) = 1.59 ± 1.58mm
(Jiang et al., 2020)	4DCT	Lung cancer	DIR-LAB, SPARE	32	3D	MJ-CNN	Clipped [-1000,-200 HU], Normalisation [0,0.2]	75 (SPARE, DIR-LAB) / 25 (DIR-LAB)	TRE = 1.58 ± 1.19mm
(de Vos et al., 2019)	4DCT, CT	Lung cancer	DIR-LAB, NLST	2070	3D	DLIR framework ConvNet	Clipped [-1000,-200 HU], Normalisation [0,1]	99 (NLST) / 1 (NLST, DIR-LAB)	DSC (NLST) = 0.75 ± 0.08 HD (NLST) = 19.34 ± 13.41 TRE (DIR-LAB) = 5.12 ± 4.64mm
(Sokooti et al., 2017)	CT	COPD		19	3D	RegNet CNN	Synthetic DVFs for data augmentation, Initial affine registration	63/37 (SPREAD)	TRE = 4.39 ± 7.54mm
(Sokooti et al., 2019)	CT, 4DCT	Lung cancer, COPD	SPREAD, DIR-LAB	39	3D	RegNet CNN, U-Net	Synthetic DVFs for data augmentation, Initial affine registration	54 (SPREAD, DIR-LAB COPD) / 46 (SPREAD, DIR-LAB)	TRE (DIR-LAB) = 1.86 ± 2.12mm
(Blendowski and	CT	COPD	DIR-LAB	10	3D	CNN	Cropped to lung region	LOOCV (DIR-LAB)	TRE = 3.00 ± 0.48mm

Heinrich, 2019)									
(Qin et al., 2019)	CT, MRI	COPD	COPDGene	1000	2D	UMDIR-LaGAN	Cross modality registration, deformed to domain invariant latent space	90/10 (COPDGene)	DSC = 0.967 ± 0.03 HD = 8.257 ± 4.43mm MCD = 0.71 ± 0.44mm
(Galib et al., 2020)	CT, CBCT	Healthy, COPD, Lung cancer	DIR-LAB, VCU	27	3D	CNN	Normalisation [0,1]	37 (DIR-LAB) / 63(VCU)	AUC-ROC = 0.882 ± 0.11 CI = 68%
(Ferrante et al., 2018)	X-ray	Healthy, Lung nodule	JSRT	247	2D	U-Net	Normalisation [0-1], Domain adaption Cardiac MR	81/19 (JSRT)	MAD ≈ 6.3 CMD ≈ 5 mm DSC ≈ 0.9
(Mahapatra et al., 2018)	X-ray	Multiple	NIH-ChestXray14	420	2D	JRSNet (cycleGAN with U-Net)	Joint segmentation and registration	NR (SCR, NIH-ChestXray14)	TRE = 7.75mm
(Stergios et al., 2018)	MRI	Systemic sclerosis, healthy		41	3D	CNN + transformation layer	Clipped [0, 1300], Normalisation [0,1]	68/32	DSC = 0. 915 ± 2.33 Euclydian error = 4.358mm

Abbreviations: Target registration error (TRE), Dice similarity coefficient (DSC), Jaccard similarity coefficient (JSC), Mean absolute error (MAE), Area under curve-receiver operator characteristic (AUC-ROC), Mean absolute differences (MAD), Contour mean distance (CMD), Visual geometry group (VGG), Markovian random field (MRF), Cross-validation (CV), Leave-one-out cross-validation (LOOCV), Hausdorff distance (HD), Mean contour distance (MCD), Chronic obstructive pulmonary disorder (COPD), Convolutional neural network (CNN), Deep learning image registration (DLIR), Hounsfield unit (HU). *The training dataset includes internal validation data.

3.6.4 Reconstruction

Image reconstruction is the process of generating a usable image from the raw data acquired by a scanner. CT and SPECT reconstruction use different reconstruction algorithms to MRI. DL-based CT reconstruction, as with segmentation, is further developed than DL-based MRI reconstruction applications. CT and SPECT reconstruction use analytic (e.g. filtered back-projection) or iterative algorithms to produce 3D images from projections taken at multiple angles around a subject. MRI reconstruction, in contrast, produces images by transforming raw k-space data via Fourier transforms. Full details of image reconstruction methods have been described elsewhere (Willemink and Noel, 2019; Ye, 2019) Studies describing DL-based lung image reconstruction applications are summarised in Table 3.5.

CT/SPECT images can be reconstructed accurately using Monte-Carlo based iterative reconstruction (Norberg *et al.*, 2007); however, this process is computationally expensive and time-consuming (El Bitar *et al.*, 2006). In addition, multiple studies have demonstrated the success of analytical methods such as filtered-back projection (Willemink and Noel, 2019). Building upon this, CNNs have been used to speed up the process of filtered back-projection to shorten reconstruction times (Dietze *et al.*, 2019). The results suggest DL can accurately reconstruct SPECT images in under 10 seconds. Furthermore, the authors compare clinical metrics, such as the lung shunting fraction (LSF), between methods in a specific time frame. DL produced an LSF of 4.7% comparable to 5.8% for Monte-Carlo methods, indicating the potential for use in clinical applications (Dietze *et al.*, 2019). Multiple studies have employed DL for MRI reconstruction (Hammernik *et al.*, 2018) but only one published study has applied it to pulmonary MRI (Duan *et al.*, 2019). MRI of the lungs can take upwards of 10 seconds to acquire, often requiring that patients maintain inflation levels for a significant period; this can be particularly challenging for patients with severe lung pathologies. Compressed sensing can be used to reconstruct randomly undersampled k-space in conjunction with regularisation methods to produce accurate reconstructions in hyperpolarised gas MRI (Ajraoui *et al.*, 2010; Sheikh *et al.*, 2016) and enables reduced acquisition time without significantly reducing image quality. A coarse-to-fine neural network has been proposed to yield an accurate hyperpolarised gas MRI scan with an accelerating factor of 8 (undersampled 1/8 of k-space) (Duan *et al.*, 2019). The method can also improve inherent spatial co-registration accuracy when acquiring proton and hyperpolarised gas MRI in the same breath (Wild *et al.*, 2011), possibly alleviating the need for substantial post-acquisition image registration.

Tangentially related to the goal of image reconstruction, images can also be improved further using image enhancement at the post-acquisition stage. Multiple studies have shown the effectiveness of using CNNs combined with gradient regularisation and super-resolution modules to enhance low-dose CT images with noise and artefacts, potentially limiting radiation exposure without degrading image quality (Gou *et al.*, 2019; Umehara *et al.*, 2018).

3.6.5 Synthesis

Image synthesis, also referred to as regression, is the process of generating artificial images of unknown target images from given source images. Synthesis has been applied to a range of applications, such as generating functional or metabolic images from structural images. For example, estimating contrast-based functional images from routinely acquired non-contrast structural modalities reduces the need for additional scans, specialised equipment and administration of contrast agents. Even within traditional model-based techniques, accurate synthesis has proved challenging due to the complex mathematical functions mapping input to output images. Studies describing DL-based lung image synthesis applications are summarised in Table 3.6.

The development of DL architectures such as GANs enables a more unsupervised approach, which lends itself to the complex problem of synthesis (Kazeminia *et al.*, 2018). DL has been used to generate synthetic fluorine-18-fluorodeoxyglucose (FDG) PET images from CT images via a GAN (Bi *et al.*, 2017). The GAN's inputs were varied to include either a CT image, label, or both CT and corresponding label; the multi-channelled GANs (M-GAN) provided the most accurate synthetic PET images, demonstrating that multiple inputs increase synthesis accuracy. To explore this further, the authors also evaluate the synthetic PET images by feeding them into a network as training data. The network aims to delineate tumours by learning relationships from the training data; the data was then divided into real PET images and synthetic PET images. The trained model was then evaluated on unseen tumour detection problems. The synthetic PET-trained network produced 2.79% lower recall accuracy. This indicates that, as a whole, the synthetic PET images are closely related to the real images in terms of tumour identification. The paper posits that synthetic PET images can be used as additional training data in other DL tasks. However, it is unclear if synthetic PET images can be used in treatment planning and other clinical tasks with this level of accuracy (Bi *et al.*, 2017). GANs have continued to show promise in synthesis problems (Jang *et al.*, 2019). CT images have been used to generate SPECT images via a conditional GAN (cGAN) instead of a CNN (Ren *et al.*, 2019). The method used a 2D GAN with 49 patients consisting of 3054 2D images as training data; the testing data contains five patients. cGANs differ from the regular GAN architecture by using both the observed image and a random noise vector, mapping these to the output image instead of only the noise vector. The generator used is based on the U-Net architecture with multiple inputs. Synthetic and real SPECT images were compared using the multiscale structural similarity index measure (MS-SSIM), yielding MS-SSIM=0.87. Further analysis used a Gamma index with a passing rate of 97.7±1.2% with 2%/2mm. The authors note qualitatively that errors occur more frequently at the base of the lungs, possibly caused by the increased deformation in this region. A key limitation for synthesis methods is the errors introduced by the registration of source and target images. Consequently, it has been suggested that images that are not matched anatomically due to breathing discrepancies are excluded (Jang *et al.*, 2019). complicating validation for clinical adoption (Jang *et al.*, 2019; Ren *et al.*, 2019).

A major application of DL image synthesis is for MR-guided radiotherapy. The current paradigm in radiotherapy is to derive electron density information required for dose calculations directly from CT scans; MRI does not directly provide this information. DL has been invoked to generate pseudo-CT images for use in MR-guided stereotactic body radiotherapy using GANs, precluding the need for CT (Olberg *et al.*, 2018). (Zhong *et al.*, 2019a) used a CNN to synthesise ventilation images from 4DCT scans. Whilst good performance was observed, the major limitation of this study is that the target images in the training phase were CT-based surrogates of ventilation generated from aligned inspiratory and expiratory CT scans via deformable registration and computational modelling. These images are still the subject of intense validation efforts (Kipritidis *et al.*, 2019). Using more direct measures of regional lung function, such as hyperpolarised gas MRI, and larger datasets are critical to the success of future work in structure-to-function DL synthesis applications.

Table 3.5 Summary of reviewed studies using deep learning for lung image reconstruction.

Study	*Modality*	*Disease*	*Number of patients*	*Dimensions*	*Architecture*	*Pre-processing*	*Percentage data split (training*/testing)*	*Performance*
(Beaudry et al., 2019)	4D Cone beam CT	Lung cancer	16	2D	Sino-Net (Modified U-Net)	Cropped to same input size, Sinogram Normalisation [0,1]	88/12	RMSE Translational = 1.67mm
(Lee et al., 2019)	CT	COPD	60	2D	FCN	No sinogram used	Dataset 1: 80/20 Dataset 2: 40/60	Mean RMSE (Dataset 1) = 65.7 ± 15.8% Mean RMSE (Dataset 2) = 59.6 ± 5.5%
(Ge et al., 2020)	CT	Liver lesion	5413	2D	ADAPTIVE-NET CNN	Convert from HU to linear attenuation coefficient	90/10	PSNR = 43.15 ± 1.9 SSIM = 0.968 ± 0.013 Normalised RMSE = 0.0071 ± 0.002
(Duan et al., 2019)	HP Gas MRI	COPD, nodule, PTB, healthy, asthma	72	2D	C-Net and F-Net (U-Net based)	Under sampled K-space (AF =4), Removed SNR below 6.6, Normalisation [0,1]	NR	MAE = 4.35% SSIM = 0.7558 VDP bias = 0.01 ± 0.91%
(Dietze et al., 2019)	^{99}mTc-MAA SPECT	Liver Cancer	128	2D	CNN	Initial filtered back projection	94/6	LSF = 5.1% CNR = 12.5

Abbreviations: Root mean square error (RMSE), Structural similarity index metric (SSIM), Mean absolute error (MAE), Volume defect percentage (VDP), Lung shunting fraction (LSF), Contrast to noise ratio (CNR), Electrical impedance tomography (EIT), Peak signal to noise ratio (PSNR), Ventilation defect percentage (VDP), Convolutional neural network (CNN|), Chronic obstructive pulmonary disorder (COPD), Pulmonary tuberculosis (PTB), Hounsfield unit (HU), Technetium-99m macroaggregated albumin (^{99}mTc-MAA). *The training dataset includes internal validation data.

Table 3.6 Summary of reviewed studies using deep learning for lung image synthesis.

Study	*Modality (original ⇒ target)*	*Disease*	*Number of subjects*	*Dimensions*	*Model*	*Pre-processing*	*Percentage data split (training*/testing)*	*Performance*
(Bi et al., 2017)	CT ⇒ FDG PET	Lung cancer	50	2D	Multichannel-GAN (U-Net)	Manual segmentation of tumour/lymph nodes, axial slices with tumours only	50/50	MAE = 4.6 PSNR = 28.06
(Jang et al., 2019)	CT ⇒ ^{99}mTc-MAA SPECT perfusion	Lung cancer	54	2D	Conditional GAN	Resized images, segmentation and removal of bone, soft tissue and heart	91/9	MS-SSIM = 0.87 Gamma index 2%/2mm = 97.7% ± 1.2%
(Zhong et al., 2019a)	4DCT ⇒ CT ventilation	Lung cancer, COPD	82	2D	Deep CNN	Images cropped to ROI	10-fold CV	MSE = 7.6% Gamma index 5%/5mm = 80.6% ± 1.4% SSIM = 0.880 ± 0.035
(Liu et al., 2020)	4DCT ⇒ ^{99}mTc-Technegas SPECT ventilation	Lung cancer, oesophageal cancer	50	2D	U-Net	Pre-computed lung mask, normalisation [0,1], post-processing normalisation [90th percentile]	10-fold CV	Spearman's ρ = 0.73 ± 0.17 DSC = 0.73 ± 0.09
(Ren et al., 2019)	CT ⇒ ^{99}mTc-MAA SPECT perfusion	Lung cancer	30	3D	U-Net	Clipped [-1000,-300 HU] for segmentation, normalisation [0,1]	83/17	Correlation coefficient = 0.53 ± 0.14

(Preiswerk et al., 2018)	Ultrasound ⇒ MRI	NR	7	3D	LRCN	PCA = 10 components	66/33 (conducted in time segments)	SSE = 39.0 ± 12
(Olberg et al., 2018)	MRI ⇒ CT	NR	41	NR	GAN (U-Net)	Normalisation [NR], pre-computed body mask	90/10	3D Gamma index passing rate 99.2% Lung V20% difference = 0.11%
(Ren et al., 2022)	CT ⇒ SPECT	Lung cancer	170	3D	UNet	Clipped [-1000, -300 HU], transfer learning	3-fold CV	Voxel-wise Spearman’s correlation = 0.64
(Capaldi et al., 2020)	MRI ⇒ HP gas MRI	Various	114	2D	UNet	Normalisation [0, 255]	6-fold CV	Pearson correlation = 0.87

Abbreviations: Fluorine-18-fluorodeoxyglucose (FDG), Mean absolute error (MAE), Peak signal to noise ratio (PSNR), Technetium-99m macroaggregated albumin (^{99}mTc-MAA), Multi-scale structural similarity index metric (MS-SSIM), Mean square error (MSE), Sum of squared error (SSE), Generative adversarial network (GAN), Convolutional neural network (CNN), Not reported (NR), Principle component analysis (PCA), Hounsfield unit (HU), Region of interest (ROI), Long-term recurrent convolutional network (LRCN), Chronic obstructive pulmonary disease (COPD), Hyperpolarised (HP). *The training dataset includes internal validation data.

3.7 Future research directions

The studies reviewed show that DL has significant potential to outperform more traditional methods in a wide range of lung image analysis applications. Novel ways of using DL to synthesise more training examples (Salehinejad *et al.*, 2019) or combine segmentation and registration in one process (Mahapatra *et al.*, 2018) have been shown to enhance performance. The scope of such innovation is still in its infancy, providing an opportunity for novel technical developments. As shown through the improved performance observed by combining traditional approaches with machine learning and DL for registration, great synergy can be achieved by combining DL and conventional image processing approaches (Blendowski and Heinrich, 2019).

In image synthesis, researchers have developed techniques to synthesise CT images from MRI scans of the brain (Nie *et al.*, 2016); similar advancements in lung imaging would allow patients to receive less radiation exposure as well as reduce the cost and time for additional scans. Using synthesis to generate functional lung images from routinely acquired structural images would allow clinicians to understand which areas of the lungs are ventilated or perfused without the need to acquire dedicated functional scans, which often require contrast agents and specialised equipment, reducing costs and acquisition times. Such applications require further DL research in architectural development and the input of lung imaging experts. Using DL for CT enhancement to reduce radiation dose or improve compressed sensing methods in MRI has the potential to reduce scan times, improving image quality and patient compliance.

However, in relation to image synthesis, there are important considerations which must be addressed before widespread clinical translation. Numerous challenges exist, including the ability to reliably generate synthetic surrogates, which limits its usefulness as a diagnostic tool in image analysis applications. Many research studies report inconsistent results on datasets of limited size, precluding their translation. In addition, this unreliability in conjunction with the lack of explainability reduces trust in synthetic images. This is particularly a concern in applications where information not present in the input images must be generated by the neural network as this can lead to the addition of incorrect information.

Thus, it is likely that DL-based synthesis solutions will mainly be used in a triaging capacity rather than as a specific diagnostic tool.

Promising results have been shown for both proton MRI and hyperpolarised gas MRI segmentation (Tustison *et al.*, 2019); however, further work is required to demonstrate accurate MRI segmentation in an independent multicentre validation. The importance of collaborative research to boost training data and inject heterogeneity of centre and scanner will lead to more robust and generalisable models. The paucity of published DL studies in functional lung imaging (only 12.9% of reviewed studies here) provides significant opportunities for innovations and further research in this field.

The literature on CT segmentation provides a positive picture of the success of DL methods in providing fast, accurate automatic segmentations. However, producing impressive results in a research setting is no substitute for clinical validation. Long-term clinical case studies are required with large numbers of patients before these novel developments have a real impact. The 'black box' nature of DL methods and the lack of explainability of generated outputs can undermine clinicians and patients' trust, despite, or even because of, an unprecedented level of hype. Another challenge is transparency; although most software used for DL is well documented and open source, a requirement for continued use, the open-source nature also generates safety concerns relating to software edits and bugs. Developing a standardised literature consensus on validation and evaluation procedures is key to ensuring transparency. All of these challenges need to be overcome before DL can live up to its full potential.

3.8 Conclusion

We have reviewed the role of DL for several lung image analysis tasks, including segmentation, registration, reconstruction and synthesis. CT-based lung segmentation was the most prevalent application where exceptional performance has been demonstrated. However, research in other applications and modalities, including functional lung imaging, is still in its infancy. A concerted effort from the research community is required to develop the field further. Before widespread clinical adoption is achievable, challenges remain concerning validation strategies, transparency and trust.

Chapter 4
Large-scale investigation of deep learning approaches for ventilated lung segmentation using multi-nuclear hyperpolarised gas MRI

Respiratory diseases are leading causes of mortality and morbidity worldwide. Pulmonary imaging is an essential component of the diagnosis, treatment planning, monitoring, and treatment assessment of respiratory diseases. Insights into numerous pulmonary pathologies can be gleaned from functional lung MRI techniques. These include hyperpolarised gas ventilation MRI, which enables visualization and quantification of regional lung ventilation with high spatial resolution. Segmentation of the ventilated lung is required to calculate clinically relevant biomarkers. Recent research in deep learning (DL) has shown promising results for numerous segmentation problems. Here, we evaluate several 3D convolutional neural networks to segment ventilated lung regions on hyperpolarised gas MRI scans. The dataset consists of 759 helium-3 (^{3}He) or xenon-129 (^{129}Xe) volumetric scans and corresponding expert segmentations from 341 healthy subjects and patients with a wide range of pathologies. We evaluated segmentation performance for several DL experimental methods via overlap, distance and error metrics and compared them to conventional segmentation methods, namely, spatial fuzzy c-means (SFCM) and K-means clustering. We observed that training on combined ^{3}He and ^{129}Xe MRI scans using a 3D nn-UNet outperformed other DL methods, achieving a mean±SD Dice coefficient of 0.963±0.018, average boundary Hausdorff distance of 1.505±0.969mm, Hausdorff 95th percentile of 5.754±6.621mm and relative error of 0.075±0.039. Moreover, limited differences in performance were observed between ^{129}Xe and ^{3}He scans in the testing set. Combined training on ^{129}Xe and ^{3}He yielded statistically significant improvements over the conventional methods ($p<0.0001$). In addition, we observed very strong correlation and agreement between DL and expert segmentations, with Pearson correlation of 0.99 ($p<0.0001$) and Bland-Altman bias of -0.8%. The DL approach evaluated provides accurate, robust and rapid segmentations of ventilated lung regions and successfully excludes non-lung regions such as the airways and artefacts. This approach is expected to eliminate the need for, or significantly reduce, subsequent time-consuming manual editing.

4.1 Preface

The majority of the material in this chapter was originally published as a full-length article in the journal *Nature Scientific Reports*:

> **Astley J.R.**, Biancardi A.M., Hughes P.J.C., Marshall H., Smith L.J., Collier G.J., Eaden J.A., Weatherley N.D., Hatton M.Q., Wild J.M. and Tahir B.A. (2022). Large-scale investigation of deep learning approaches for ventilated lung segmentation using multi-nuclear hyperpolarized gas MRI. *Scientific Reports* 12, 10566. https://doi.org/10.1038/s41598-022-14672-2.

This article was published under an Open Access Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format (https://creativecommons.org/licenses/by/4.0/). Slight modifications have been made to the published version. The work contained within this chapter has also been published as conference proceedings at the following conferences:

> **Astley J.R.**, Biancardi A.M., Hughes P.J.C., Smith L.J., Marshall H., Eaden J.A., Weatherley N.D., Bray J., Collier G.J., Wild J.M. and Tahir B.A. (2020). 3D Deep Convolutional Neural Network-Based Ventilated Lung Segmentation Using Multi-Nuclear Hyperpolarized Gas MRI. MICCAI: In: Petersen J. et al. (eds) Thoracic Image Analysis. TIA 2020. Lecture Notes in Computer Science, vol 12502. Springer, Cham. https://doi.org/10.1007/978-3-030-62469-9_3
>
> **Astley J.R.**, Biancardi A.M., Hughes P.J.C., Smith L.J., Marshall H., Mussell G.T., Eaden J.A., Weatherley N., Collier G.J., Wild J.M. and Tahir B.A. (2020). Automatic Segmentation of Hyperpolarized Gas MRI via Deep Learning. Institute of pure and applied mathematics: deep learning and medical applications (IPAM-DLM) 2020. *Los Angeles, CA, USA.*
>
> **Astley J.R**., Biancardi A.M., Hughes P.J.C., Smith L.J., Marshall H., Mussell G.T., Eaden J.A., Weatherley N., Collier G.J., Wild J.M. and Tahir B.A. (2020). Automatic Segmentation of Hyperpolarized Gas MRI via Deep Learning. The international society for magnetic resonance in medicine (ISMRM) 2020. *Online.*
>
> **Astley J.R.**, Biancardi A.M., Hughes P.J.C., Smith L.J., Marshall H., Eaden J.A., Weatherley N.D., Collier G.J., Wild J.M. and Tahir B.A. (2021). Automatic Segmentation of Hyperpolarised Gas MRI via Deep Learning. The University of Sheffield Medical School Research Day (MSR) 2021. *Sheffield, UK.*

Astley J.R., Biancardi A.M., Hughes P.J.C., Smith L.J., Marshall H., Collier G.J., Eaden J.A., Weatherley N.D., Wild J.M. and Tahir B.A. (2021). Comparison of 3D deep convolutional neural networks and loss functions for ventilated lung segmentation using multi-nuclear hyperpolarized gas MRI. The international society for magnetic resonance in medicine (ISMRM) 2021. *Online.*

Astley J.R., Biancardi A.M., Hughes P.J.C., Marshall H., Eaden J.A., Weatherley N., Collier G.J., Wild J.M. and Tahir B.A. (2021). Comparison of 3D deep convolutional neural networks and training strategies for ventilated lung segmentation using multi-nuclear hyperpolarized gas MRI. American association of physicists in medicine (AAPM) 2021. *Online.*

Additional material that could not be included within the journal article or within conference proceedings is also contained within this chapter.

4.1.1 Author contributions

J.R.A., J.M.W. and B.A.T. made substantial contributions to the conceptualisation of the work. A.M.B., P.J.C.H., H.M. L.J.S., G.J.C., J.A.E., N.D.W., M.Q.H., J.M.W. and B.A.T. were involved with patient recruitment, image acquisition and/or generating expert segmentations. J.R.A. performed the deep learning experiments, interpreted data, and conducted statistical analyses. J.R.A. drafted the manuscript. B.A.T. substantively revised the manuscript. All authors reviewed and approved the submitted manuscript.

4.2 Introduction

Respiratory diseases are leading causes of mortality and morbidity worldwide with 339 million experiencing asthma, 65 million people with chronic obstructive pulmonary disease (COPD) (Vos *et al.*, 2017; GBD15 *et al.*, 2016) and 1.8 million new lung cancer cases diagnosed every year (Torre *et al.*, 2015). Pulmonary imaging, using various modalities, is an essential part of the diagnosis, treatment planning, monitoring, and treatment assessment of respiratory diseases. The acquisition, processing, and interpretation of pulmonary images are critical components of patient management and are essential in reducing mortality and morbidity.

Currently, computed tomography (CT) is the clinical gold standard for pulmonary imaging due to its exceptional spatial and temporal resolution, and its ubiquitous availability. CT is a structural imaging modality that provides exquisite detail of morphological changes in the lung parenchyma but employs ionising radiation. Although proton magnetic resonance imaging (^{1}H MRI) has historically been susceptible to the low proton density in lungs, recent advances in pulse sequences and hardware with ultra-short and zero echo times have enabled ^{1}H MRI to compete with CT with the added benefit of no ionising radiation (Togao *et al.*, 2010; Bae *et al.*, 2019). However, whilst structural imaging modalities facilitate the assessment of changes in lung tissue density, they do not directly provide an accurate picture of regional lung function.

Although nuclear imaging modalities such as single-photon emission computed tomography (SPECT) can provide regional lung function information (Petersson *et al.*, 2007), they require harmful ionising radiation, reducing the ability to conduct regular scans during clinical care. This is particularly important when imaging children, as developing tissue is more sensitive to ionising radiation. Moreover, SPECT is limited by poor temporal and spatial resolution and images acquired using ^{99m}Tc-diethylenetriamine pentaacetate (DTPA) aerosols, one of the most commonly used radiotracers for ventilation imaging with SPECT, are subject to clumping artefacts (Petersson *et al.*, 2007; Yuan *et al.*, 2011). In contrast, unparalleled insights into respiratory diseases can be gleaned from non-ionising functional lung MRI modalities, such as dynamic contrast-enhanced lung perfusion MRI and hyperpolarised gas ventilation MRI. Hyperpolarised gas MRI provides visualisation and quantification of regional lung ventilation with high spatial resolution within a single breath (Fain *et al.*, 2007). Quantitative biomarkers derived from this modality, including the ventilated defect percentage (VDP) and coefficient of variation, provide further insights into regional ventilation (Woodhouse *et al.*, 2005; Tzeng *et al.*, 2009; Hughes *et al.*, 2019). To facilitate the computation of such biomarkers, segmentation of ventilated regions of the lungs is required (Tustison *et al.*, 2011).

Previous approaches for hyperpolarised gas MRI ventilation segmentation employed classical image processing and machine learning approaches, such as hierarchical K-means (Kirby *et al.*, 2012a) and spatial fuzzy c-means (SFCM) clustering (Hughes *et al.*, 2018). However, as these methods rely on voxel intensities and thresholding, they only provide semi-automatic segmentations; as such, they are prone to generate errors in regions where voxel intensities are similar to those of the ventilated lung region (e.g., airways and artefacts). Consequently, they frequently require significant time to manually correct.

Deep learning (DL), which utilises artificial neural networks with multiple hidden layers, has shown tremendous promise in medical image segmentation applications (Hesamian *et al.*, 2019). Although DL was initially theorised over half a century ago, the field only received widespread acclaim in 2012 when AlexNet, a form of an artificial neural network referred to as a convolutional neural network (CNN), triumphed in the ImageNet Large Scale Visual Recognition Challenge (Krizhevsky *et al.*, 2012). Subsequently, CNNs, and DL more generally, have become mainstream in the medical image segmentation field. UNet and VNet CNNs have demonstrated their profound impact in numerous medical image segmentation problems (Bakator and Radosav, 2018; Lundervold and Lundervold, 2019). Adoption has been enhanced through transfer learning to cope with limited datasets common in the medical imaging field (Tajbakhsh *et al.*, 2016). In a recent review of DL-based lung image analysis studies, Astley et al. identified a significant gap in DL-based lung MRI segmentation studies (n=7) with only one published conference proceeding (Astley *et al.*, 2020a) and one journal article (Tustison *et al.*, 2019) evaluating DL for hyperpolarised gas MRI segmentation. Tustison et al. used a 2D UNet for hyperpolarised gas MRI segmentation on a dataset of 113 images, developing a novel template-based method to augment the limited lung imaging data alongside pre-processing techniques, including N4 bias correction and adaptive denoising. A mean±SD DSC between DL and manual segmentations of 0.94±0.03 was achieved (Tustison *et al.*, 2019). However, the application of DL on a more extensive dataset with a broader range of pathologies is required prior to clinical adoption.

In this work we conducted extensive parameterisation experiments to determine the best-performing 3D CNN architecture, loss function and pre-processing techniques for hyperpolarised gas MRI segmentation. We further evaluated five DL methods using the best performing configuration to accurately, robustly and rapidly segment ventilated lungs on hyperpolarised gas MRI scans. Using a diverse testing set, with both helium-3 (^{3}He) and xenon-129 (^{129}Xe) noble gas scans and corresponding expert segmentations, we evaluated and compared performance using a range of evaluation metrics. We also investigated the effect of the noble gas on DL performance. Furthermore, we compared the best performing

DL method to conventional approaches. Finally, ventilated lung volume correlation and agreement were assessed for the best-performing DL method compared to expert-derived volumes.

4.3 Materials and Methods

4.3.1 Hyperpolarised gas MRI acquisition

All subjects underwent 3D volumetric ^{3}He or ^{129}Xe hyperpolarised gas MRI with full lung coverage in the coronal plane at 1.5T on a HDx scanner (GE Healthcare, Milwaukee, WI) using 3D steady-state free precession (SSFP) sequences as previously described (Horn *et al.*, 2014; Stewart *et al.*, 2015; Tahir *et al.*, 2018). Flexible quadrature radiofrequency coils were employed for transmission and reception of MR signals at the Larmor frequencies of ^{3}He and ^{129}Xe. In-plane (x-y) resolution of scans for both gases was 4x4mm^{2} (matrix = 512x512). ^{129}Xe scans ranged from 16-34 slices with a mean of 23 slices and slice thickness of 10mm. The ^{129}Xe acquisition protocol used the following settings: repetition time/echo time of 6.7/2.2 milliseconds, in-plane resolution of ~4x4 mm^{2} with a slice thickness of 10 mm. A ~40 cm field of view with a flip angle of 9° or 10° at a bandwidth of ±8kHz was used. ^{3}He scans ranged from 34-56 slices with a mean of 45 slices and slice thickness of 5mm. The ^{3}He acquisition protocol used the following settings: repetition time/echo time of 1.9/0.6 milliseconds, in-plane resolution of ~4x4 mm^{2} with a slice thickness of 5 mm. A ~40 cm field of view with a flip angle of 10° at a bandwidth of ±166.6kHz was used.

4.3.2 Dataset

The imaging dataset used in this study was pooled retrospectively from several research studies and clinical studies of patients referred for hyperpolarised gas MRI scans. Data use was approved by the Institutional Review Boards at the University of Sheffield and the National Research Ethics Committee. All data was anonymised and all investigations were conducted in accordance with the relevant guidelines and regulations.

The dataset consisted of 759 volumetric hyperpolarised gas MRI scans (23265 2D slices), with either ^{3}He (264 scans, 11880 slices) or ^{129}Xe (495 scans, 11385 slices), from 341 subjects. The slices were distributed approximately 50:50 between ^{3}He and ^{129}Xe. The dataset contained healthy subjects and patients with various pulmonary pathologies: asthma, COPD, asthma/COPD overlap, bronchiectasis, interstitial lung disease (ILD), idiopathic pulmonary fibrosis (IPF), lung cancer, cystic fibrosis (CF), children born prematurely, and patients investigated for possible airway disease. Demographic and clinical data for these subjects are summarised in Table 4.1. The dataset contains ^{129}Xe and ^{3}He scans acquired using either free-breathing or breath-hold acquisitions at a range of inflation levels, including functional residual capacity (FRC), total lung capacity (TLC) residual volume (RV).

Each of the 759 scans in the dataset has a corresponding, manually-edited expert segmentation, representing the ventilated region of the lungs. These scans and segmentations were collected from numerous retrospective studies; consequently, the segmentations were generated using several semi-automated methods (Hughes *et al.*, 2018; Biancardi *et al.*, 2018) and edited by multiple expert observers. Quality control was conducted by an experienced imaging scientist who identified potential errors and manually corrected them to ensure segmentation accuracy; the airways were removed down to the third generation, and it was ensured that no voxels were outside of the lung parenchymal region defined by a structural ^{1}H MRI scan, thereby removing background noise.

Table 4.1 Summary of demographics, clinical characteristics and image dataset information stratified by disease.

Disease	Total number of scans	Number of patients	Number of HP gas scans		Sex*		Median (range) age*	Mean±SD ventilated lung volume (litres)*
			^{3}He	^{129}Xe	Male	Female		
Healthy	43	33	1	42	15	13	12 (9, 76)	3.78 ± 1.18
Asthma	169	81	4	165	28	52	50 (13, 73)	4.23 ± 1.03
Asthma / COPD overlap	11	5	0	11	0	5	56 (45, 67)	4.13 ± 0.68
Bronchiectasis	3	3	1	2	1	1	15 (9, 29)	3.76 ± 1.00
CF	247	58	134	113	29	28	16 (6, 48)	3.65 ± 1.05
COPD	62	23	56	6	4	5	64 (52, 80)	4.43 ± 0.71
Non-IPF ILD**	77	41	0	77	25	16	69 (39, 83)	3.78 ± 0.80
Investigation for possible airways disease	38	21	5	33	2	16	49 (36, 69)	3.89 ± 1.05
IPF	46	20	45	1	17	3	72 (52, 80)	3.87 ± 0.71
Lung cancer	22	16	14	8	10	6	69 (34, 85)	4.12 ± 0.86
Preterm birth	41	40	4	37	15	25	12 (9, 14)	2.75 ± 0.55

*Data for 25 patients was unavailable.

**Contains connective tissue disease-associated ILD (CTD-ILD), hypersensitivity pneumonitis and drug-induced ILD (DI-ILD).

Abbreviations: hyperpolarised (HP), cystic fibrosis (CF), chronic obstructive pulmonary disease (COPD), interstitial lung disease (ILD), idiopathic pulmonary fibrosis (IPF), standard deviation (SD)

4.3.3 Parameterisation experiments

Parameterisation is rarely conducted in the DL literature due to its time-consuming nature, computational cost and the possibilities of biasing future investigations using the same dataset. To this end, we used a subset of the data as the parameterisation dataset comprising 431 hyperpolarised gas MRI scans (55% of the total data), with either ^{3}He (n=173) or ^{129}Xe (n=258) from a subset of the diseases present in the total dataset. 29 scans were used for the parameterisation testing set. In this study, we conducted three distinct parameterisation experiments that were completed sequentially; the outcome of the previous experiment influenced the inputs used in the subsequent experiment. We assessed the following hyperparameters:

Fully convolutional network architecture - The network architecture concerns the structure of the neural network. All the networks tested are 3D fully convolutional networks, which represent a subset of CNN architectures. The choice was made to evaluate CNNs as they are by far the most common network architectures in the lung image segmentation literature (Astley *et al.*, 2020b). Here, we focused on 3D CNNs due to the volumetric nature of features present in hyperpolarised gas MRI, lending itself towards analysis in a 3D view.

CNN loss function - Loss functions, or cost functions, are used to optimise the network. The loss function dictates how a network's weights and biases are updated for a given training pass; therefore, they represent a key component to any neural network and are important for generating accurate segmentations.

Pre-processing - Pre-processing is defined as an action taken to modify the training and testing images before they are passed to the network architecture. Pre-processing is often used to enhance features of the image, to accentuate the differences between foreground and background voxels or to remove noise from the image. Numerous studies in the DL-based MRI segmentation literature have employed several pre-processing strategies, including normalisation, denoising and N4 bias correction (Astley *et al.*, 2020b).

Each experiment trained a CNN for 30,000 iterations on an NVIDIA Tesla V100 graphical processing unit (GPU). Performance was assessed at intervals of 5000 iterations to determine the optimal number of iterations for each network. Shapiro-Wilk tests were performed for each experiment to determine normality and appropriate parametric or non-parametric statistical tests were conducted accordingly. The first parameterisation experiment compared the following four 3D CNNs:

VNet: A 3D fully convolutional neural network trained end-to-end on volumetric MRI. The network consists of convolutional compressions and subsequent decompression stages until the image is the original size. Each operation is conducted with valid padding. Convolution operations decrease in size from 5x5x5 with a stride of 1 to 2x2x2 with a stride of 2 (Milletari *et al.*, 2016). The network uses a spatial window size of [96, 96, 32] and a batch size of 6.

Dense VNet: Similar to VNet, this CNN employs convolution and deconvolution operations (Milletari *et al.*, 2016) with the addition of batch-wise spatial dropout, dense feature stacks and an explicit spatial prior (Gibson *et al.*, 2018a). Similar to the VNet, we used a spatial window size of [96, 96, 32] and a batch size of 6.

nn-UNet: The UNet is a common 2D encoder-decoder network; here, we used a 3D implementation of the UNet modified to reduce memory constraints, allowing 30 feature channels (Isensee *et al.*, 2019). Convolution operations vary in size from 3x3x3 to 1x1x1 depending on the layer of the network. The network also makes use of instance normalisation. An isotropic spatial window size of [96, 96, 96] was used with a batch size of 2.

HighResNet: A 3D fully convolutional neural network containing 20 layers, the first seven of which used 3x3x3 kernels to capture low-level features. Subsequent layers are dilated by either 2 or 4, with the number of kernels increasing from 16 to 64 to capture high-level features (Li *et al.*, 2017). Every two layers are grouped with residual connections to form a residual block. An isotropic spatial window size of [96, 96, 96] was used with a batch size of 2.

Figure 4.1 displays the results of the four 3D CNNs, showing mean performance on the parameterisation testing set. All networks show improved performance as iterations increase for the DSC and average Hausdorff distance at the boundary (Avg HD) metrics. At 30,000 iterations, we compared performance using DSC and Avg HD across the four network architectures. A Friedman test indicated that there was a statistically significant difference between architectures $X^2(4)=50.75$, $p<0.0001$. Pairwise comparisons were conducted with Bonferroni correction for multiple comparisons. Post-hoc analysis using the DSC and Avg HD metrics showed that both the nn-UNet and the VNet significantly outperformed the other networks tested (p<0.05); however, no statistical difference was observed between the nn-UNet and VNet architectures. Consequently, we could not conclude which network generates the most accurate hyperpolarised gas MRI segmentations for the parameterisation dataset. Hence, the loss function experiments were performed for both network architectures.

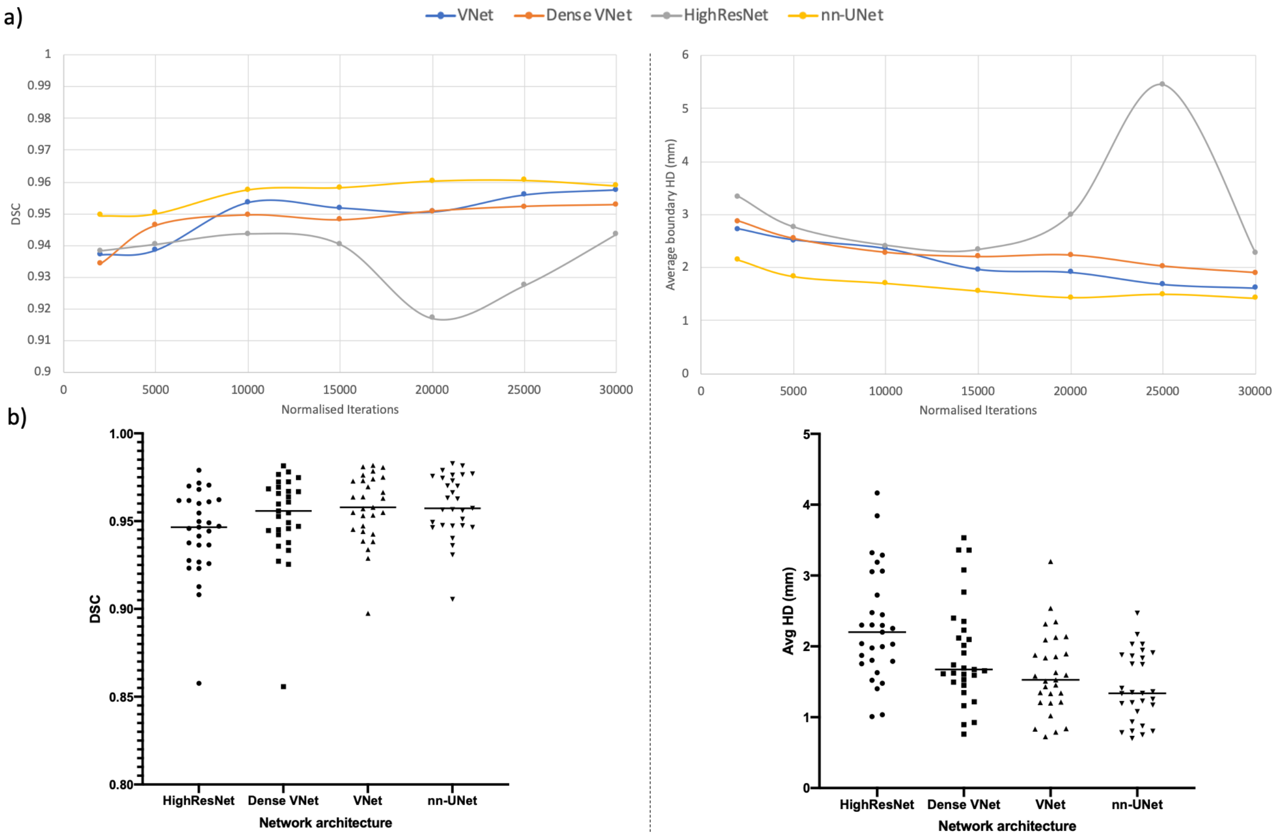


Figure 4.1 Results for the network architecture parameterisation experiments showing performance at (a) intervals of 5000 iterations and (b) 30000 iterations for four 3D CNNs in terms of DSC (left) and Avg HD (right).

We compared three common loss functions, namely, the binary cross-entropy (BCE) loss, the dice loss and the Tversky loss. The BCE loss function is defined below:

$$\mathrm{BCE(PR, GT)} = -\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}[\mathrm{gt}_i\, log(\mathrm{pr}_i) + (1 - \mathrm{gt}_i)\, log(1 - \mathrm{pr_i})] \tag{4.1}$$

where $\mathrm{GT} = \{\mathrm{gt_i} \in \mathrm{GT}\}$ denotes the manually-edited ground truth segmentation, $\mathrm{PR} = \{\mathrm{pr_i} \in \mathrm{PR}\}$ the predicted segmentation by the network and i represents the voxel location within the image, which is assumed to have N number of voxels. The Dice loss, which has shown promise in medical image segmentation tasks (Milletari *et al.*, 2016) is defined below:

$$\mathrm{Dice}(\mathrm{PR},\mathrm{GT}) = \frac{2\sum_i^{\mathrm{N}} \mathrm{pr}_i \mathrm{gt}_i}{\sum_i^{\mathrm{N}} \mathrm{pr}_i^2 + \sum_i^{\mathrm{N}} \mathrm{gt}_i^2} \tag{4.2}$$

The Tversky loss (Salehi *et al.*, 2017) provides a similar function as the Dice loss; however, it can be weighted to bias the loss function in favour of false positives and false negatives. The Tversky loss is defined below:

$$\mathrm{Tversky}(\mathrm{PR},\mathrm{GT},\alpha\,,\beta) = \frac{\sum_i^{\mathrm{N}} \mathrm{pr}_i \mathrm{gt}_i}{\sum_i^{\mathrm{N}} \mathrm{pr}_i \mathrm{gt}_i + \alpha \sum_i^{\mathrm{N}} \frac{\mathrm{pr}_i}{\mathrm{gt}_i} + \beta \sum_i^{\mathrm{N}} \frac{\mathrm{gt}_i}{\mathrm{pr}_i}} \tag{4.3}$$

where α and β are constants that weight the network's performance towards false positives or false negatives. For this work, we used α + β = 1, which reduces the Tversky loss to a set of F_β scores. This has been shown to work well for imbalanced data (Salehi *et al.*, 2017). Results for all three loss functions are shown in Figure 4.2.

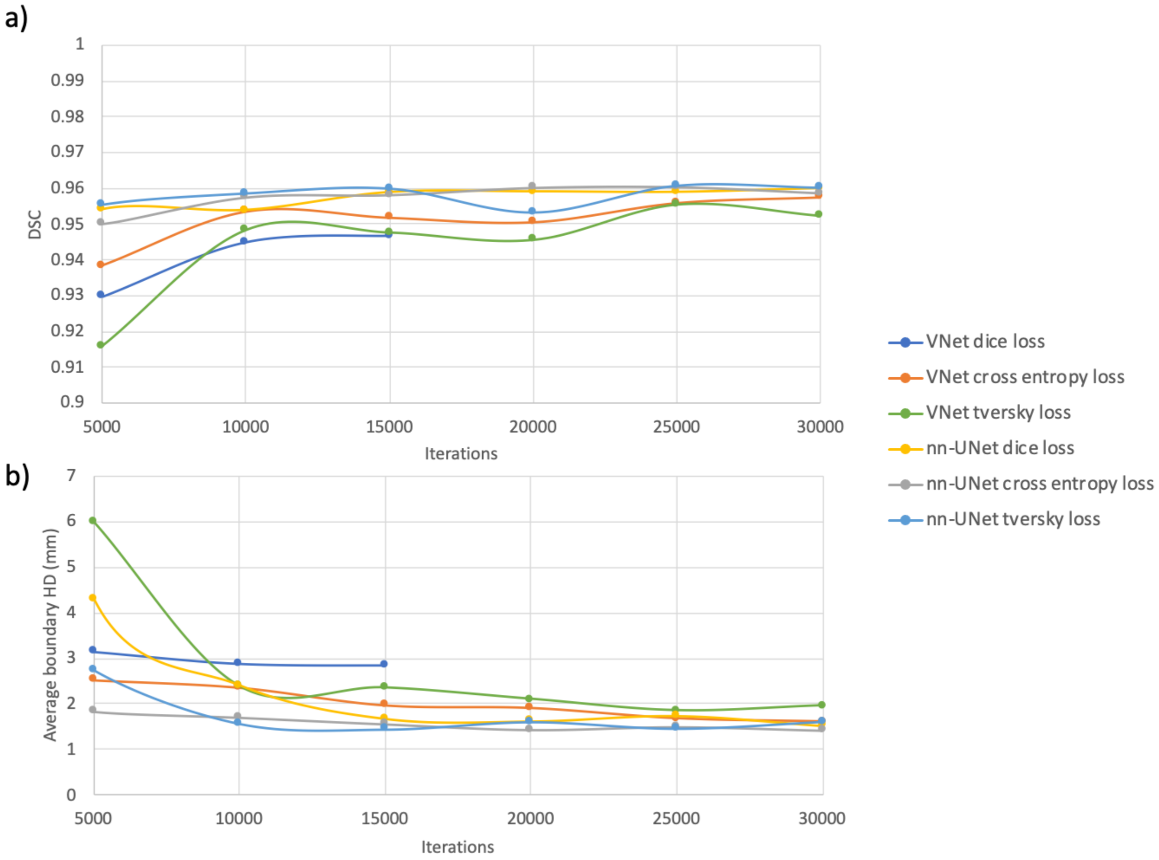


Figure 4.2 Results for both the nn-UNet and VNet architectures using three common loss functions, namely, the binary cross-entropy loss, dice loss and Tversky loss evaluated with a) DSC and b) Avg HD metrics at intervals of 5000 iterations up to 30000 iterations.

Figure 4.2 indicates no meaningful difference at 30000 iterations between any of the loss functions on either network, except for the VNet with Dice loss which exhibited inferior performance. This configuration failed due to an exploding gradient; hence, results are only available up to 15000 iterations. A Friedman test was performed indicating no significant differences between the nn-UNet architectures or the VNet with BCE loss. Consequently, we used the BCE loss and continued to evaluate both the nn-UNet and VNet architectures further for the impact of image pre-processing on performance.

We evaluate the impact of three commonly used pre-processing techniques for hyperpolarised gas MRI, namely, normalisation, denoising (Manjon *et al.*, 2010) and N4 bias correction (Tustison *et al.*, 2010). In addition, we compared the previous strategies to a combination of all pre-processing techniques and unprocessed images. Figure 4.3a shows that the segmentations produced by the combination of all pre-processing methods perform significantly worse than the other pre-processing methods alone and the images with no pre-processing. Figure 4.3b indicates the performance of each pre-processing method at 30000 iterations.

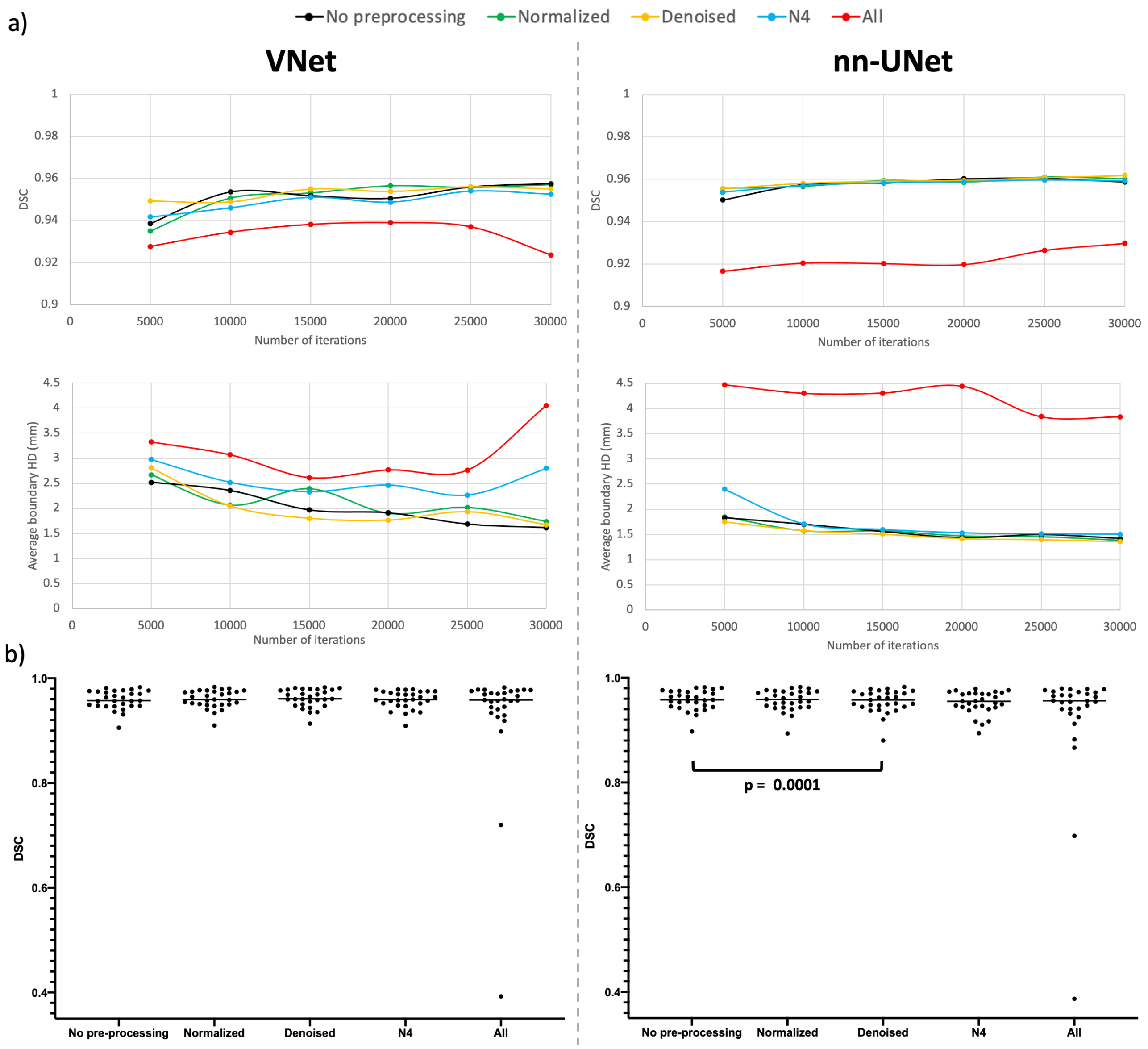


Figure 4.3 a) Results for both the nn-UNet and VNet architectures using common pre-processing strategies evaluated with DSC and Avg HD at intervals of 5000 iterations up to 30000 iterations. b) Results for pre-processing experiments comparing performance at 30000 iterations in terms of DSC.

Friedman tests were conducted for both the VNet ($X^2(5)=23.28$, p=0.0001) and nn-UNet ($X^2(5)=25.08$, p<0.0001) architectures, indicating significant differences between pre-processing strategies. Pairwise comparisons were conducted with Bonferroni correction for multiple comparisons. No pre-processing was selected as the control group for this experiment. Post-hoc comparisons of the VNet architecture indicated that no pre-processing method provides a statistically significant improvement over scans without pre-processing. Post-hoc comparisons of the nn-UNet architecture indicated that denoising provides a statistically significant improvement over images without pre-processing ($p<0.05$); no significant differences were observed for the other pre-processing strategies. Accordingly, we compared the nn-UNet denoised model with the VNet unprocessed model using a

Wilcoxon signed-rank test and observed that the nn-UNet denoised model exhibited superior performance (p<0.0001). Performance at 5000 iteration intervals for the nn-UNet denoised and the VNet unprocessed models are shown in Figure 4.4.

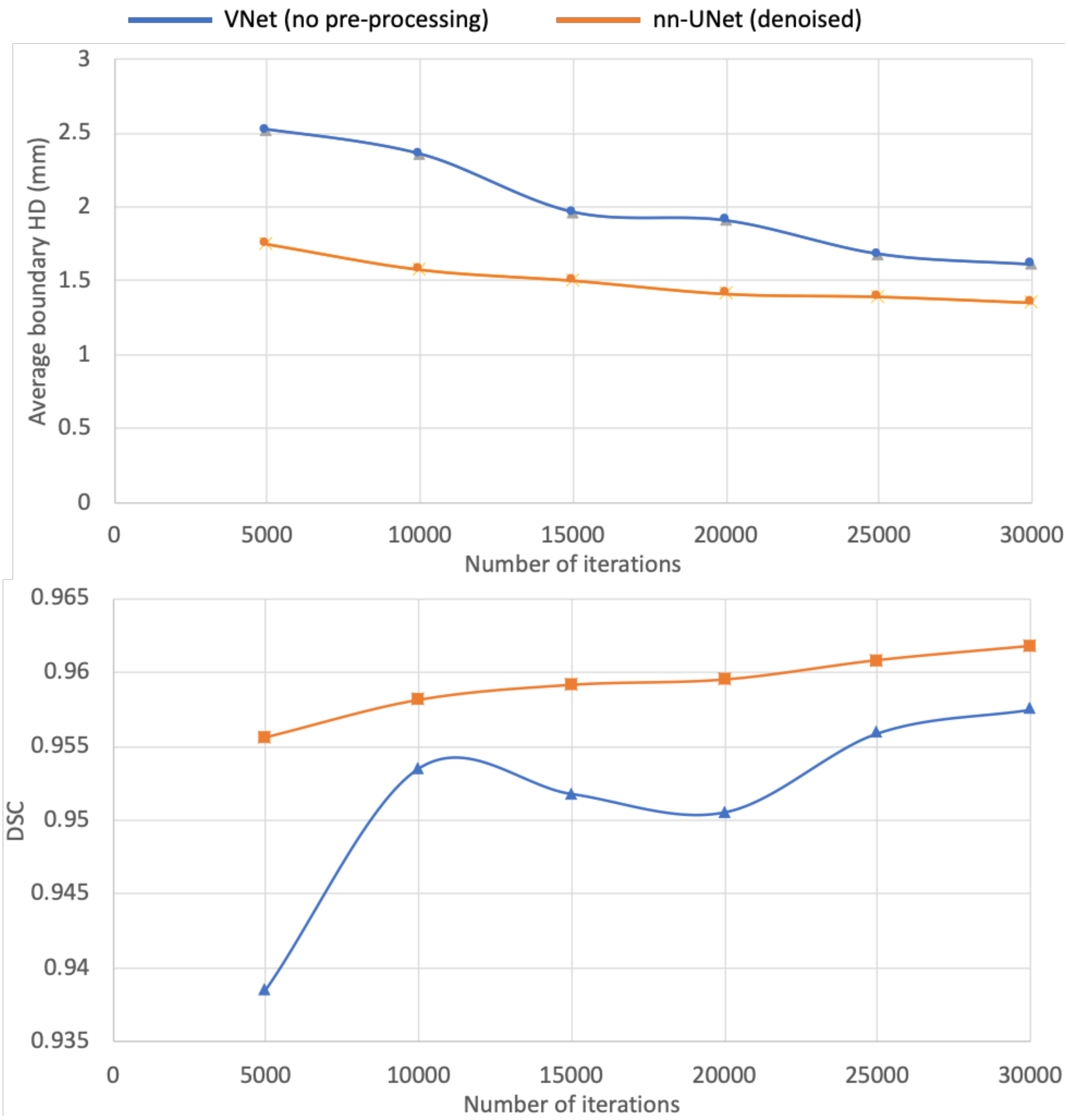


Figure 4.4 Comparison of performance for the nn-UNet denoised and VNet no pre-processing models using Avg HD (top) and DSC (bottom) at intervals of 5000 iterations up to 30000 iterations.

Based on the parameterisation experiments conducted on a subset of our available data, we determined that for our hyperpolarised gas MRI segmentation problem, the nn-UNet architecture with BCE loss using denoised images generates the best performing segmentations and, therefore, constitutes the optimal configuration for future investigations. Conducting these experiments on a subset of the total data allows for optimisation of parameters without introducing potential biases to specific training and testing sets. The following section describes the data split and DL parameters, informed by the above investigations, used in the remainder of this work.

4.3.4 Convolutional neural network

We used the nn-UNet fully convolutional neural network which processes 3D scans using volumetric convolutions. The network is trained end-to-end using hyperpolarised gas MRI volumetric scans. We use a 3D implementation of the UNet which has been modified to reduce memory constraints, allowing 30 feature channels (Isensee *et al.*, 2019). Convolution operations vary in size from 3x3x3 to 1x1x1 depending on the layer of the network. The network also makes use of instance normalisation. An isotropic spatial window size was used of [96, 96, 96] with a batch size of 2. A high-level visual representation of the 3D nn-UNet, specific to the spatial window sizes used, is shown in Figure 4.5.

The network utilises a non-linear PReLU activation function (He *et al.*, 2015) and is optimised using a binary cross-entropy loss function. ADAM optimisation was used to train the CNN (Kingma and Ba, 2015) and instance normalisation was conducted for each pass. The spatial window size was [96, 96, 96] with a batch size of 2. A learning rate of 1x10^{-5} was used for initial training and 0.5x10^{-5} for subsequent fine-tuning methods.

Each hyperpolarised gas MRI scan was pre-processed using spatially adaptive denoising, designed to consider both Rician noise and spatially varying patterns of noise. Denoising was implemented with ANTs 2.1.0 using the *DenoiseImage* function across three dimensions. Standard parameters were used (Manjon *et al.*, 2010). Constrained random rotation and scaling was used for data augmentation. Rotation with limits -10° to 10° and scaling of -10% to 10%, where a random rotation or scaling were applied at an interval within those limits, were used. A different random value was computed for each rotation axis and scaling factor.

All networks were trained using the medical imaging DL framework NiftyNet 0.6.0 (Gibson *et al.*, 2018b) built on top of TensorFlow 1.14 (Abadi *et al.*, 2016). Training and inference were performed on an NVIDIA Tesla V100 GPU with 16 GB of RAM.

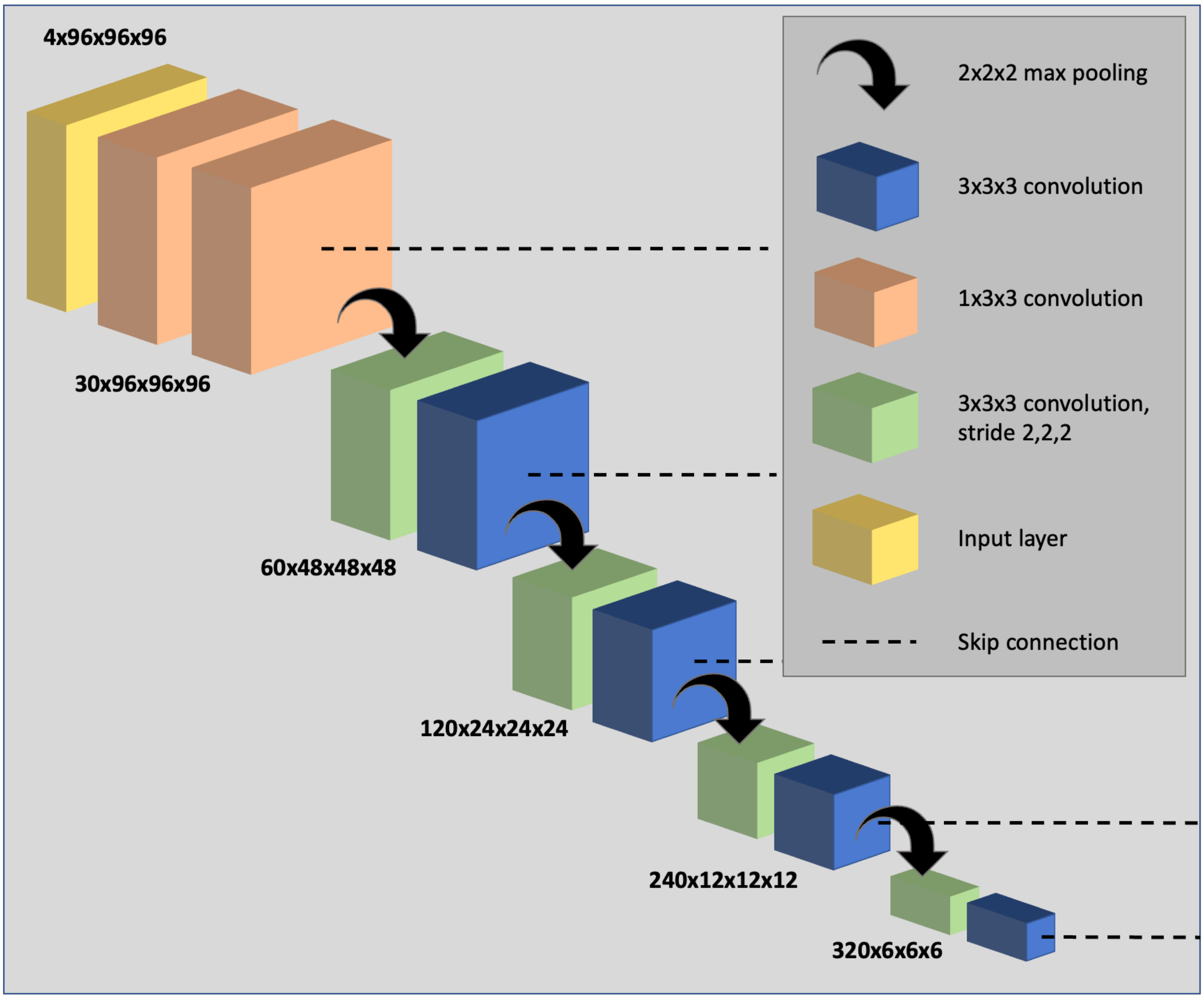


Figure 4.5 Visual representation of the modified 3D nn-UNet network used in this work. The deconvolution side of the network is omitted as it follows the same structure as the convolutional path, however, with the addition of a 1x1x1 SoftMax layer.

The dataset was randomly split into training and testing sets with 75% and 25% of the data respectively, in terms of the number of subjects. The training set, therefore, contained 237 ^{3}He scans (10902 slices) and 436 ^{129}Xe scans (10028 slices) from a total of 255 subjects. 86 scans, each from a different subject, were selected for the testing set (^{3}He: 27 scans (1242 slices); ^{129}Xe: 59 scans (1357 slices)). Repeat or longitudinal scans from multiple visits for the same patient were contained in the training set; however, no subject was present in both the training and testing sets, with the testing set containing only one scan from each patient. This was ensured by randomly selecting only one scan from each subject in the testing set and discarding the remaining scans; these scans are not included in Table 4.1. The range of diseases in the testing set was representative of the dataset as a whole. In addition, it was specified that there would be no overlap between the newly defined testing set and the previous testing set used for parameterisation experiments in terms of either patient or scan.

4.3.5 DL experimental methods

Five DL experimental methods were performed to train the network:

(1) The model was trained on 237 ^{3}He scans for 30000 iterations.

(2) The model was trained on 436 ^{129}Xe scans for 30000 iterations.

(3) The model was trained on 237 ^{3}He scans for 20000 iterations; these weights were used to initialise a model trained on 436 ^{129}Xe scans for 10000 iterations.

(4) The model was trained on 436 ^{129}Xe scans for 20000 iterations; these weights were used to initialise a model trained on 237 ^{3}He scans for 10000 iterations.

(5) The model was trained on 436 ^{129}Xe and 237 ^{3}He scans for 30000 iterations.

The five experimental methods were applied to the data split defined above using the same testing set for each method, facilitating comparison between the five methods to identify the best performing training method across multiple metrics.

4.3.6 Comparison to conventional methods

For further benchmarking, the best-performing DL method was compared against two other conventional machine learning methods, namely, SFCM and K-means clustering. The methods used are described as follows:

1) The k-means clustering algorithm used here was previously modified for hyperpolarised gas MRI segmentation (Kirby *et al.*, 2012a). This method attempts to find k data points, given the integer k, in an n-dimensional space R^{n} given m data points. These k data points are known as centres/centroids and the aim is to minimise the distance from each data point (m) to its centre/centroid (Kanungo *et al.*, 2002). The previously developed method (Kirby *et al.*, 2012a) attempts to delineate the image data into a number of clusters that can best represent a radiologist's analysis of the ventilation image with clusters defined from defects to hyperintense signal. The first stage of this method requires image normalisation into the range of 0-255, following which the cluster initial centres are set at 25% intervals between these values. A two-stage clustering process was applied with four clusters in the first stage, the lowest of which contains both signal void and hypointense signal. In the second stage, the clustering was reapplied to the lowest cluster from the first stage to define background, ventilation defect and hypointense signal regions.

2) The SFCM method used in this work has been reported previously (Hughes *et al.*, 2018); images are initially filtered to remove noise and maintain edges using a bilateral filter (Tomasi and Manduchi, 1998b). The standard FCM algorithm assigns *N* pixels to *C* clusters via Fuzzy memberships. The key assumption of the Spatial Fuzzy C-means is that pixels spatially close will have high correlation and hence have similarly high membership to the same cluster. This spatial information will modify the membership value only if, for example, the pixel is noisy and would have been incorrectly classified. The SFCM method makes use of nearby pixels during the iteration process by taking into account the membership of voxels within a predefined window (5x5 in this work) and will weight the central pixel depending on the provided weighting variables (Li *et al.*, 2011). The optimal number of clusters was manually selected by the observer.

4.3.7 Evaluation metrics

The testing set results for each of the five DL experimental methods and two conventional methods were evaluated using several metrics. The DSC was used to assess overlap between the sets of voxels in the ground truth (X) and DL-generated (Y) segmentations (Dice, 1945) and is defined as:

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

(4.4)

Two distance metrics, average boundary Hausdorff distance ($\mathrm{Avg\ HD}$) and 95th percentile Hausdorff distance ($\mathrm{HD95}$) were used (Beauchemin *et al.*, 1998) which assess the difference between the sets of boundary voxels in X and Y defined as $\mathrm{X_B}$ and $\mathrm{Y_B}$, respectively. The $\mathrm{HD95}$ is frequently used in the image segmentation literature to remove the impact of outlier voxels and is defined as the following:

$$\mathrm{HD95} = 95^{th}\ percentile(d(\mathrm{Y_B}, \mathrm{X_B})\,, d(\mathrm{X_B}, \mathrm{Y_B}))$$

(4.5)

where the $95^{th}\ percentile$ represents the bottom 95% of directed Hausdorff distances and $d(\mathrm{Y_B}, \mathrm{X_B})$ and $d(\mathrm{X_B}, \mathrm{Y_B})$ represent the directed Hausdorff distances between $\mathrm{X_B}$ and $\mathrm{Y_B}$ given by:

$$d(\mathrm{Y_B}, \mathrm{X_B}) = \frac{1}{\mathrm{N}} \sum_{\mathrm{y_B} \in \mathrm{Y_B}} \min_{\mathrm{x_B} \in \mathrm{X_B}} \|\mathrm{y_B} - \mathrm{x_B}\|$$

(4.6)

where $\mathrm{x_B}$ and $\mathrm{y_B}$ represent individual voxels in the set of $\mathrm{X_B}$ and $\mathrm{Y_B}$, N is the number of observations in $|\mathrm{Y_B}|$ and $\|\mathrm{y_B} - \mathrm{x_B}\|$ is the Euclidean distance between $\mathrm{y_B}$ and $\mathrm{x_B}$. The $\mathrm{Avg\ HD}$ is defined similarly as:

$$\mathrm{Avg\ HD} = \frac{1}{2}\left(d(\mathrm{Y_B}, \mathrm{X_B}) + d(\mathrm{X_B}, \mathrm{Y_B})\right)$$

(4.6)

The $\mathrm{Avg\,HD}$ reduces sensitivity to outliers and is regarded as a stable metric for segmentation evaluation (Shapiro and Blaschko, 2004). Furthermore, a relative error metric (XOR) was used to evaluate segmentation errors (Biancardi and Wild, 2017) as follows:

$$\mathrm{XOR} = \frac{|\mathrm{Y} \cap \mathrm{X}'| + |\mathrm{Y}' \cap \mathrm{X}|}{|\mathrm{X}|}$$

(4.8)

where Y' and X' are the complements of Y and X, respectively. The metric was used specifically because it is expected to correlate with the manual editing time required to correct the segmentation outcome.

4.3.8 Statistical analysis

Data were tested for normality using Shapiro-Wilk tests; when normality was not satisfied, non-parametric tests were conducted. One-way repeated-measure ANOVA or Friedman tests were conducted as appropriate with Bonferroni correction for post-hoc multiple comparisons to assess statistical significances of differences between experimental DL-based methods. Independent t-tests or Mann-Whitney U tests were used to compare differences between ^{3}He and ^{129}Xe segmentations in the testing set, assessing the effect of the noble gas. The best performing DL method was compared to other conventional segmentation methods using one-way repeated-measure ANOVA or Friedman tests with

Bonferroni correction for post-hoc multiple comparisons. Pearson or Spearman correlations and Bland-Altman analysis were conducted to compare volumes of DL-generated and expert segmentations. Statistical analysis was performed using Prism 8.4 (GraphPad, San Diego, CA) and SPSS Statistics 26.0 (IBM Corporation, Armonk, NY).

4.4 Results

4.4.1 Qualitative results

Segmentations for each of the five DL methods were generated for 86 testing set scans. Figure 4.6 shows examples of segmentation quality for a healthy subject and patients with four different pathologies across the five DL experimental methods using ^{3}He or ^{129}Xe. The original scans and expert segmentations are included to facilitate comparison. It can be observed that, in general, there are negligible voxels outside the lung parenchyma classed as ventilated and that the CNNs accurately excluded ventilation defects, as shown in the examples of the asthma and lung cancer patients. Case 4, of a healthy subject, represents an interesting case due to the presence of a zipper artefact caused by electronic noise in the hardware from external source of electromagnetic radiation. Zipper artefacts often present as a vertical band of alternating high and low intensity regions that occur across multiple slices in both ^{129}Xe and ^{3}He scans. It can be observed that some models are able to accurately exclude this artefact, whilst others remain unable to distinguish between the zipper artefact and ventilated lung voxels.

4.4.2 Quantitative evaluation

Figure 4.7 shows distributions of all four metrics for each DL method. The assumption of normality for each metric was not satisfied for all DL methods, as assessed by Shapiro-Wilk's tests ($p < 0.05$). As such, Friedman tests were run, determining that there were differences between DL methods for each metric. Post-hoc pairwise comparisons were performed for each metric with Bonferroni correction for multiple comparisons. The combined ^{3}He and ^{129}Xe method yielded statistically significant improvements over all DL methods using the DSC, XOR and HD95 metrics ($p<0.05$). However, using the Avg HD metrics, the combined ^{3}He and ^{129}Xe method significantly outperformed all but one DL method.

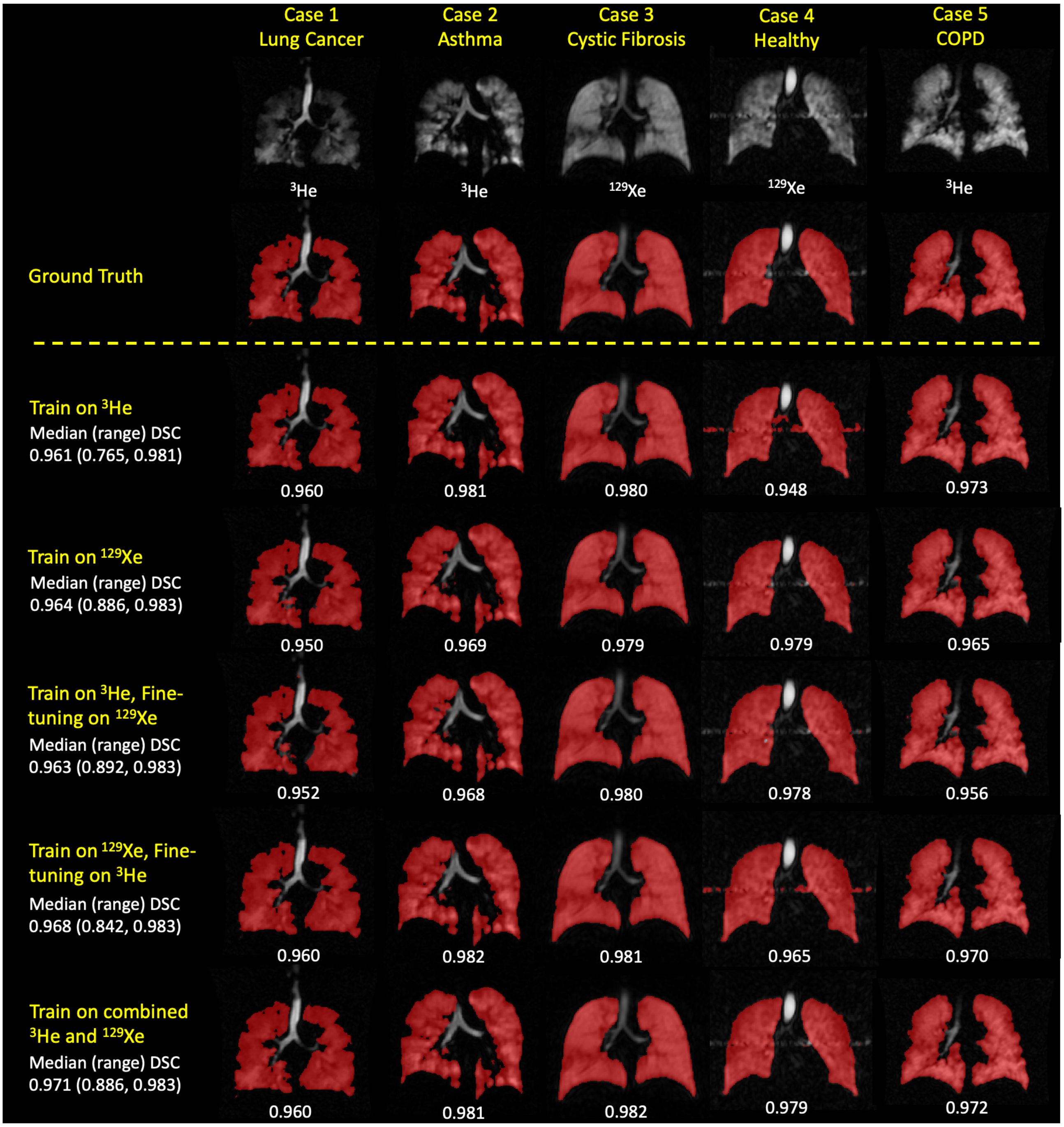


Figure 4.6 Example coronal slices for a healthy subject and four cases with different pathologies for each DL experimental method. Individual, and median (range), DSC values are displayed.

Table 4.2 summarises segmentation performance for the five DL experimental methods. The Combined ^{3}He and ^{129}Xe method generated the best segmentations using all four metrics.

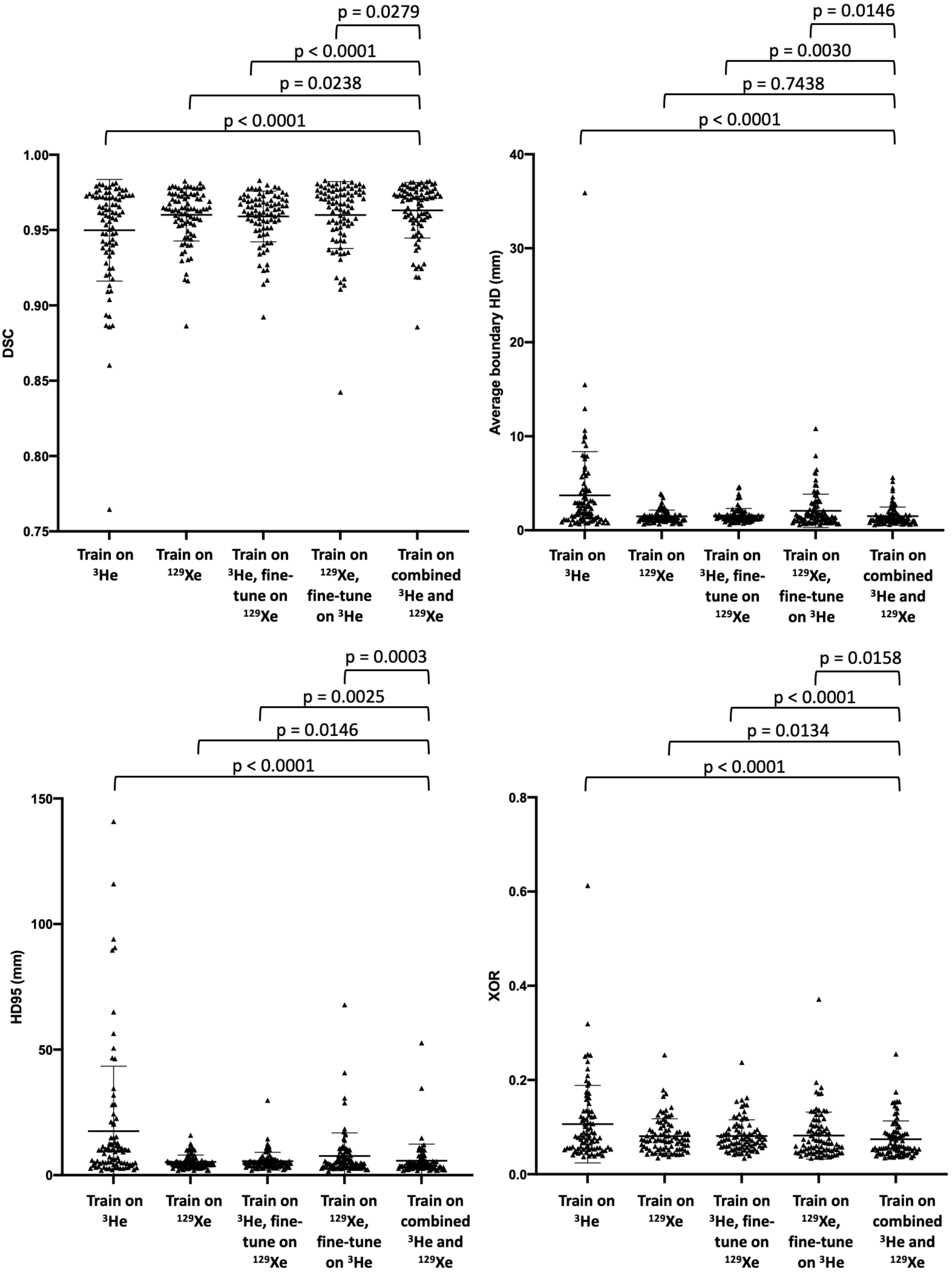


Figure 4.7 Comparison of segmentation performance on 86 testing scans for five DL experimental methods using the DSC, Avg HD, HD95 and XOR metrics (left to right). P-values are displayed for Friedman tests with Bonferroni correction for multiple comparisons, comparing the combined ^{3}He and ^{129}Xe DL method to the other DL methods. Mean and standard deviation values are marked by a bold line and whiskers, respectively.

Table 4.2 Comparison of segmentation performance for the five DL training methods for all scans in the testing set. Medians (ranges) are given; the best result for each metric is in bold.

Experimental DL methods	**Evaluation metrics: Median (range)**			
	DSC	Avg HD (mm)	HD95 (mm)	XOR
Train on ^{3}He	0.961 (0.765, 0.981)	2.335 (35.91, 0.644)	10.00 (140.9, 1.934)	0.079 (0.613, 0.037)
Train on ^{129}Xe	0.964 (0.886, 0.983)	1.341 (3.911, 0.675)	4.809 (15.90, 1.875)	0.072 (0.253, 0.035)
Train on ^{3}He, fine-tuned on ^{129}Xe	0.963 (0.892, 0.983)	1.384 (4.628, 0.636)	4.971 (29.80, 1.934)	0.075 (0.238, 0.034)
Train on ^{129}Xe, fine-tuned on ^{3}He	0.968 (0.842, 0.983)	1.483 (10.84, 0.596)	4.935 (67.85, 1.563)	0.066 (0.372, 0.034
Combined ^{3}He and ^{129}Xe training	**0.971 (0.886, 0.983)**	**1.234 (5.630, 0.594)**	**4.193 (52.70, 1.875)**	**0.059 (0.255, 0.035)**

Figure 4.8 shows the segmentation performance for the testing set stratified by noble gas (^{129}Xe or ^{3}He) using the DSC and Avg HD metrics. The majority of models show no significant difference between ^{129}Xe and ^{3}He for both metrics. Only two methods, namely, the 'Train on ^{3}He' and 'Train on ^{129}Xe, fine-tune on ^{3}He' methods, showed a significant difference between noble gases across both metrics.

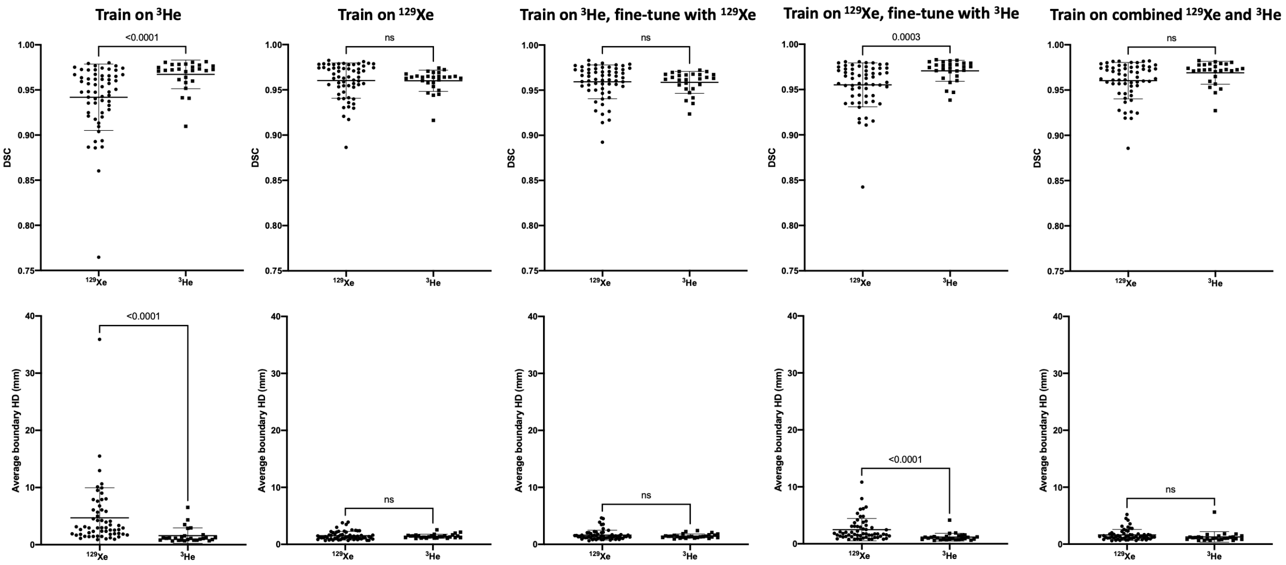


Figure 4.8 Comparison of DSC (top) and Avg HD (bottom) values for ^{129}Xe and ^{3}He testing scans for five DL methods. P-values between ^{129}Xe and ^{3}He using Mann-Whitney tests are shown. Mean and standard deviation values are marked by a bold line and whiskers, respectively.

4.4.3 Validation on 2D ^{3}He hyperpolarised gas MRI scans

Based on the results of the five experimental methods, the combined ^{3}He and ^{129}Xe DL model was identified as the most accurate DL ventilated lung segmentation method due to statistically significant improvements over all other methods using the DSC, HD95 and XOR metrics. Consequently, we tested the combined ^{3}He and ^{129}Xe DL model on 31 2D spoiled gradient-echo ^{3}He hyperpolarised gas MRI ventilation scans which differ in MRI sequence and acquisition parameters. We employed a dataset of 2D spoiled gradient-echo ^{3}He hyperpolarised gas MRI ventilation scans from 31 patients with either asthma (Tahir *et al.*, 2016) (n=12) or cystic fibrosis (Marshall *et al.*, 2017) (n=19) acquired at FRC+1L with full lung coverage at 1.5T on a HDx scanner (GE Healthcare, Milwaukee, WI). Helium was polarised on-site to around 25% polarisation (GE Healthcare, Amersham, England). Flexible quadrature radiofrequency coils were employed for transmission and reception of MR signals at the Larmor frequency of ^{3}He (Clinical MR Solutions, Brookfield, WI) with the following parameters: resolution of ~3x3x10mm^{2}, TR/TE equal to 3.6/1.1 milliseconds, field of view of 30-40cm, flip angle of 8° and bandwidth of ±63kHz. The ^{3}He hyperpolarised gas MRI scans differ from the scans used in the primary investigation in terms of both MRI sequence and acquisition parameters.

Figure 4.9 shows examples of segmentation quality for one asthma and one CF patient using the combined ^{3}He or ^{129}Xe DL trained model. The original scans and expert segmentations are included to facilitate comparison. The proposed model accurately excludes subtle and gross ventilation defects in the spoiled gradient-echo hyperpolarised gas MRI scans and excludes airways. Quantitative results of segmentation performance on the 2D spoiled gradient-echo hyperpolarised gas MRI scans are displayed in Table 4.3. The results indicate that the combined ^{3}He and ^{129}Xe DL trained model generated segmentations which largely agree with expert ground truth segmentations across a range of metrics.

Table 4.3 Summary of 2D spoiled gradient-echo results.

Segmentation method	**Evaluation metrics: Median (range)**			
	DSC	Avg HD (mm)	HD95 (mm)	XOR
Combined ^{3}He and ^{129}Xe DL model	0.965 (0.947, 0.983)	1.389 (2.030, 0.602)	4.323 (8.203, 1.934)	0.069 (0.107, 0.034)

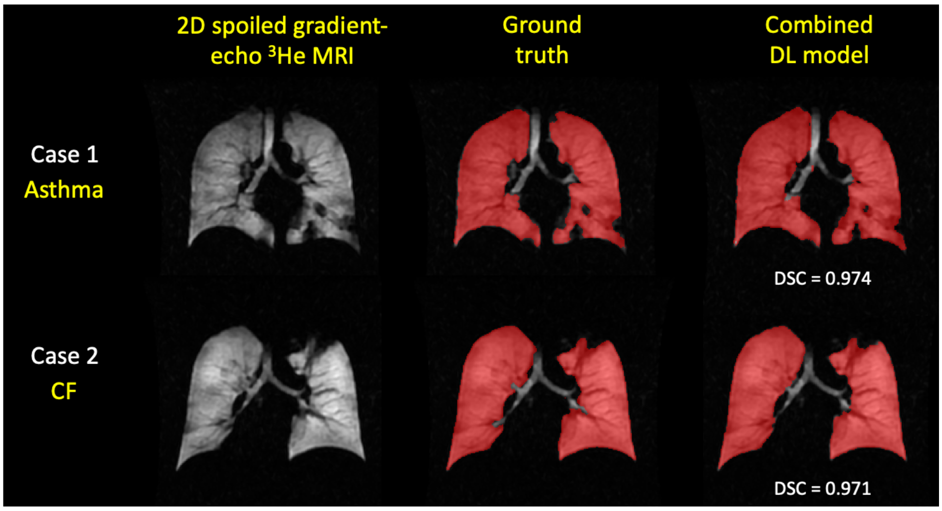


Figure 4.9 Example coronal slices for an asthma and CF case with ground truth segmentations and the segmentations generated by the combined ^{3}He and ^{129}Xe DL model. DSC values are displayed.

These results indicate that the proposed DL model, trained on a combined 3D SSFP ^{3}He and ^{129}Xe dataset, is generalisable to hyperpolarised gas MRI ventilation scans acquired with a sequence, namely, a 2D spoiled gradient-echo sequence. The segmentation performance on 2D spoiled gradient-echo hyperpolarised gas MRI scans, across all four metrics, are highly similar to the results observed on the 3D SSFP testing set.

4.4.4 Ventilated volume

Furthermore, ventilated volume was assessed for the combined ^{3}He and ^{129}Xe method. The assumption of normality was satisfied for DL and expert ventilated volume, as assessed by Shapiro-Wilk's tests (p>0.05). Pearson correlation and Bland-Altman analysis are shown in Figure 4.10 for the combined ^{129}Xe and ^{3}He model; the DL segmentation volume is highly correlated (r=0.99) with the expert segmentation volume and exhibits minimal bias (-0.8%).

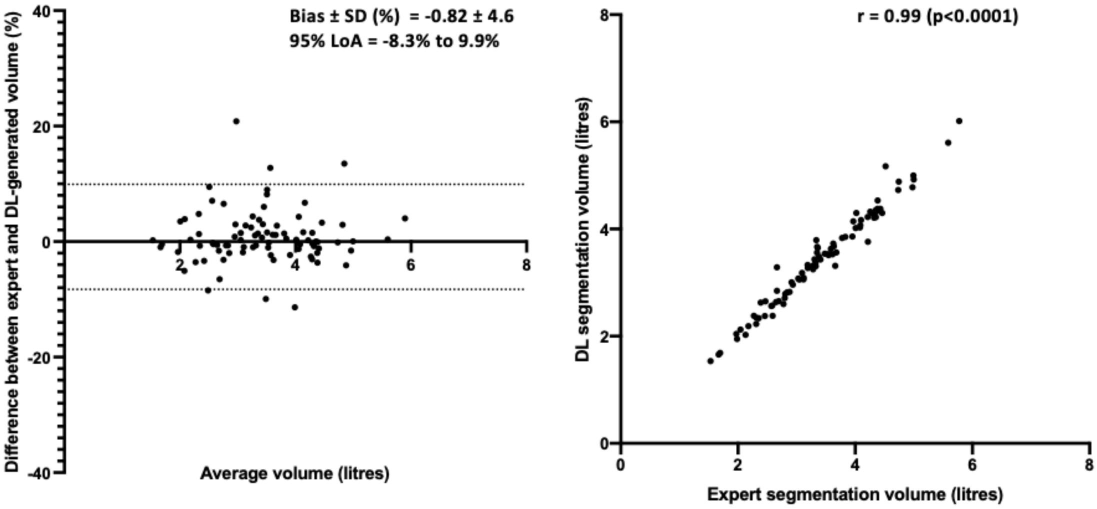


Figure 4.10 Pearson correlation and Bland-Altman analysis of lung volumes for 86 testing set cases compared to volumes derived from expert segmentations for the combined ^{3}He and ^{129}Xe DL.

4.4.5 Comparison to machine learning approaches

Figure 4.11 shows qualitative and quantitative performance for the DL combined ^{3}He and ^{129}Xe training method with two conventional segmentation methods, namely K-means clustering and SFCM across three cases. The assumption of normality for the DSC metric was not satisfied for conventional and DL approaches, as assessed by Shapiro-Wilk's tests. Post hoc Friedman's tests were performed with Bonferroni correction for multiple comparisons ($X^2(3)$, $p<0.0001$). The DL segmentation method exhibited significant improvements over conventional methods ($p<0.0001$), accurately excluding low-level noise and artefacts (e.g. Case 2) as well as non-lung regions such as the trachea and bronchi.

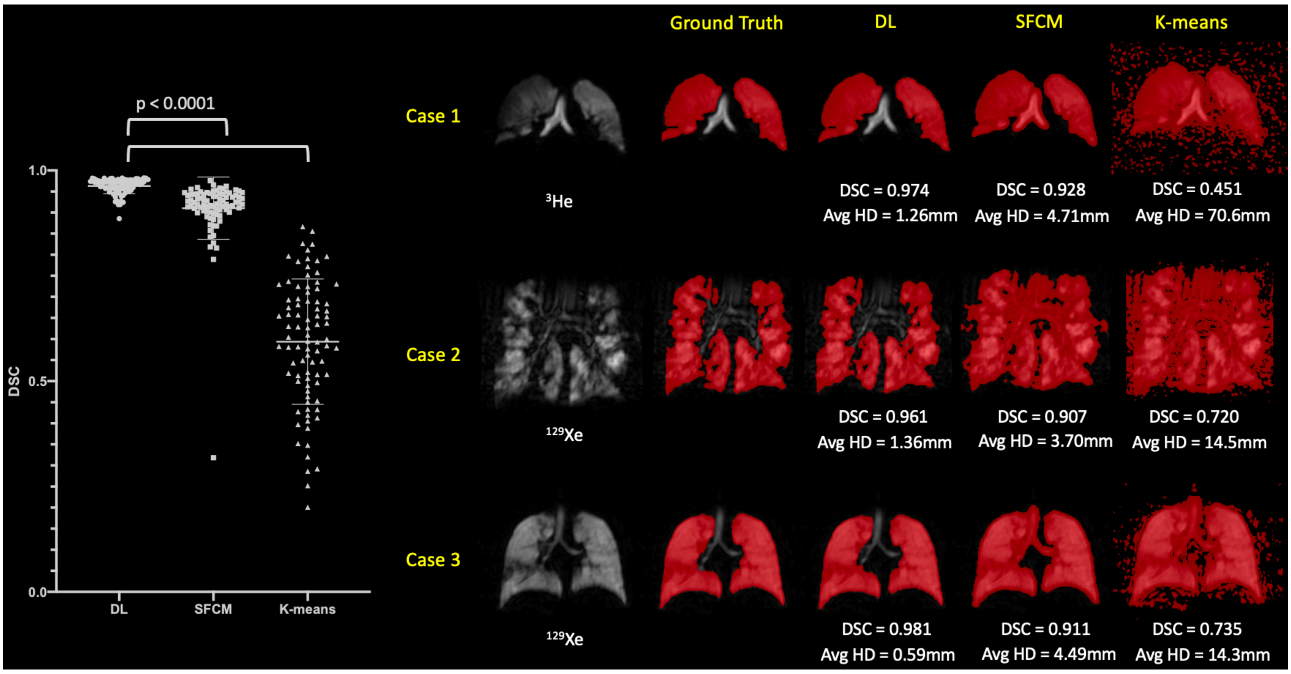


Figure 4.11 Comparison of performance on testing scans between the combined ^{129}Xe and ^{3}He DL method and conventional segmentation methods (SFCM and K-means) with P-values for Friedman tests with Bonferroni correction for multiple comparisons. Mean and standard deviation values are marked by a bold line and whiskers, respectively. Individual DSC and Avg HD values for each method are displayed for three cases.

4.5 Discussion

The DL segmentation methods yielded highly accurate segmentations across a range of evaluation metrics on the dataset used. To the best of the authors' knowledge, the hyperpolarised gas MRI dataset used here is the largest to date for ventilated lung segmentation, comprising 759 scans from patients with a wide range of lung pathologies. This is advantageous for preserving generalisability as it enables algorithms to learn features present in a range of diseases independent of the noble gas. Compared with ^{129}Xe MRI, ^{3}He MRI has an intrinsically stronger MRI signal due to the difference in gyromagnetic ratios between the two nuclei. Generally, lung ventilation information of similar diagnostic quality has been obtained with the two nuclei; despite this, there are known differences in lung diffusivity as well as differences in spatial resolution between the nuclei (Stewart *et al.*, 2018; Kirby *et al.*, 2012b). This is particularly important for deep learning applications as the resolutions of our ^{3}He and ^{129}Xe MRI scans differ greatly in the z-direction whereby ^{3}He and ^{129}Xe MRI scans have a slice thickness and an inter-slice distance of ~5mm and ~10mm, respectively. Therefore, it remains important to understand the performance of deep learning segmentation applications across the two nuclei.

The combined ^{3}He and ^{129}Xe DL method showed statistically significant improvements over all other methods using the DSC, HD95 and XOR metrics; however, using the Avg HD metric, no significant difference between the combined ^{3}He and ^{129}Xe method and ^{129}Xe only method was observed, perhaps attributable to an outlier case. Some statistically significant differences were observed in performance when comparing ^{3}He and ^{129}Xe testing set scans; however, the combined ^{3}He and ^{129}Xe method exhibited identical performance independent of the noble gas used. This indicates that, for a given ^{3}He or ^{129}Xe scan, the combined ^{3}He and ^{129}Xe method is unlikely to be biased towards a specific noble gas. Due to the current paucity and unpredictable supplies of ^{3}He worldwide, the field, in general, has transitioned towards the use of ^{129}Xe as the predominant noble gas for hyperpolarised gas MR ventilation imaging. As this trend continues, it may be pertinent in future work to assess the impact of training and testing solely on ^{129}Xe scans. In addition, external testing indicated the proposed model's ability to generalise across MRI sequence and acquisition parameters not seen in the training set, further reinforcing that the model is using functional features from hyperpolarised gas MRI to produce accurate segmentations.

The CNN produced more accurate segmentations than the two conventional approaches for all evaluation metrics. In particular, the CNN was able to deal with images containing background noise and artefacts, as well as successfully excluding ventilation defects and airways. In comparison, the SFCM method was unable to distinguish airways or artefacts and segmented these areas erroneously. As such, it is highly probable that the CNN eliminates or dramatically reduces the manual-editing time required after automatic segmentation. The K-means clustering algorithm exhibited poorer than expected performance, possibly attributable to the lack of an available proton MRI. This represents a benefit of the CNN-based method as only the hyperpolarised gas MRI scan is required as an input. Previous work in the literature that aimed to employ DL for hyperpolarised gas MRI segmentation used a 2D UNet and achieved a mean DSC of 0.94 (Tustison *et al.*, 2019). In comparison, our combined ^{3}He and ^{129}Xe method trained via a 3D nn-UNet yielded a mean DSC value of 0.96. The 3D CNN allows the model to treat the segmentation as a 3D volume and learns features present across multiple slices e.g. ventilation defects. Several pre-processing techniques have previously been used in the literature for lung image segmentation (Astley *et al.*, 2020b). The work of Tustison *et al.* (2019) utilises a novel template-based data augmentation strategy with N4 bias correction and denoising, which are computationally expensive and time-consuming; however, the impact of such techniques is not assessed in their work. In this study, we observed that N4 bias correction provided no significant benefit, while denoising yielded significant improvements.

All DL methods were trained and tested using a single GPU. Training required approximately nine days, while inference took 27 seconds per ^{129}Xe scan and 35 seconds per ^{3}He scan, corresponding to approximately one second per slice for both gases. Compared with conventional methods, such as SFCM, the time taken to generate automatic segmentations is significantly reduced from approximately 5 minutes to around 30 seconds, indicating the time-saving benefits of DL-based methods. Moreover, accurate automatic segmentation of hyperpolarised gas MRI ventilation scans through CNN-based approaches will eliminate or reduce manual editing time, thus improving clinical throughput. To further improve clinical translation of DL-based techniques, we have provided the trained DL model along with necessary files, enabling members of the pulmonary imaging community to apply the trained model in their own research.

The specific dataset used is unique within the context of pulmonary imaging due to the presence of numerous features such as different noble gases, longitudinal scans, repeat scans and pre- and post-treatment scans. The variation in the number of repeat or longitudinal scans and slice thicknesses between 3D ^{3}He and ^{129}Xe scans impeded us from achieving a training and testing set split equally between both gases; notwithstanding, the number of 2D slices were approximately equal between gases. Although multiple scans from the same patient were included in the training set to increase dataset numbers, to enhance the robustness of the evaluation, no scan of the same patient was present both in the training and testing sets. This study also represents the first large-scale investigation of architectures, loss functions and pre-processing techniques within the field of lung MRI. Selecting a subset of the data allowed us to perform parameterisation experiments to determine the ideal choice of network architecture, loss function and pre-processing technique, without creating optimisation biases in subsequent experiments. The conclusions of the parameterisation experiments were partially limited due to multiple factors; the same exact parameters cannot be used for each network due to the spatial imaging constraints of the specific network, such as requiring isotropic resolutions or the varying memory requirements of each architecture. This means that the windowing, batch size and bordering varies between architectures and can, therefore, make comparisons potentially difficult. However, where possible, we aimed to maintain consistent parameters across all networks tested. Further investigation may aim to optimise other hyperparameters that could be deemed equally important as the experiments conducted related to architecture, loss function and pre-processing; these may include the choice of activation function or optimisation algorithm. Furthermore, parameterisation results will vary based on the specific datasets used and, consequently, limit conclusions to the particular data used in these experiments.

Currently, segmentations edited by expert observers are the gold-standard for training supervised DL algorithms. Studies have shown that manual segmentations are susceptible to inter-observer variability (Mukesh *et al.*, 2012). Numerous research projects have employed techniques to create generalisable DL models across multiple institutions and observers (Balachandar *et al.*, 2020). A limitation of our study is the presence of only one expert segmentation per scan, which precludes the ability to evaluate intra- and inter-observer variability. Various studies have aimed to account for annotator intra- and inter-observer variability (Zhang *et al.*, 2020a; Tanno *et al.*, 2019; Zhang *et al.*, 2020b). However, the wide range of expert observers used to generate and manually edit the expert segmentations in this work led to significant variability in the training and testing sets. Hence, the CNN can learn a robust segmentation method invariant to the specific semi-automated method used to generate the ground truth or the expert observer who manually corrected it. In future work, multiple expert segmentations may be used to train the algorithm and allow the evaluation of inter-observer variability.

For the evaluation of certain clinically relevant metrics such as VDP (Woodhouse *et al.*, 2005), the whole-lung cavity volume is required in addition to ventilated lung volumes, most commonly computed from a whole-lung segmentation generated from a structural proton MRI scan. In this work, we showed that ventilated lung volumes derived from CNN-generated segmentations have a significant Pearson correlation of 0.99 and a minimal Bland-Altman bias of -0.8% with expert volumes. However, evaluation of DL-based methods using not only ventilated lung volume, but also VDP, would further the extensive validation required for clinical adoption.

4.6 Conclusion

In conclusion, we evaluated a 3D fully connected CNN using the nn-UNet architecture that is capable of producing accurate, robust and rapid hyperpolarised gas MRI segmentations on a large, diverse dataset. We compared five experimental DL methods and observed that combining ^{3}He and ^{129}Xe scans during training produces significantly improved segmentations. Compared with expert segmentations, this CNN-based method also showed a strong Pearson correlation and limited bias using Bland-Altman analysis. In addition, the CNN-based segmentation method significantly outperformed two conventional segmentation methods commonly used in the literature.

Chapter 5
Implementable deep learning for multi-sequence proton MRI lung segmentation: a multi-centre, multi-vendor, and multi-disease study

Background: Recently, deep learning via convolutional neural networks (CNNs), has largely superseded conventional methods for proton (^{1}H)-MRI lung segmentation. However, previous deep learning studies have utilised single-centre data and limited acquisition parameters.

Purpose: Develop a generalizable CNN for lung segmentation in ^{1}H-MRI, robust to pathology, acquisition protocol, vendor, and centre.

Study type: Retrospective.

Population: 809 ^{1}H-MRI scans from 258 participants with various pulmonary pathologies (median age (range): 57 (6–85); 42% females) and 31 healthy participants (median age (range): 34 (23–76); 34% females) that were split into training (593 scans (74%); 157 participants (55%)), testing (50 scans (6%); 50 participants (17%)) and external validation (164 scans (20%); 82 participants (28%)) sets.

Field strength/sequence: 1.5-T and 3-T / 3D spoiled-gradient recalled and ultrashort echo-time ^{1}H-MRI.

Assessment: 2D and 3D CNNs, trained on single-centre, multi-sequence data, and the conventional spatial fuzzy c-means (SFCM) method were compared to manually-delineated expert segmentations. Each method was validated on external data originating from several centres. Dice similarity coefficient (DSC), average boundary Hausdorff distance (Avg HD), and relative error (XOR) metrics to assess segmentation performance.

Statistical tests: Kruskal-Wallis tests assessed significances of differences between acquisitions in the testing set. Friedman tests with post-hoc multiple comparisons assessed differences between the 2D CNN, 3D CNN and SFCM. Bland-Altman analyses assessed agreement with manually-derived lung volumes. A p-value of <0.05 was considered statistically significant.

Results: The 3D CNN significantly outperformed its 2D analogue and SFCM, yielding a median (range) DSC of 0.961 (0.880–0.987), Avg HD of 1.63mm (0.65–5.45) and XOR of 0.079 (0.025–0.240) on the testing set and a DSC of 0.973 (0.866–0.987), Avg HD of 1.11mm (0.47–8.13) and XOR of 0.054 (0.026–0.255) on external validation data.

Data conclusion: The 3D CNN generated accurate ^{1}H-MRI lung segmentations on a heterogenous dataset, demonstrating robustness to disease pathology, sequence, vendor, and centre.

5.1 Preface

Work contained within this chapter has been submitted as a Journal article to the in Journal of Magnetic Resonance Imaging:

> **Astley J.R.**, Biancardi A.M., Hughes P.J.C, Marshall H., Collier G.J., Chan H.F., Saunders L.C., Smith L.J., Brook M.L., Thompson R., Rowland-Jones S., Skeoch S., Bianchi S.M., Hatton M.Q., Rahman N.M., Ho L.P., Brightling C.E., Wain L.V., Singapuri A., Evans R.A., Moss A.J., McCann G.P., Neubauer S., Raman B.; C-MORE/PHOSP-COVID Collaborative Group; Wild J.M., Tahir B.A. Implementable Deep Learning for Multi-sequence Proton MRI Lung Segmentation: A Multi-center, Multi-vendor, and Multi-disease Study. *Journal of Magnetic Resonance Imaging*. 2023 Feb 17. doi: 10.1002/jmri.28643. Epub ahead of print. PMID: 36799341.

The work contained within this chapter has also been published as conference proceedings at the following conference:

> **Astley J.R.**, Biancardi A.M., Marshall H., Smith L.J., Collier G.J., Hughes P.J.C., Walker M., Hatton M.Q., Wild J.M. and Tahir B.A. (2021). Generalizable deep learning for multi-resolution proton MRI lung segmentation in multiple diseases. The international society for magnetic resonance in medicine (ISMRM) 2021. *Online.*

Additional material that could not be included within the conference proceeding or the journal article is also contained within this chapter.

5.1.1 Author contributions

J.R.A., J.M.W. and B.A.T. made substantial contributions to the conceptualisation of the work. A.M.B., H.M. L.J.S., G.J.C., P.J.C.H., M.Q.H., J.M.W. and B.A.T. were involved with patient recruitment, image acquisition and/or analysis. J.R.A. conceptualised and performed the DL experiments and validation. J.R.A interpreted data and conducted statistical analyses. J.R.A. drafted the manuscript. B.A.T. substantively revised the manuscript. All authors reviewed and approved the submitted manuscript.

5.2 Introduction

Imaging of the lungs is a key component in the management of patients with respiratory diseases and facilitates their diagnosis, treatment planning, monitoring, and assessment. Imaging modalities such as computed tomography (CT) and proton MRI (^{1}H-MRI) enable the visualization and quantification of anatomical features within the lungs (Ivanovska *et al.*, 2016; Zeng *et al.*, 2019). High-resolution CT has traditionally represented the reference standard in clinical practice for structural lung imaging due to its impeccable resolution (~1mm^3) and ubiquitous availability (Whiting *et al.*, 2015). ^{1}H-MRI has historically been limited in the management of patients with respiratory diseases due to the low proton density and fast signal decay within the lungs which pose inherent challenges for the modality (Wild *et al.*, 2012). However, recent advances in sequence development and coil design have improved structural detail via ultrashort and zero echo-time sequences which increase the resolution to approximately that of CT (~1.5mm^3), enabling the use of ^{1}H-MRI in numerous pulmonary imaging applications (Voskrebenzev and Vogel-Claussen, 2021). Furthermore, ^{1}H-MRI uses non-ionising radiation and therefore can be utilised for paediatric patient care and treatment monitoring where longitudinal imaging studies are required.

Segmentation of the lungs in ^{1}H-MRI is required to delineate the lung cavity from other nearby features and has numerous applications, such as disease characterization (Liu *et al.*, 2021a), treatment planning (Crockett *et al.*, 2021) and longitudinal assessment (Pennati *et al.*, 2021). Lung segmentation is also required for the computation of quantitative dynamic contrast-enhanced and oxygen-enhanced MRI which evaluate lung perfusion and ventilation, respectively (Voskrebenzev and Vogel-Claussen, 2021). In addition, surrogates of ventilation can be derived from non-contrast, multi-inflation ^{1}H-MRI, requiring the segmentation of the lung parenchyma at different volumes (Kjorstad *et al.*, 2017). Segmentation of pathological lungs, in particular, represents a challenge due to the relative similarity in signal intensity between aerated and non-aerated lung tissue and the presence of various pathological patterns such as ground glass opacities, consolidation, and bronchiectasis.

Conventional image processing and machine learning approaches have traditionally been used for lung segmentation in ^{1}H-MRI; these include semi-automatic thresholding, clustering and region growing methods (Ivanovska *et al.*, 2016). Spatial fuzzy c-means (SFCM) is a clustering method that employs spatial information to modify cluster membership and has been used successfully as a semi-automated ^{1}H-MRI lung segmentation method (Hughes *et al.*, 2018; Biancardi *et al.*, 2018). However, although these methods achieved varying degrees of success, they remain semi-automated in nature. Time-consuming manual correction is often required to modify semi-automated methods based on MRI sequence or readout parameters.

In recent years, deep learning (DL) has largely superseded classical image processing, such as thresholding, and conventional machine learning, such as clustering, for medical image segmentation applications. Convolutional neural networks (CNNs) have emerged as the dominant DL approach and have been used in numerous pulmonary image segmentation applications. A recent review of DL applications in lung image segmentation indicated that studies predominantly utilised CT imaging and single-centre datasets (Astley *et al.*, 2020b). This leads to reduced performance when deploying DL models across multiple centres due to variations in training and testing set distributions (Raschka, 2018). Due to variations in MR acquisition protocols or vendor, the large-scale segmentation of ^{1}H-MRI represents a significant challenge for the deployment of implementable DL models. Multi-centre datasets have been used for other DL-based lung segmentation applications such as the use of the COPDGene dataset in CT fissure detection and segmentation (Gerard *et al.*, 2021); however, large-scale DL investigations are yet to be conducted for ^{1}H-MRI lung segmentation. Consequently, there is a pressing need for a multi-centre implementable approach to ^{1}H-MRI segmentation that can be deployed regardless of specific MR imaging parameters or patient pathology.

In this study, we hypothesised that a generalizable DL-based segmentation algorithm can accurately delineate the lung cavity across a multi-centre, multi-vendor, and multi-disease ^{1}H-MRI dataset. We aimed to develop and compare ^{1}H-MRI DL segmentation networks with a conventional segmentation approach to automatically segment the lungs on ^{1}H-MRI scans.

5.3 Materials and methods

5.3.1 Patient data

All studies received ethical approval from the relevant institutional review boards with participants (or their guardians) providing informed written consent. Appropriate consent and permissions have been granted by the sponsors to utilize this data for retrospective purposes. All data were anonymised, and all investigations were conducted in accordance with the appropriate guidelines and regulations.

^{1}H-MRI scans used in this study were retrospectively collected from several research imaging studies and patients referred for clinical pulmonary MRI scans. The dataset comprised 809 ^{1}H-MRI scans from 31 healthy participants with a median age (range) of 34 (23, 76); 66% males, 34% females and 258 participants with various pulmonary pathologies with a median age (range) of 57 (6, 85); 58% males, 42% females. Scans acquired at different inflation levels, longitudinal, and intrasession reproducibility scans were included in the dataset, resulting in a larger number of 3D scans than participants. A breakdown of patient data and demographics, stratified by disease, is included in Table 5.1.

Table 5.1 Summary of patient data.

Disease	Number of subjects	Number of scans	Age[a] Median (range)	Sex[a] Frequency (%)
Asthma	17	89	50 (15, 73)	5M (29%), 12F (71%)
Post COVID-19	147	376	57 (21, 83)	97M (66%), 49F (34%)
Cystic fibrosis	26	82	18 (6, 48)	12M (46%), 14F (54%)
Healthy	31	103	34 (23, 76)	19M (66%), 10F (34%)
ILD[b]	46	83	69 (44, 83)	25M (54%), 21F (46%)
Possible airways disease	4	15	50 (46, 64)	0M (0%), 4F (100%)
Lung cancer	18	59	72 (35, 85)	11M (61%), 7F (39%)
Total	289	809	56 (6, 85)	168M (59%), 117F (41%)

[a]Patient demographic data was unavailable for four participants.

[b]Contains connective tissue disease-associated interstitial lung disease (CTD-ILD), hypersensitivity pneumonitis (HP), idiopathic pulmonary fibrosis (IPF) and drug-induced ILD (DI-ILD).

Abbreviations: M, male; F, female; ILD, interstitial lung disease.

5.3.2 ^{1}H-MRI acquisition

The dataset used in this study contained ^{1}H-MRI acquired with a range of sequences and readout parameters from three distinct centres in the United Kingdom. ^{1}H-MRI acquisition details are summarised in Table 5.2.

Spoiled-gradient echo (SPGR) and ultrashort echo-time (UTE) ^{1}H-MRI scans were collected from Centre 1 and originated from several research and clinical studies conducted between 2014 and 2022. The data was used for training and testing DL networks containing a total of 643 scans from 207 participants and included five distinct MR sequence and readout parameter configurations (see Table 5.2). These acquisitions included differences in scanner manufacturer, sequence, field strength, lung inflation level, in-plane resolution, and slice thickness.

SPGR ^{1}H-MRI scans collected from Centre 2 and Centre 3 and originated from a single clinical study conducted between 2021 and 2022. They were used for external validation with a total of 110 scans from 55 participants (Centre 2) and 54 scans from 27 participants (Centre 3) acquired 3 to 12 months after hospitalization due to COVID-19. Each participant underwent an inspiratory and expiratory scan, resulting in two scans per subject. Acquisition details are provided in Table 5.2.

5.3.3 ^{1}H-MRI segmentations

All ^{1}H-MRI scans (n=809) had corresponding, manually-edited segmentations, representing the lung parenchyma. These segmentations were used as ground-truth delineations of the lung cavity volume, exclusive of major airways. Segmentations were pooled retrospectively and were originally generated manually or using a variety of semi-automated methods (Hughes *et al.*, 2018; Woodhouse *et al.*, 2005; Horn *et al.*, 2014). Subsequently, they were manually reviewed and edited by several experienced observers (B.A.T had 10 years, H.M had 7 years, G.J.C had 6 years, P.J.C.H had 5 years, A.M.B had 5 years, H.F.C had 4 years, L.J.S had 3.5 years, and J.R.A had 3 years, of experience in editing lung segmentations) with each observer segmenting different cases within the dataset using the ITK-SNAP software (ITK-SNAP, University of Pennsylvania, PA, USA). Airways were removed down to the third generation, and care was taken to ensure that no more than two connected components were present in the segmentations, thus removing any potentially incorrect stray voxels.

Table 5.2 ^{1}H-MRI acquisition details.

	Acquisition 1	**Acquisition 2**	**Acquisition 3**	**Acquisition 4**	**Acquisition 5**	**External validation 1**	**External validation 2**
Centre:	Centre 1	Centre 1	Centre 1	Centre 1	Centre 1	Centre 2	Centre 3
Scanner:	GE HDx	Philips Ingenia	GE HDx	GE HDx	GE HDx	Siemens Skyra	Siemens Prisma
Field strength:	1.5T	3T	1.5T	1.5T	1.5T	3T	3T
Coil:	8-channel cardiac	Body	8-element cardiac	Body	Body coil	Body coil	Body coil
Sequence:	UTE	SPGR	SPGR	SPGR	SPGR	SPGR	SPGR
Sequence dimension:	3D	3D	3D	3D	3D	3D	3D
Acquisition orientation:	Axial	Coronal	Coronal	Coronal	Coronal	Coronal	Coronal
Inflation level:	FRC (free-breathing gated on expiration)	INSP / EXP	RV, TLC, FRC+bag[a]	FRC+bag[a]	FRC+bag[a]	INSP / EXP	INSP / EXP
Slice thickness (mm):	~1.5	5	3 or 4	5	10	3	3
Inter-slice distance (mm):	~1.5	2.5	3 or 4	5	10	3	3
In-plane resolution (mm^2):	~1.5 x 1.5	~2 x 2	~3 x 3 or ~4 x 4	~4 x 4	~4 x 4	~3.13 x 3.13	~3.13 x 3.13
TR / TE (milliseconds):	2.8 / 0.078	1.9 / 0.6	1.8 / 0.7	1.9 / 0.6	1.9 / 0.6	1.9 / 0.7	1.9 / 0.7
Flip angle (°):	4	3	3	5	5	3	3
Field of view (cm):	~35-48	~38-40	~35-48	~35-48	~35-48	~40	~40
Bandwidth (kHz):	±125	±321.4	±166.6	±166.6	±166.6	±200.3	±200.3

[a]bag volume was titrated based on standing height and ranges from 400mL to 1L.

Abbreviations: FRC, functional residual capacity; RV, residual volume; TLC, total lung capacity; INSP, inspiratory; EXP, expiratory; SPGR, spoiled-gradient recalled echo; UTE, ultrashort echo time.

5.3.4 Convolutional neural network

The proposed networks consisted of a 2D and 3D implementation of the UNet CNN (Isensee *et al.*, 2018). All networks were trained using the medical imaging DL framework NiftyNet (0.6.0) (Gibson *et al.*, 2018b) built on top of TensorFlow (1.14) (Abadi *et al.*, 2016). To ensure an adequate comparison between the two CNNs, training was performed on an NVIDIA Tesla V100 graphical processing unit (GPU) (Nvidia Corporation, Santa Clara, CA, USA) with 16 GB of RAM for the same length of time, thereby normalising the performance in terms of computational efficiency and resources. Each network was trained for 120 hours.

2D UNet: A 2D UNet (Ronneberger *et al.*, 2015) architecture was used with varying kernel sizes from 3x3x3 to 1x1x1 depending on the layer of the network. An input spatial window size of 128x128x1 in the coronal plane and a volume padding size of 24x24x0 was implemented to maintain consistent image dimensions. Each network was trained with a partial rectified linear unit (PReLU) activation function (He *et al.*, 2015), Adam optimization (Kingma and Ba, 2015) and binary cross-entropy loss function. A learning rate of $1x10^{-5}$ and batch size of 1 was used for 123 training epochs. A decay of $1x10^{-6}$ and L2 regularization were implemented to minimize overfitting.

3D UNet: A 3D implementation of the UNet, referred to as the nn-UNet was used (Isensee *et al.*, 2018). Convolution operations varied in kernel size from 3x3x3 to 1x1x1 depending on the layer of the network. The network also made use of instance and batch normalization to reduce the covariate shift between network layers. An isotropic spatial window size of 96x96x96 was used. Each network was trained with a PReLU activation function (He *et al.*, 2015), Adam optimization (Kingma and Ba, 2015) and binary cross-entropy loss function. A learning rate of $1x10^{-5}$ and batch size of 2 were used for 227 training epochs. A decay of $1x10^{-6}$ and L2 regularization were selected to minimize overfitting.

Data augmentation: Data augmentation was employed before 3D scans were fed into the network to increase the variability of the training images. The augmentation method did not increase the total size of the dataset but instead used random rotation and scaling factors to modify scans before entering the network. Rotation angles of -10° to 10° and scaling values of -10% to 10% were applied for each epoch, selected based on previous research investigations (Astley *et al.*, 2022). Augmentation techniques were constrained to the above limits to produce physiologically plausible scans.

Training and testing sets: 50 scans from 50 participants, with 10 scans from each distinct acquisition in Centre 1, were randomly selected as a testing set. This constituted approximately 8% of the total number of scans from Centre 1 and 25% of the total number of participants. This was done to ensure that no participant was included concurrently in the training and testing sets and that only one scan per participant was included in the testing set. In addition, two external validation cohorts from Centres 2 and 3 were used to further validate the DL frameworks. Therefore, as a proportion of the total dataset, approximately 27% and 46% of the data in terms of scans and participants were used for testing, respectively. Numbers of scans and participants in the training, testing and external validation datasets are shown in Table 5.3.

5.3.5 Conventional approach: Spatial fuzzy c-means

A conventional approach commonly used for ^{1}H-MRI segmentation, namely, SFCM, was used (Hughes *et al.*, 2018). Images were initially bilaterally filtered to remove noise and maintain edges (Tomasi and Manduchi, 1998). SFCM differs from generic FCM algorithms in that it assumes that voxels in close spatial proximity will have a high correlation with each other and hence have similarly high membership to the same cluster. This spatial information will modify the membership value if, for instance, the voxel is noisy yet highly spatially correlated and consequently would have been incorrectly classified. The optimal number of clusters was manually selected by A.M.B based on previous experience in the clinical translation of this technique. Traditional FCM methods assign *N* pixels to *C* clusters via fuzzy memberships yet do not make use of nearby pixels during the iteration process. By taking into account the membership of voxels within a predefined window (5x5 in this work), SFCM will weigh the central voxel depending on the provided weighting variables (Li *et al.*, 2011) and thus is expected to generate more accurate segmentations (Hughes *et al.*, 2018).

Table 5.3 Breakdown of training and testing strategy with external validation

Image Acquisition		No. of scans	No. of participants
Training	Total	593	157[a]
	Acquisition 1	89	44
	Acquisition 2	78	39
	Acquisition 3	242	65
	Acquisition 4	99	26
	Acquisition 5	85	33
Testing	Total	50	50
	Acquisition 1	10	10
	Acquisition 2	10	10
	Acquisition 3	10	10
	Acquisition 4	10	10
	Acquisition 5	10	10
External validation	Total	164	82
	External validation 1	110	55
	External validation 2	54	27

[a]The number of unique participants in the training set. The totals for each acquisition in the training set are greater than this as some participants have scans from multiple acquisitions.

5.3.6 Quantitative evaluation

Segmentations generated by DL and SFCM were compared to manually-annotated segmentations and quantitatively evaluated using several common voxel-based evaluation metrics. The overlap-based Dice similarity coefficient (DSC) metric was used to assess overlap between the sets of voxels in the ground truth (X) and computationally-generated (Y) segmentations (Dice, 1945) and is defined as:

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

Average boundary Hausdorff distance $(\mathrm{Avg\ HD})$ in millimetres (Shapiro and Blaschko, 2004) assesses the conformity of boundaries between the sets of boundary voxels in X and Y defined as $\mathrm{X_B}$ and $\mathrm{Y_B}$, respectively, and is defined as follows:

$$\mathrm{Avg\ HD} = \frac{1}{2}\left(d(\mathrm{Y_B}, \mathrm{X_B}) + d(\mathrm{X_B}, \mathrm{Y_B})\right) \tag{5.2}$$

where $\mathrm{x_B}$ and $\mathrm{y_B}$ represent individual voxels in the set of $\mathrm{X_B}$ and $\mathrm{Y_B}$, respectively, and $d(\mathrm{Y_B}, \mathrm{X_B})$ and $d(\mathrm{X_B}, \mathrm{Y_B})$ represent the directed average Hausdorff distances given by:

$$d(\mathrm{Y_B}, \mathrm{X_B}) = \frac{1}{\mathrm{N}} \sum_{\mathrm{y_B} \in \mathrm{Y_B}} \min_{\mathrm{x_B} \in \mathrm{X_B}} \|\mathrm{y_B} - \mathrm{x_B}\| \tag{5.3}$$

where N is the number of observations in $|\mathrm{Y_B}|$ and $\|\mathrm{y_B} - \mathrm{x_B}\|$ is the Euclidean distance between $\mathrm{y_B}$ and $\mathrm{x_B}$.

The relative error metric (XOR) is an error-based metric which is expected to correlate with the manual editing time required to correct the segmentation (Biancardi and Wild, 2017) and is defined as follows:

$$\mathrm{XOR} = \frac{|\mathrm{Y} \cap \mathrm{X}'| + |\mathrm{Y}' \cap \mathrm{X}|}{|\mathrm{X}|} \tag{5.4}$$

where Y' and X' are the complements of Y and X, respectively.

5.3.7 Statistical analysis

The normality of the data was assessed using Shapiro-Wilk tests; if normality was not satisfied, non-parametric tests were conducted. Kruskal-Wallis tests for multiple comparisons were used to assess differences in segmentation performance between Centre 1 image acquisitions (see Table 5.2). One-way repeated-measures analysis of variance (ANOVA) with Tukey's test or Friedman tests with corrected Dunn's method for post-hoc multiple comparisons were used to assess differences in segmentation performance between the 2D UNet, 3D UNet and SFCM methods for Centre 1 data. Bland-Altman analyses were conducted to compare the 2D UNet-, 3D UNet- and SFCM-generated segmentations on external validation data. ANOVA or Friedman tests were used to assess differences between segmentation methods on external validation cohorts from Centres 2 and 3. Furthermore, independent t-tests with Welch's correction or Mann-Whitney U tests were used to assess differences between expiratory and inspiratory segmentations in external validation data. Statistical analyses were conducted using GraphPad Prism 9.2.0 (GraphPad Software, San Diego, CA). A p-value of <0.05 was considered statistically significant.

5.4 Results

5.4.1 Qualitative evaluation

Figure 5.1 shows the segmentations generated by the 2D UNet, 3D UNet and SFCM methods in comparison to the manually-edited segmentations for six cases, where a range of pulmonary pathologies, centres, and MR sequences were chosen to demonstrate each method's performance. For all cases, the 3D UNet exhibited improved performance over its 2D analogue and the SFCM method; this superior performance was maintained for the external validation dataset. Cases with challenging features such as artifacts, ground glass opacities, consolidation and bronchiectasis are displayed in Figure 5.2 along with expert, DL and SFCM segmentations. The 3D UNet exhibited improved performance on these cases compared to the other approaches tested; however, some differences were observed with expert segmentations, particularly when areas of high signal intensity were adjacent to the border of the lung cavity.

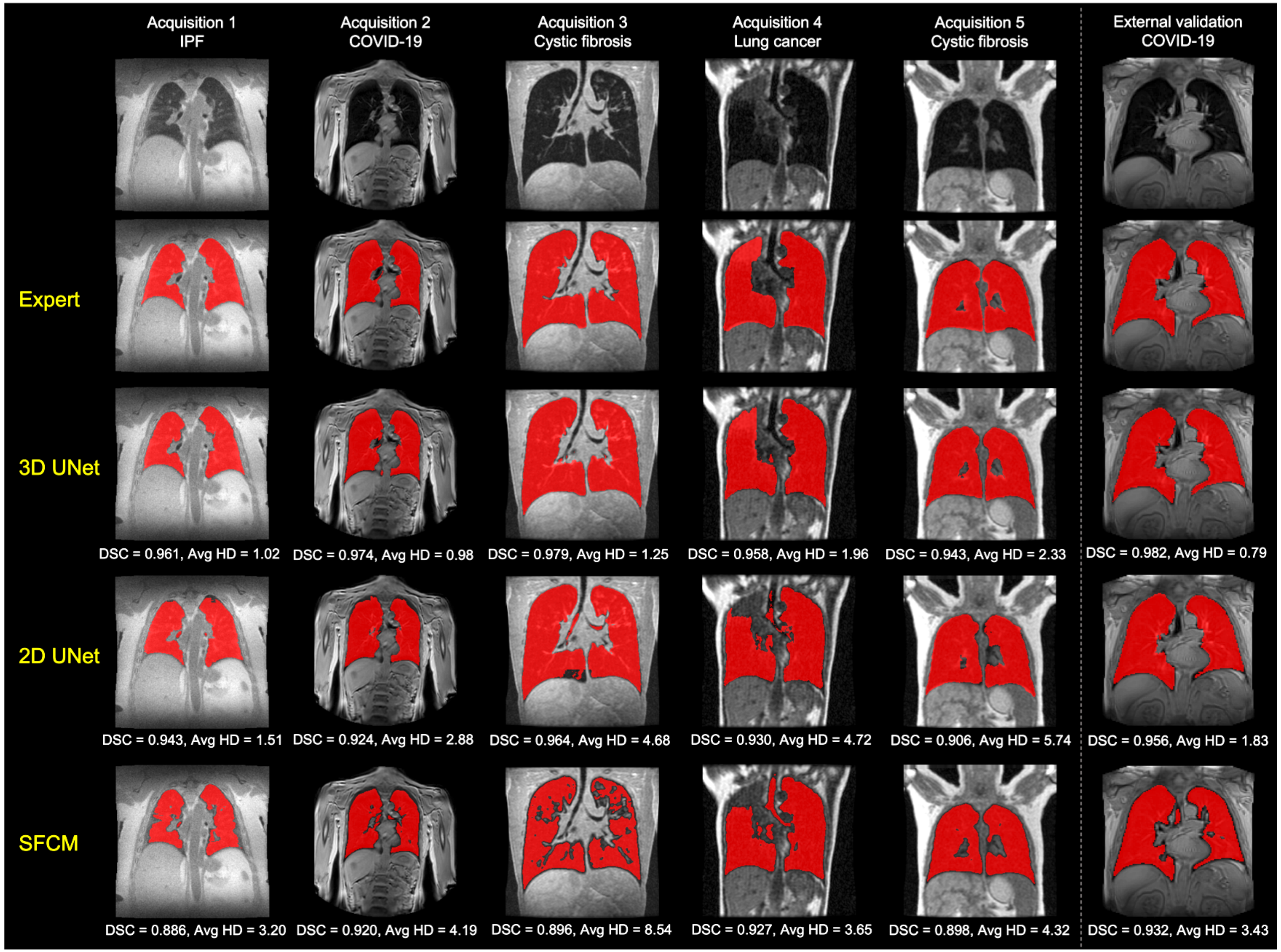

Acquisition 1
IPF
Acquisition 2
COVID-19
Acquisition 3
Cystic fibrosis
Acquisition 4
Lung cancer
Acquisition 5
Cystic fibrosis
External validation
COVID-19
Expert
3D UNet
DSC = 0.961, Avg HD = 1.02
DSC = 0.974, Avg HD = 0.98
DSC = 0.979, Avg HD = 1.25
DSC = 0.958, Avg HD = 1.96
DSC = 0.943, Avg HD = 2.33
DSC = 0.982, Avg HD = 0.79
2D UNet
DSC = 0.943, Avg HD = 1.51
DSC = 0.924, Avg HD = 2.88
DSC = 0.964, Avg HD = 4.68
DSC = 0.930, Avg HD = 4.72
DSC = 0.906, Avg HD = 5.74
DSC = 0.956, Avg HD = 1.83
SFCM
DSC = 0.886, Avg HD = 3.20
DSC = 0.920, Avg HD = 4.19
DSC = 0.896, Avg HD = 8.54
DSC = 0.927, Avg HD = 3.65
DSC = 0.898, Avg HD = 4.32
DSC = 0.932, Avg HD = 3.43

Figure 5.1 Example coronal slices showing the ^{1}H-MRI scans (row 1) and the ^{1}H-MRI scans overlaid with manual segmentations (row 2) and segmentations generated by the 3D UNet, 2D UNet and SFCM methods (rows 3-5) for six cases. DSC and Avg HD values are provided for each case. Example slices were left uncropped to display differences in field of view and arm position between acquisitions.

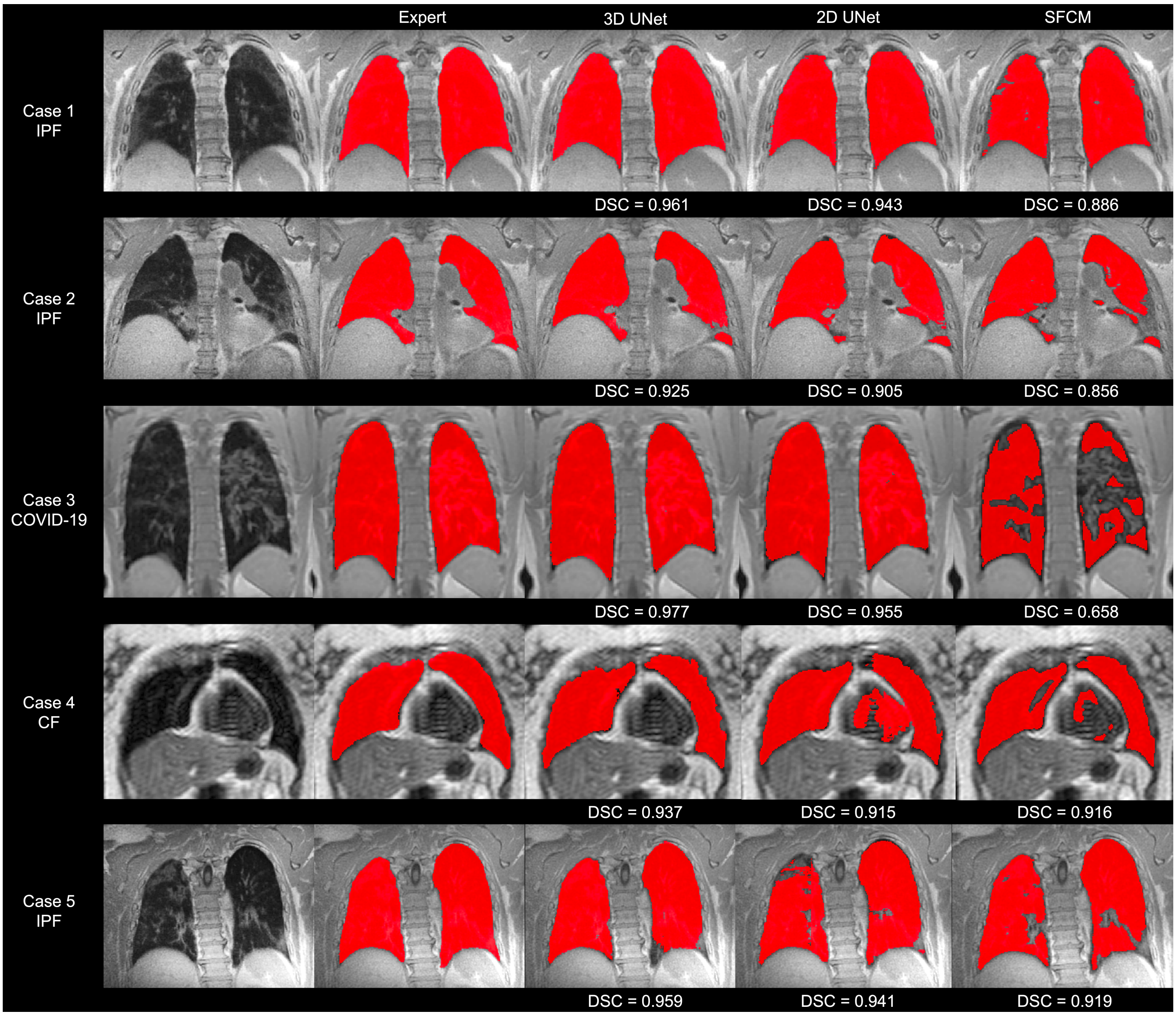
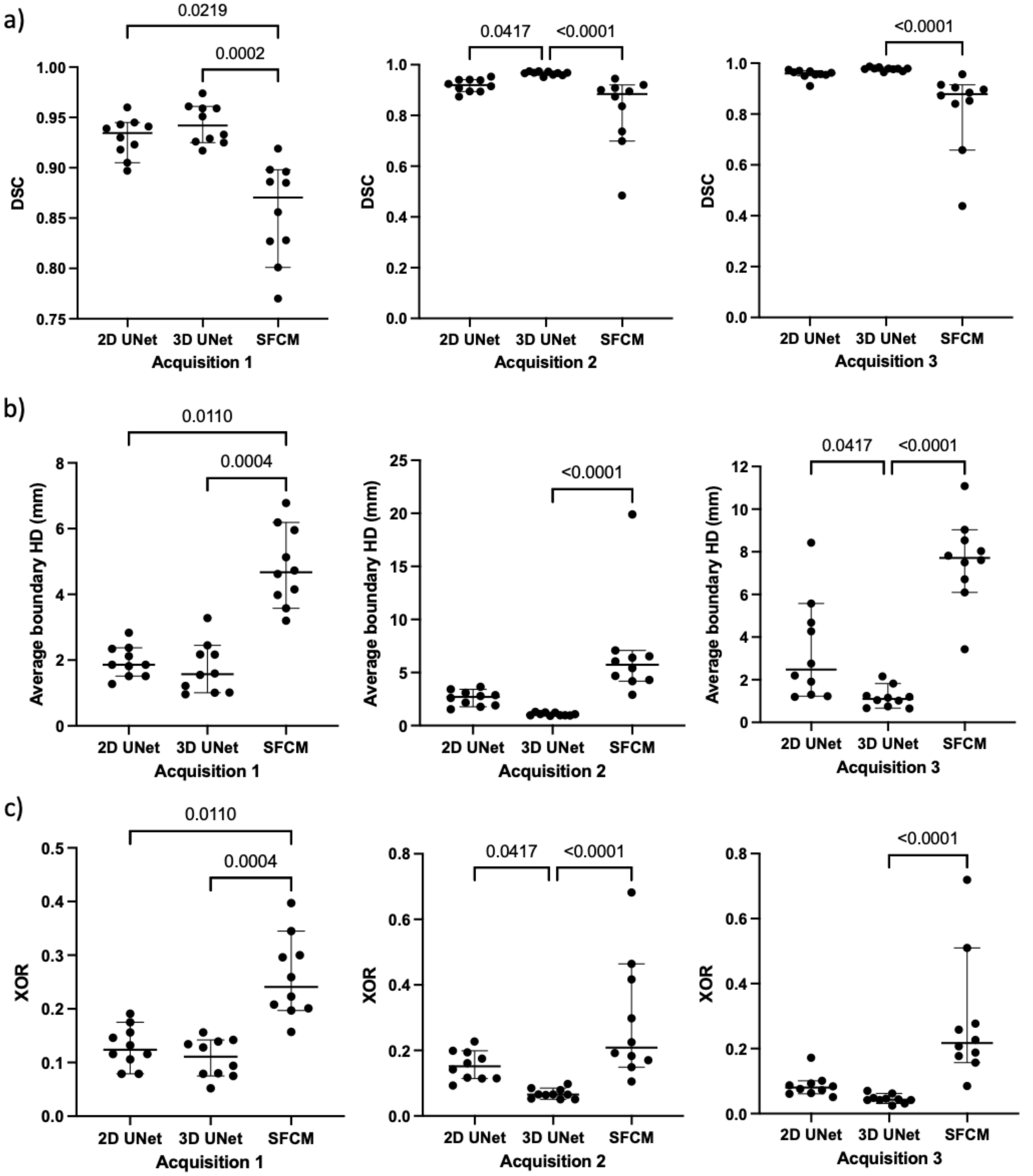


Figure 5.2 Example coronal slices showing ^{1}H-MRI scans that exhibit challenging features such as artifacts, ground glass opacities, consolidation, and bronchiectasis for five cases with corresponding expert, deep learning, and SFCM segmentations. DSC values are provided for each case and method.

5.4.2 Centre 1 evaluation

Quantitative results for the 2D UNet, 3D UNet and SFCM method are displayed in Table 5.4. Results demonstrated that the 3D UNet generated superior segmentations across all three metrics for each acquisition. The 3D UNet achieved a median (range) DSC, Avg HD and XOR of 0.961 (0.880, 0.987), 1.63mm (0.65, 5.45) and 0.079 (0.025, 0.240), respectively, on testing data from Centre 1. Both the DL-based approaches outperformed the SFCM method across all three metrics. Our 3D UNet trained model is publicly available at https://github.com/POLARIS-Sheffield/1H-MRI-segmentation. In Figure 5.3, performance between segmentation methods is shown per MR acquisition configuration for all metrics. The 3D UNet significantly outperformed the SFCM method in all comparisons and the 2D

UNet in almost all comparisons. The 2D UNet statistically outperformed the SFCM on Acquisition 1 data only. Network training performance and convergence for the 3D and 2D UNets are illustrated graphically in Figure 5.4.



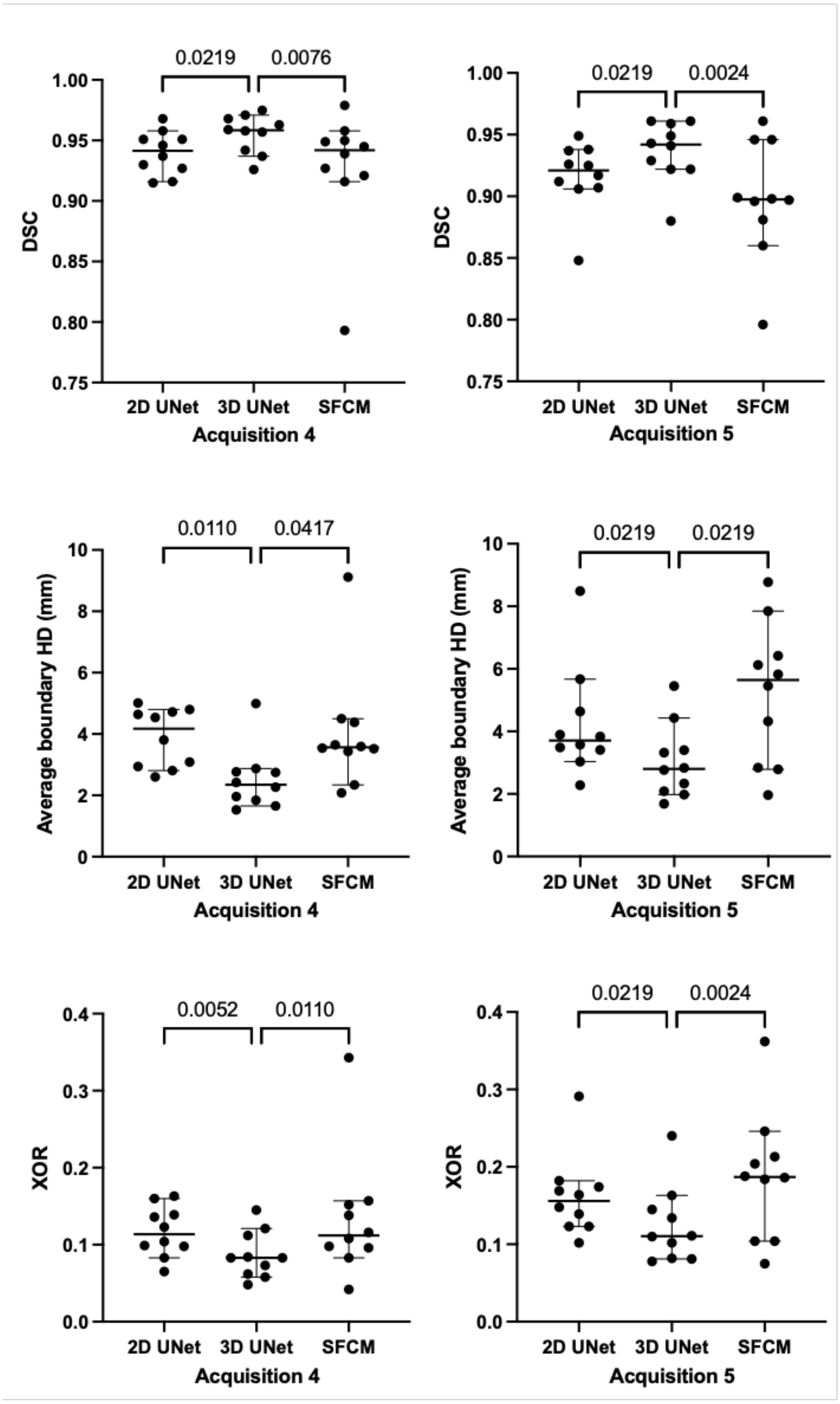
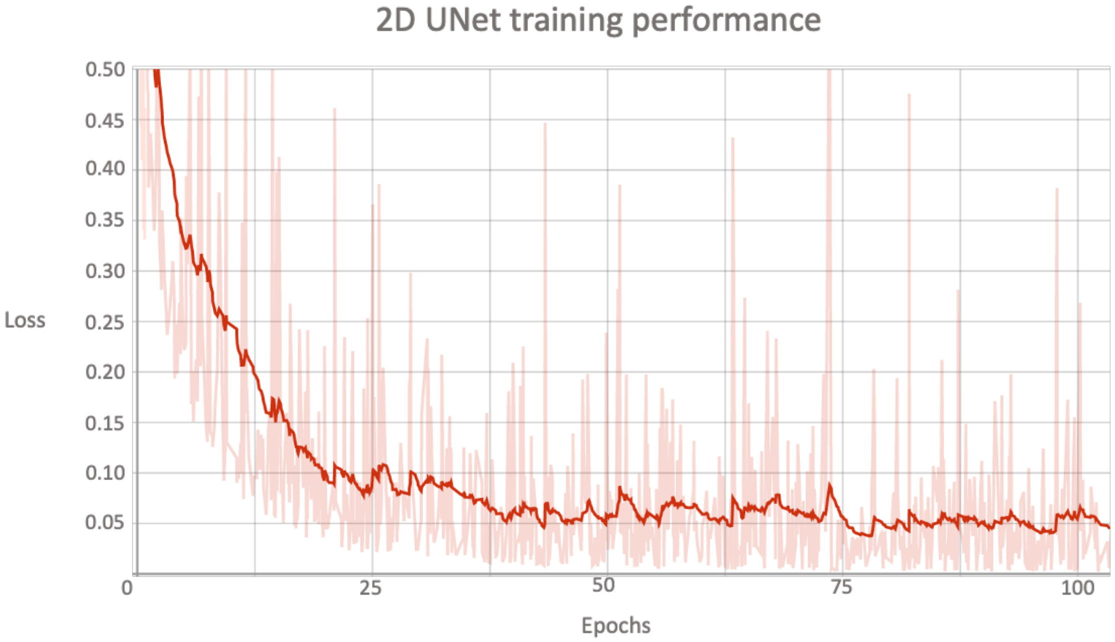


Figure 5.3 Comparison of segmentation performance for each of the methods using the a) DSC, b) Avg HD, and c) XOR metrics. Significances of differences between DL methods and SFCM as assessed by Friedman tests with Bonferroni correction are displayed for each metric.

Table 5.4 Quantitative results for the testing set (n=50), external validation 1 (n=110) and external validation 2 (n=54) using the DSC, Avg HD (mm) and XOR metrics for the SFCM, 2D UNet and 3D UNet methods. Median (range) values are provided for each acquisition protocol, the combined testing set, and the external validation sets.

Acquisition	SFCM			2D UNet			3D UNet		
	DSC	Average HD (mm)	XOR	DSC	Average HD (mm)	XOR	DSC	Average HD (mm)	XOR
	Median (range)	Median (range)	Median (range)	Median (range)	Median (range)	Median (range)	Median (range)	Median (range)	Median (range)
Acquisition 1	0.871 (0.770, 0.919)	4.67 (3.20, 6.78)	0.241 (0.157, 0.397)	0.935 (0.897, 0.960)	1.86 (1.27, 2.83)	0.124 (0.079, 0.191)	0.942 (0.917, 0.974)	1.57 (0.96, 3.28)	0.111 (0.052, 0.156)
Acquisition 2	0.885 (0.484, 0.945)	5.74 (2.90, 19.9)	0.209 (0.105, 0.682)	0.920 (0.874, 0.953)	2.70 (1.54, 3.66)	0.152 (0.093, 0.227)	0.968 (0.951, 0.974)	1.03 (0.93, 1.30)	0.065 (0.051, 0.098)
Acquisition 3	0.879 (0.438, 0.956)	7.71 (3.43, 11.1)	0.217 (0.085, 0.719)	0.960 (0.910, 0.974)	2.48 (1.20, 8.43)	0.080 (0.051, 0.172)	0.979 (0.964, 0.987)	1.10 (0.65, 2.16)	0.043 (0.025, 0.070)
Acquisition 4	0.942 (0.793, 0.979)	3.57 (2.08, 9.12)	0.112 (0.042, 0.343)	0.942 (0.915, 0.968)	4.17 (2.60, 5.01)	0.114 (0.065, 0.163)	0.959 (0.926, 0.975)	2.35 (1.53, 4.99)	0.083 (0.048, 0.145)
Acquisition 5	0.898 (0.796, 0.961)	5.64 (1.96, 8.78)	0.187 (0.075, 0.362)	0.921 (0.848, 0.949)	3.71 (2.28, 8.49)	0.156 (0.102, 0.291)	0.942 (0.880, 0.961)	2.80 (1.68, 5.45)	0.111 (0.078, 0.240)
Testing total	0.896 (0.438, 0.979)	5.28 (1.96, 19.9)	0.195 (0.042, 0.719)	0.938 (0.848, 0.974)	2.86 (1.20, 8.49)	0.123 (0.051, 0.291)	0.961 (0.880, 0.987)	1.63 (0.65, 5.45)	0.079 (0.025, 0.240)
External validation 1	0.831 (0.295, 0.949)	5.07 (2.82, 54.1)	0.290 (0.097, 0.918)	0.894 (0.477, 0.959)	4.58 (1.64, 16.7)	0.197 (0.080, 0.688)	0.973 (0.866, 0.986)	1.19 (0.53, 8.13)	0.054 (0.028, 0.255)
External validation 2	0.808 (0.170, 0.925)	5.88 (3.35, 71.9)	0.324 (0.141, 0.907)	0.902 (0.272, 0.954)	3.47 (1.79, 44.8)	0.185 (0.090, 0.912)	0.972 (0.914, 0.987)	0.96 (0.47, 3.86)	0.056 (0.026, 0.159)
External validation total	0.819 (0.170, 0.949)	5.36 (2.82, 71.9)	0.307 (0.097, 0.918)	0.894 (0.272, 0.959)	4.08 (1.64, 44.8)	0.197 (0.080, 0.912)	0.973 (0.866, 0.987)	1.11 (0.47, 8.13)	0.054 (0.026, 0.255)

Abbreviations: SFCM, spatial fuzzy c-means; DSC, Dice similarity coefficient; Average HD, average boundary Hausdorff distance; XOR, relative error metric.



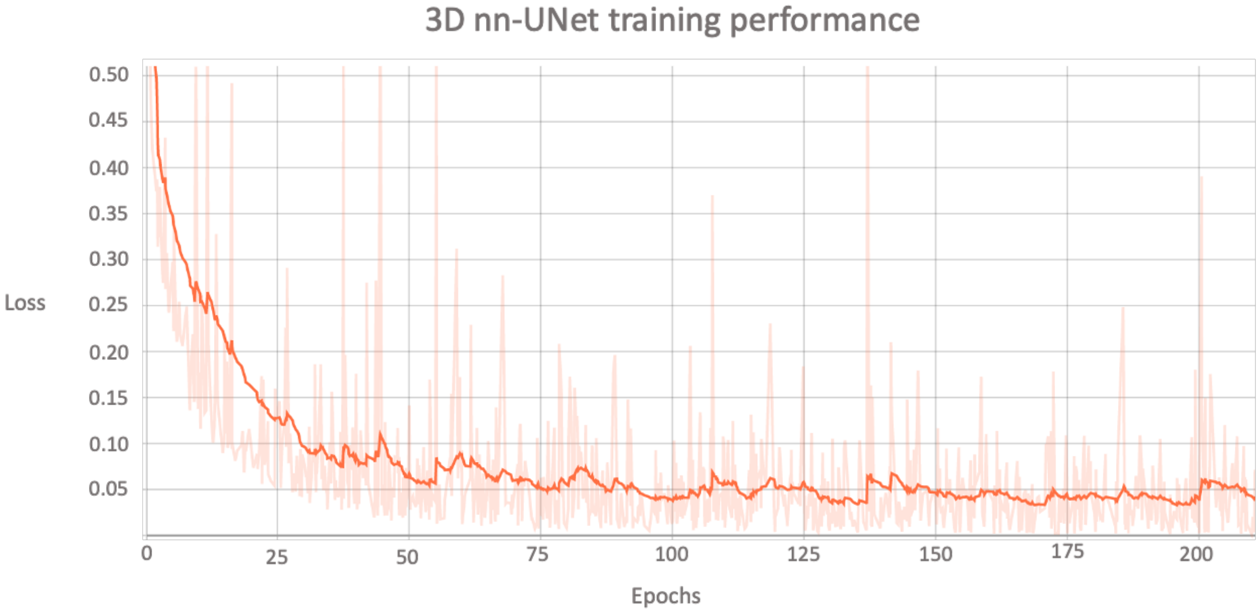
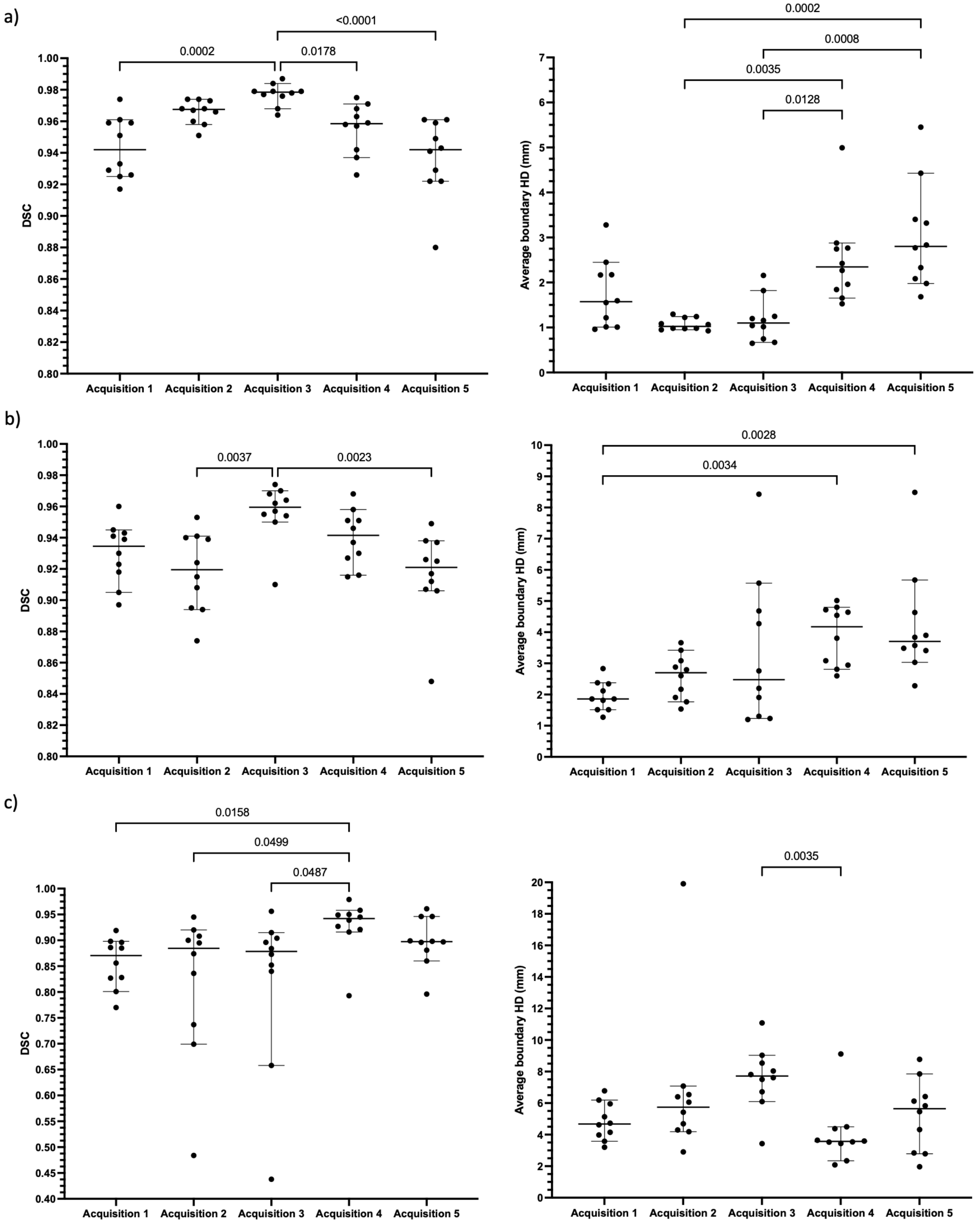


Figure 5.4 (Top) 2D UNet and (bottom) 3D nn-UNet training performance and convergence.

Figure 5.5 displays graphically the performance of the (a) 3D UNet, (b) 2D UNet and (c) SFCM methods for each metric. All methods exhibited statistically significant differences between some of the acquisitions; however, the 3D UNet exhibited the smallest range between least and best performing MR acquisition. The 3D UNet produced the most accurate segmentations for a single acquisition (Acquisition 3) when using all three metrics; in contrast, the 2D UNet and SFCM methods did not consistently exhibit superior performance for a specific acquisition across metrics.


a)
0.0002
<0.0001
0.0178
DSC
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.84
0.82
0.80
Acquisition 1
Acquisition 2
Acquisition 3
Acquisition 4
Acquisition 5
0.0002
0.0008
0.0035
0.0128
Average boundary HD (mm)
7
6
5
4
3
2
1
0
Acquisition 1
Acquisition 2
Acquisition 3
Acquisition 4
Acquisition 5
b)
0.0037
0.0023
DSC
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.84
0.82
0.80
Acquisition 1
Acquisition 2
Acquisition 3
Acquisition 4
Acquisition 5
0.0028
0.0034
Average boundary HD (mm)
10
9
8
7
6
5
4
3
2
1
0
Acquisition 1
Acquisition 2
Acquisition 3
Acquisition 4
Acquisition 5
c)
0.0158
0.0499
0.0487
DSC
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
Acquisition 1
Acquisition 2
Acquisition 3
Acquisition 4
Acquisition 5
0.0035
Average boundary HD (mm)
20
18
16
14
12
10
8
6
4
2
0
Acquisition 1
Acquisition 2
Acquisition 3
Acquisition 4
Acquisition 5

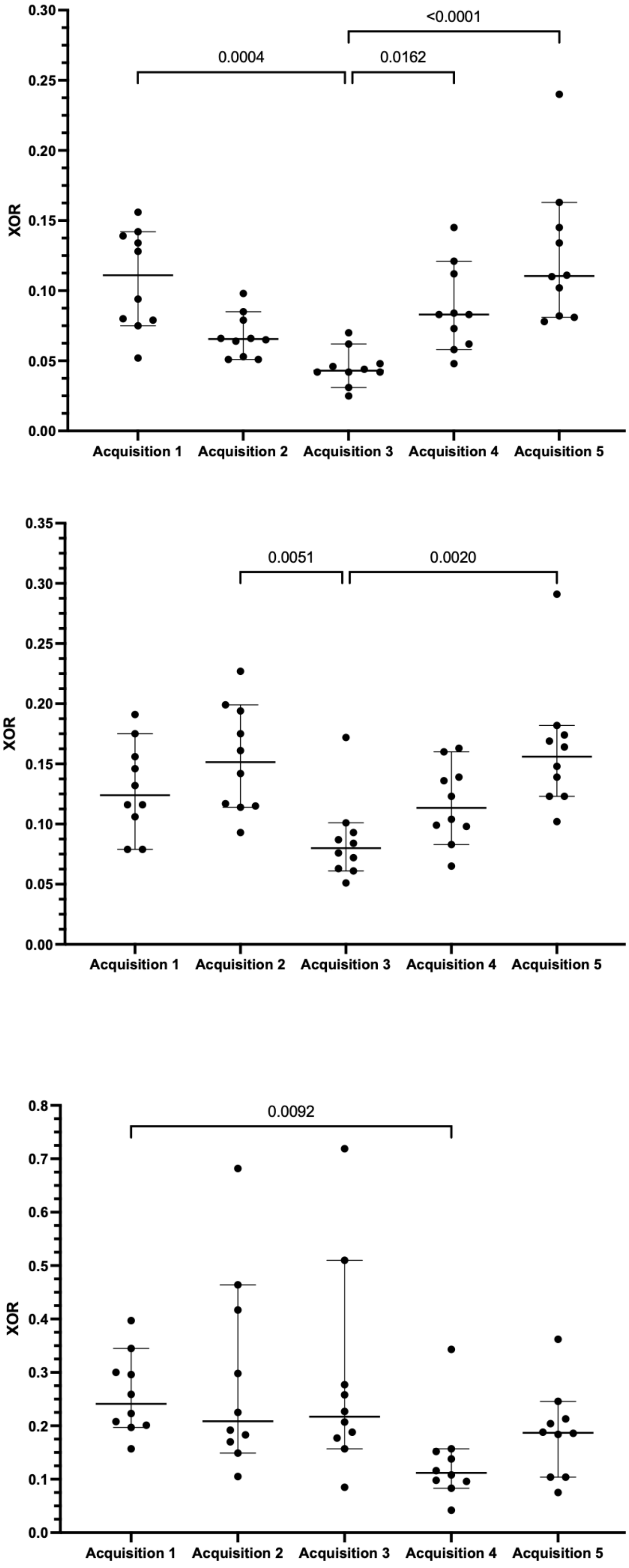


Figure 5.5 Comparison of segmentation performance across acquisition protocols for the DSC, Avg HD and XOR metrics for a) 3D UNet, b) 2D UNet and c) SFCM methods. Significant differences between image acquisitions were assessed by Kruskal-Wallis tests are given for each metric.

5.4.3 External Validation

As shown in Table 5.4, improved performance over Centre 1 testing data was exhibited on the external validation cohorts, achieving a median (range) DSC, Avg HD and XOR of 0.973 (0.866, 0.987), 1.11mm (0.47, 8.13) and 0.054 (0.026, 0.255), respectively. The 3D UNet significantly outperformed the 2D UNet and SFCM for all three metrics across 164 external validation scans using the DSC, Avg HD and XOR metrics; distribution and comparison of segmentation performance is displayed in Figure 5.6. Figure 5.7 shows Bland-Altman analyses comparing the lung parenchymal volume of DL methods and SFCM to manually-derived lung volumes for the 164 external validation scans from Centres 2 and 3. The 3D UNet exhibited a significantly reduced bias compared to other methods tested and achieved a bias of 0.063 litres with limits of agreement (LoA) -0.099 to 0.225 litres.

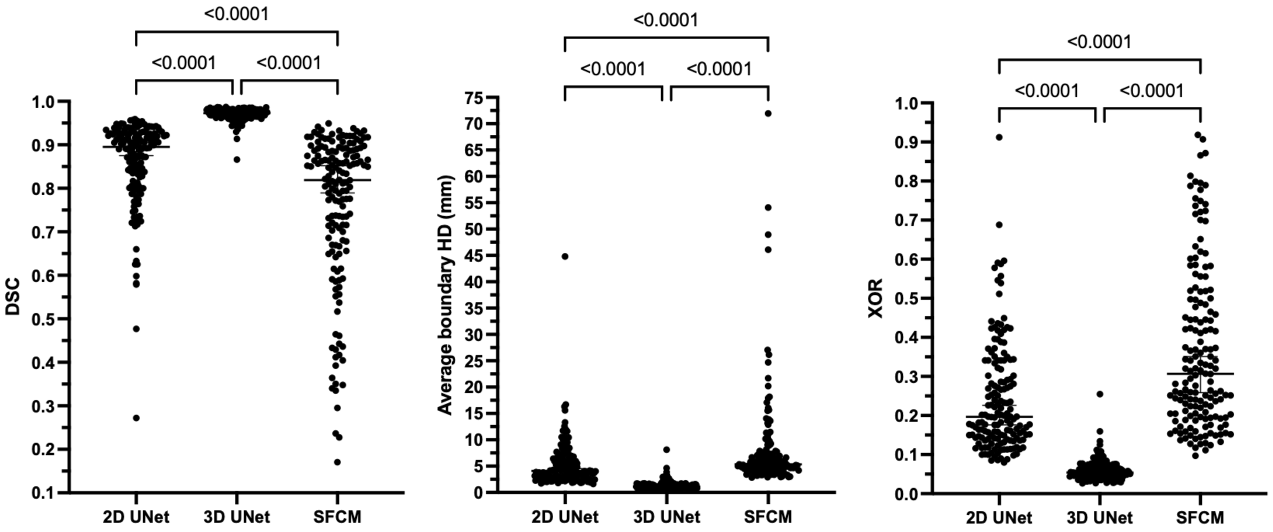


Figure 5.6 Comparison of segmentation performance on the combined external validation datasets for each of the methods using the DSC (left), Avg HD (centre) and XOR (right) metrics. Significances of differences between DL methods and SFCM as assessed by Friedman tests with Bonferroni correction for multiple comparisons are displayed for each metric.

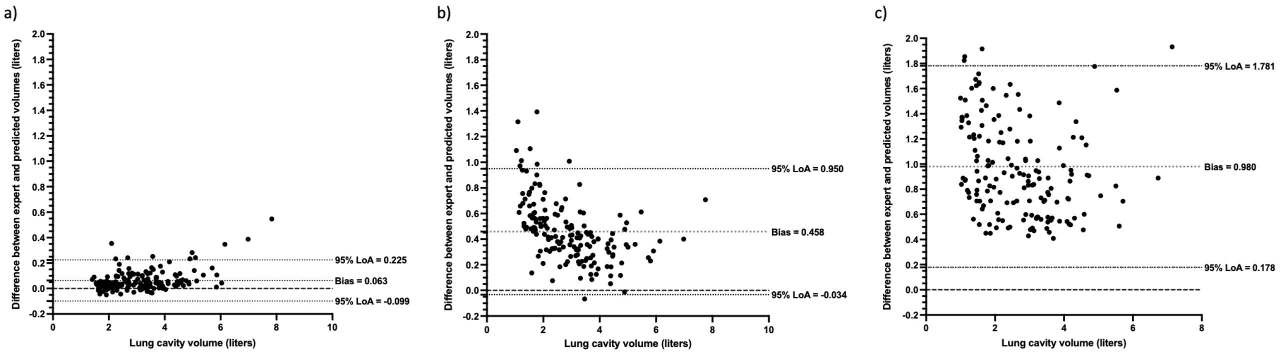


Figure 5.7 Bland-Altman agreement analysis of lung volumes for 164 external validation set cases compared to volumes derived from manual segmentations for a) 3D UNet b) 2D UNet and c) SFCM methods.

Figure 5.8 displays a comparison of segmentation performance between expiratory and inspiratory scans in data from Centres 2 and 3 for all metrics used. For the 2D UNet and the SFCM methods, inspiratory scans were segmented more accurately than expiratory scans for all metrics. This was replicated for the 3D UNet using the DSC and XOR metrics; however, no difference was observed between inspiratory and expiratory scans using the Avg HD metric (p=0.06).

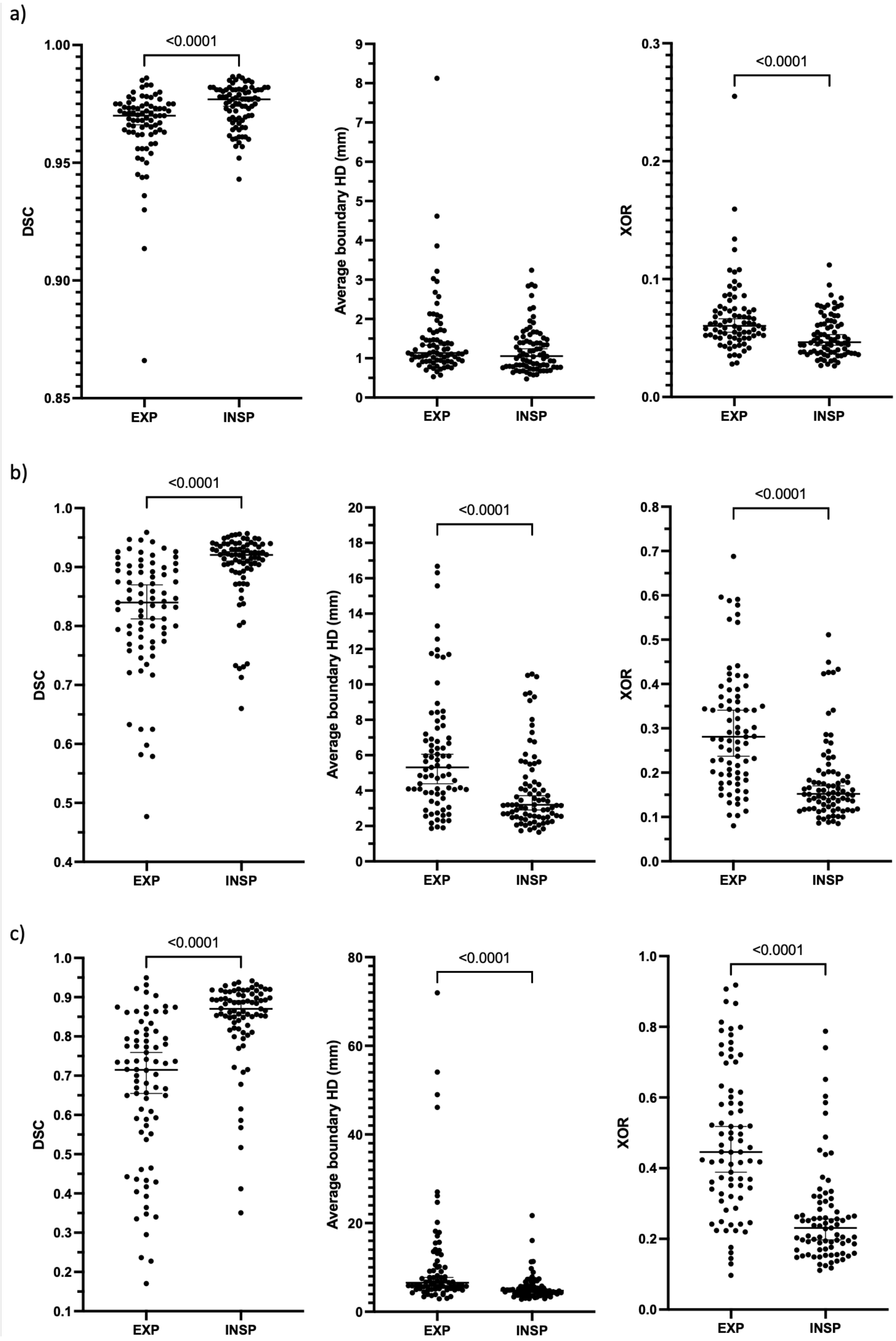


Figure 5.8 Comparison of the combined external validation datasets stratified by inspiratory and expiratory scans using the DSC, Avg HD and XOR metrics for a) 3D UNet, b) 2D UNet and c) SFCM methods. P-values between inspiratory and expiratory scans are shown.

5.5 Discussion

In this study, the proposed implementable DL segmentation algorithm produced accurate lung segmentations on a large, multi-centre, multi-acquisition, multi-disease ^{1}H-MRI dataset. Our proposed 3D CNN significantly outperformed a 2D CNN and a conventional machine learning segmentation method. In addition, it was validated on external data from two centres, acquired on different vendor scanners, demonstrating minimal bias compared to manually-edited lung volumes. Differences in lung segmentation performance were observed between scans acquired at inspiratory and expiratory inflation levels.

The dataset used is diverse in terms of pulmonary pathology, centre in which the scans were acquired, and image acquisition parameters, including sequence, field strength and vendor. This results in a segmentation network that is invariant to the specifics of the ^{1}H-MRI scans analysed, relying on relevant anatomical features present in ^{1}H-MRI scans to generate segmentations. These anatomical features remain consistent regardless of acquisition parameters in contrast to other features that varied between acquisitions, such as noise patterns, arm position, or location of the lungs within the scan. CT lung segmentation methods have adopted the large, multi-centre COPDGene dataset for validation of DL segmentation models to increase generalizability (Gerard *et al.*, 2021). In this work, we used a large multi-centre, multi-vendor ^{1}H-MRI dataset to demonstrate the generalizability of the DL model, allowing it to potentially be deployed across numerous centres; this could have a large impact on the pulmonary MRI field.

Furthermore, our proposed 3D UNet demonstrated high quality segmentations across a range of pulmonary pathologies. This performance largely extends to particularly challenging cases such as participants with idiopathic pulmonary fibrosis. Fibrotic lungs contain an increased presence of challenging pathologies, such as ground glass opacities and honeycombing, which lead to increased heterogeneity within the lung parenchyma and consequently represent challenging cases for segmentation algorithms (Mansoor *et al.*, 2015). Similarly, ^{1}H-MRI scans from participants who were previously hospitalised for COVID-19 can exhibit consolidation and reticulation patterns that reduce the difference in signal intensity between lung and non-lung tissue (Fields *et al.*, 2021), which our proposed model adequately accounts for.

Quantitative results and statistical tests indicated that, for all acquisitions, across all metrics, the 3D UNet significantly outperformed the SFCM method. For the majority of acquisitions and metrics, the 3D UNet significantly outperformed its 2D analogue. When tested on external validation data, some degree of overfitting was present in the 2D UNet exemplified by a reduction in performance compared to testing set data from Centre 1; this behaviour was not exhibited by the 3D UNet. Differences in performance between the 2D and 3D UNets are potentially due to the volumetric nature of the ^{1}H-MRI scans, which were acquired using 3D sequences. In addition, anatomical features primarily occur across multiple slices and thus a 3D approach to segmentation may better encapsulate these features. Comparison between DL networks was limited due to the differences in batch size and spatial windowing between the two CNNs as a result of differing memory constraints. It is possible that these differences may impact network comparisons; however, computational resources remained consistent between 2D and 3D CNNs and therefore the computational efficiency of the networks were assessed alongside segmentation performance.

Several investigators have leveraged CNNs for pulmonary MRI segmentation. For example, Zha *et* al. used a 2D UNet to segment the lung cavity on UTE ^{1}H-MRI scans, achieving a mean DSC of 0.96 across both lungs. However, the generalizability of this method was not demonstrated due to the small dataset of the study which only contained 45 UTE ^{1}H-MRI scans from a limited number of diseases (Zha *et al.*, 2019). Tustison *et* al. evaluated a 3D UNet CNN for isotropic ^{1}H-MRI lung cavity segmentation, achieving a mean DSC of 0.94 on a dataset of 268 scans (Tustison *et al.*, 2019). These studies employed a limited range of image acquisition parameters with ^{1}H-MRI scans acquired using the same scanner and from a single-centre. Our 3D UNet proposed here demonstrated improved performance over previous research studies on a significantly larger dataset containing scans from multiple centres with varying sequences and readout parameters. Previous works in the field of ^{1}H-MRI lung segmentation have employed either 2D (Zha *et al.*, 2019) or 3D (Tustison *et al.*, 2019) approaches; here, we directly compared differences in segmentation performance between 2D and 3D segmentation networks.

Our analysis of external validation data from Centres 2 and 3 indicated that all lung cavity segmentation methods show significantly reduced performance on scans acquired at expiration. This effect was less prevalent in segmentations generated by the 3D UNet where no significant difference between inflation levels was observed using the Avg HD metric. Differences in performance between inflation levels may be due to the reduced contrast between the lung parenchyma and other tissues as air is expelled from the lungs and the increased heterogeneity of signal within the parenchyma caused by pathophysiological air trapping at expiration observed in some patients. In addition, segmentations of exhaled lungs have a smaller volume than those of inhaled lungs; this can potentially bias quantitative results when using voxel-based evaluation metrics (Reinke *et al.*, 2021).

Accurate lung segmentation of ^{1}H-MRI plays an important role in the treatment planning, monitoring, and assessment of patients with respiratory diseases as well as other applications that require the delineation of the lung cavity such as dynamic contrast-enhanced perfusion MRI (Voskrebenzev and Vogel-Claussen, 2021). The ability to rapidly produce lung cavity segmentations can greatly reduce cumbersome manual editing, leading to a more streamlined lung imaging workflow and thus higher clinical throughput, increasing clinical translation.

5.5.1 Limitations

The ratios of MRI acquisitions present in the training set leads to potential biases towards some MR sequences or acquisitions; those with a larger number of scans may lead to improved segmentation performance for these acquisitions by the network. In particular, this study presented more Acquisition 3 scans than any other acquisition in the training set, potentially leading to the increased DSC values exhibited by the 2D and 3D UNets for this acquisition. However, using the Avg HD metric, no relationship between the number of scans in the training set and reduced segmentation performance can be established, indicating that these biases are minimal. This is further reinforced by the superior performance on external validation datasets demonstrated by the 3D UNet, despite the CNN never being exposed to ^{1}H-MRI scans from these centres or vendors during training. However, external validation data contained only one pulmonary pathology, namely, patients previously hospitalised with COVID-19.

The expert segmentations used in this work delineate only the lung parenchyma inclusive of vessels and not other relevant structures, such as the airways. Various applications require the delineation of only the lung parenchyma, including the computation of clinically relevant metrics such as the ventilation defect percentage (Woodhouse *et al.*, 2005) and as a precursor step to image registration of multi-inflation proton MRI for the generation of ^{1}H-MRI surrogates of ventilation (Capaldi *et al.*, 2018). However, in certain respiratory disorders such as obstructive sleep apnoea, the segmentation of the airways is highly relevant for studying the anatomical structure of the upper airways (Gamaleldin *et al.*, 2020). Future investigations may aim to integrate a multi-label DL solution which can segment both the lung parenchyma and airways simultaneously.

The number of MRI sequences contained within the dataset were limited. The dataset contained SPGR and UTE sequence scans i.e., proton density or T1-weighted scans only. In addition, UTE scans were acquired with a kooshball acquisition and, therefore, other possible acquisitions, such as Floret and spiral, were not assessed. Likewise, only 3D acquisition sequences were contained in the dataset, thereby limiting its implementation to 3D sequences. The inclusion of other MRI sequences, such as steady-state free-precession or fast spin echo sequences, in combination with 2D and 3D MRI sequences will help to further generalize the work. In future investigations, we will aim to further validate the model with data from an increasing number of centres and from MRI sequences not previously investigated.

In this work, ^{1}H-MRI lung segmentations were primarily evaluated using voxel-wise evaluation metrics, such as the DSC. These metrics are susceptible to reduced sensitivity in segmentation evaluation as the volume of the segmentation is increased (Taha and Hanbury, 2015). Hence, comparisons between lung inflation levels evaluated using voxel-based metrics are challenging. In future work, transfer learning could be employed to boost performance on expiratory scans or more advanced data augmentation methods could be used to increase the number of expiratory scans in the training set. Similarly, comparisons between acquisitions were limited in this study because of variations in voxel resolution, resulting in large differences in the overall number of voxels between acquisitions. Whilst the volume of the lung cavity remained largely consistent between acquisitions, the number of voxels did not; therefore, biases were introduced when using voxel-based evaluation metrics. The subject of appropriate evaluation metrics remains lively within the medical image analysis field with recent works aiming to quantify the benefits and drawbacks of each metric (Reinke *et al.*, 2021). With this in mind, in this work, we employed a range of evaluation metrics; the overlap-based DSC (Dice, 1945), the distance-based Avg HD (Shapiro and Blaschko, 2004) and the error-based XOR metric (Biancardi and Wild, 2017) which each assessed a different component of segmentation accuracy. In addition, analysis of the lung cavity volume was also undertaken when evaluating external validation data as a non-voxel-based evaluation metric to further diversify segmentation performance evaluation.

5.6 Conclusion

The DL-based implementable ^{1}H-MRI segmentation network produced accurate lung segmentations across a range of pathologies, acquisitions, vendors, and centres, which could potentially have numerous applications for pulmonary MRI quantification. A 3D CNN significantly outperformed its 2D analogue and a conventional segmentation method.

Chapter 6
A dual-channel deep learning approach for lung cavity estimation from hyperpolarised gas and proton MRI

Background: Hyperpolarised gas MRI enables quantification of regional lung ventilation via biomarkers such as the ventilation defect percentage (VDP). VDP is computed from segmentations derived from spatially co-registered functional hyperpolarised gas and structural proton (^{1}H)-MRI. Although these scans can be acquired at similar lung inflation levels, they are frequently misaligned, requiring a lung cavity estimation (LCE) mask. Recently, single channel, mono-modal deep learning (DL)-based methods have shown promise for numerous pulmonary image segmentation problems. Multi-channel approaches using multi-modal images may outperform single-channel alternatives when there are important features across multiple imaging modalities

Purpose: We hypothesise that a DL-based dual-channel approach that leverages both ^{1}H-MRI and Xenon-129-MRI (^{129}Xe-MRI) can generate LCEs more accurately than single-channel alternatives.

Study type: Retrospective.

Population: 480 corresponding ^{1}H-MRI and ^{129}Xe-MRI scans from 26 healthy participants (median age (range): 11 (8, 71); 50% males, 50% females) and 289 patients with pulmonary pathologies (median age (range): 47 (6, 83); 49% males, 51% females) were split into training (422 scans (88%); 257 participants (82%)) and testing (58 scans (12%); 58 participants (18%)) sets.

Field strength/sequence: 1.5-T, three-dimensional (3D) spoiled gradient recalled ^{1}H-MRI and 3D steady-state free-precession ^{129}Xe-MRI.

Assessment: We developed a multi-modal DL approach that integrates ^{129}Xe-MRI and ^{1}H-MRI in a dual-channel nn-UNet convolutional neural network. We compared this approach to single-channel alternatives using manually-edited LCEs as a benchmark. We further assessed a fully-automatic DL-based framework to calculate VDPs and compared it to manually generated VDPs.

Statistical tests: Shapiro-Wilk tests for normality assessment; Friedman tests with post-hoc Bonferroni correction for multiple comparisons to compare single-channel and dual-channel DL approaches using Dice similarity coefficient (DSC), average boundary Hausdorff distance (Avg HD), and relative error (XOR) metrics. Bland-Altman analysis and paired t-tests to compare manually and DL-generated VDPs. A p-value <0.05 was considered statistically significant.

Results: The dual-channel approach significantly outperformed single-channel approaches and generated realistic LCEs across numerous pulmonary pathologies achieving a median (range) DSC, Avg HD, and XOR of 0.967 (0.867–0.978), 1.68 mm (37.0–0.778), and 0.066 (0.246–0.045), respectively. Furthermore, the DL-generated VDP values were statistically indistinguishable from manually generated VDP values (p=0.710).

Data conclusion: We used a dual-channel DL approach that may allow to generate LCEs, which could be integrated with ventilated lung segmentations to produce markers such as the VDP without manual intervention.

6.1 Preface

Work contained within this chapter has been published as a Journal article in Journal of Magnetic Resonance Imaging:

> **Astley, J.R.,** Biancardi, A.M., Marshall, H., Hughes, P.J.C., Collier, G.J., Smith, L.J., Eaden, J.A., Hughes, R., Wild, J.M. and Tahir, B.A. (2023), A Dual-Channel Deep Learning Approach for Lung Cavity Estimation From Hyperpolarized Gas and Proton MRI. *Journal of Magnetic Resonance Imaging*. https://doi.org/10.1002/jmri.28519

The work contained within this chapter has also been published as conference proceedings at the following conferences:

> **Astley J.R.**, Biancardi A.M., Marshall H., Hughes P.J.C., Collier G.J., Smith L.J., Eaden J.A., Wild J.M. and Tahir B.A. (2022). A multi-channel deep learning approach for lung cavity estimation using hyperpolarized gas and proton MRI. Medical imaging and deep learning (MIDL) 2022. *Zurich, Switzerland.*

> **Astley J.R.**, Biancardi A.M., Marshall H., Hughes P.J.C., Collier G.J., Smith L.J., Eaden J.A., Blè F.X, Hughes R., Wild J.M. and Tahir B.A. (2022). A multi-channel deep learning approach for lung cavity estimation from hyperpolarized gas and proton MRI. The international society for magnetic resonance in medicine (ISMRM) 2022. *London, UK.*

Additional material that could not be included within the journal article or within conference proceedings is also contained within this chapter.

6.1.1 Author contributions

J.R.A., J.M.W. and B.A.T. made substantial contributions to the conceptualisation of the work. A.M.B., H.M., P.J.C.H., G.J.C., L.J.S., J.A.E., R.H., J.M.W. and B.A.T. were involved with patient recruitment, image acquisition and/or analysis. J.R.A. developed the dual-channel approach, performed the deep learning experiments, interpreted data, and conducted statistical analyses. J.R.A. drafted the manuscript. B.A.T. substantively revised the manuscript. All authors reviewed and approved the submitted manuscript.

6.2 Introduction

Respiratory diseases are among the leading causes of mortality and disability worldwide (Vos *et al.*, 2017; GBD15 *et al.*, 2016). Imaging plays an important role in the diagnosis, treatment planning, monitoring, and treatment assessment of respiratory diseases (Hollings and Shaw, 2002; Antony *et al.*, 2007; Kaireit *et al.*, 2017; Martini and Frauenfelder, 2018). Computed tomography (CT) is the reference standard in clinical practice for most patients with respiratory diseases (Eichinger *et al.*, 2010). Recent advances in proton MRI (^{1}H-MRI) have overcome historical challenges in using this modality for pulmonary imaging, including the low proton density and many air-tissue interfaces in the lungs (Wild *et al.*, 2012). Despite the strengths of both these modalities, they only provide structural information and not information on regional lung function. Hyperpolarised gas MRI has shown applicability for functional lung imaging including lung ventilation quantification (Woodhouse *et al.*, 2005), treatment response assessment (Horn *et al.*, 2017), and for functional lung avoidance radiotherapy (Tahir *et al.*, 2017; Ireland *et al.*, 2016). Hyperpolarised gas MRI enables quantification of regional lung ventilation with high spatial and temporal resolution (Fain *et al.*, 2007), allowing the computation of clinical biomarkers such as the ventilation defect percentage (VDP) (Woodhouse *et al.*, 2005; Hughes *et al.*, 2019).

The VDP is computed from segmentations derived from spatially co-registered, hyperpolarised gas MRI and structural ^{1}H-MRI (Stewart *et al.*, 2021). To ensure spatial alignment, both modalities are acquired consecutively and at approximately the same lung inflation level. However, the acquired scans are frequently misaligned, given that image registration, which assumes topology preservation between fixed and moving images, consistently underperforms in cases with large discrepancies in topology between functional and structural modalities (Tahir *et al.*, 2014). Consequently, the misaligned structural region of interest (the lung cavity) required for the computation of VDP poses considerable segmentation challenges. To ensure the most accurate results, particularly in cases with substantial discrepancies in inflation levels during image acquisition, a lung cavity estimation (LCE) representing the thoracic cavity volume in the spatial domain of hyperpolarised gas MRI is required. To date, no algorithm exists to automatically segment this structure and manual editing is time-consuming.

Deep learning (DL) has shown promise for numerous pulmonary image segmentation problems (Gerard *et al.*, 2020). A recent review of DL applications in lung image analysis showed that the vast majority of DL lung segmentation studies employed CT (Astley *et al.*, 2020b). The authors identified that MRI is underrepresented in DL lung segmentation applications and thus represents a gap in the literature. In the field of DL, convolutional neural networks (CNNs) have become dominant for lung image segmentation due to their ability to accurately segment various structures with computational efficiency (Astley *et al.*, 2020b). Several investigators have evaluated the use of CNNs for pulmonary MRI segmentations (Tustison *et al.*, 2019; Zha *et al.*, 2019; Jiang *et al.*, 2018; Jiang *et al.*, 2019; Sandkühler *et al.*, 2019). Tustison et al. used a three-dimensional (3D) UNet CNN to produce ^{1}H-MRI whole-lung segmentations, achieving a mean Dice similarity coefficient (DSC) of 0.94 (Tustison *et al.*, 2019). Zha et al. used a two-dimensional (2D) UNet to successfully segment ultra-short echo time (UTE) ^{1}H-MRI scans; however, this work used a relatively limited dataset, containing only 45 participants (Zha *et al.*, 2019). Astley et al. have demonstrated accurate ^{1}H-MRI segmentation on a large dataset, containing multi-resolution scans of patients with various pulmonary pathologies (Astley *et al.*, 2021). A 3D UNet was employed and achieved a mean DSC of 0.96 for whole-lung segmentation across all resolutions (Astley *et al.*, 2021). All these approaches to generate whole-lung segmentations from ^{1}H-MRI have used single-channel, mono-modal CNN-based methods, where a single image or 3D scan is used as an input to the CNN (Tustison *et al.*, 2019; Zha *et al.*, 2019; Astley *et al.*, 2021). Although these methods have shown promising results, they cannot account for the aforementioned spatial misalignments between structural and functional modalities. Multi-channel approaches using multi-modal images have shown promise in DL image analysis applications, where there are important features across multiple imaging modalities (Xu, 2019; Yang *et al.*, 2019). For example, DL has been employed for lesion segmentation using multi-modal CT and positron emission tomography (PET) images that are acquired simultaneously (Guo *et al.*, 2019). A similar problem is encountered in this work, thus motivating the investigation of dual-channel, dual-modal approaches.

We hypothesise that a dual-channel approach that leverages both ^{1}H-MRI and Xenon-129-MRI (^{129}Xe-MRI) can generate accurate LCEs across a wide range of lung pathologies. We aimed to compare this approach with single-channel CNN-based methods which do not integrate functional and structural imaging as inputs to a CNN. In addition, we aim to combine the dual-channel approach with a previously developed DL method for hyperpolarised gas MRI ventilated lung segmentation to generate clinical biomarkers, such as the VDP, without manual intervention.

6.3 Materials and methods

All prospective studies received ethical approval by the national research ethics committee with participants (or their guardians) providing informed written consent. Appropriate consent and permissions have been granted by the Sponsors to utilise this data for retrospective purposes.

6.3.1 Patient data

The dataset included in this study contained 480 corresponding ^{1}H-MRI and ^{129}Xe-MRI scans from 26 healthy participants (median age (range): 11 (8, 71); 50% males, 50% females) and 289 patients with various pulmonary pathologies (median age (range): 47 (6, 83); 49% males, 51% females). An overview of all participants, stratified by pathology, is displayed in Table 6.1. The data used in this study were pooled retrospectively from a range of prospective clinical imaging studies.

Table 6.1 Summary of patient demographic data.

Disease	Number of participants	Number of scans	Age* Median (range)	Sex* Frequency (%)	VDP* Median (range)
Asthma	92	154	50 (13, 74)	36M (40%), 55F (60%)	2.5 (0.07, 30.9)
Asthma + COPD	25	27	59 (33, 71)	15M (60%), 10F (40%)	10.4 (1.3, 29.3)
COPD	20	22	66 (48, 80)	8M (40%), 12F (60%)	18.8 (1.9, 64.8)
Cystic fibrosis	55	109	18 (6, 62)	27M (51%), 26F (49%)	6.1 (0.38, 62.0)
Healthy	26	27	11 (8, 71)	13M (50%), 13F (50%)	0.17 (0.01, 1.6)
ILD**	40	71	67 (39, 83)	21M (58%), 15F (42%)	7.9 (1.5, 30.1)
Investigation for possible airways disease	15	27	49 (11, 69)	2M (13%), 13F (87%)	6.6 (0.65, 35.0)
Preterm birth	42	43	12 (9, 14)	14M (34%), 27F (66%)	0.48 (0.01, 5.2)
Total	315	480	44 (6, 83)	138M (45%), 169F (55%)	3.6 (0.01, 62.0)

*Demographic information unavailable for eight patients. Age and VDP given at baseline.

**Contains connective tissue disease-associated interstitial lung disease (CTD-ILD), hypersensitivity pneumonitis (HP), idiopathic pulmonary fibrosis (IPF) and drug-induced ILD (DI-ILD).

Abbreviations: chronic obstructive pulmonary disease (COPD), interstitial lung disease (ILD), ventilated defect percentage (VDP), male (M), female (F).

6.3.2 Image acquisition

All participants underwent 3D volumetric ^{129}Xe-MRI and ^{1}H-MRI in the coronal plane at approximately functional residual capacity (FRC)+bag (for any given participant, the bag volume was titrated based on standing height and ranges from 400mL to 1L) or total lung capacity (TLC) with full lung coverage at 1.5T on a HDx scanner (GE Healthcare, Milwaukee, WI, USA).

^{129}Xe-MRI acquisition: The ^{129}Xe was polarised on site to approximately 25% by using an in-house developed rubidium spin-exchange polariser (Norquay *et al.*, 2018). Flexible quadrature radiofrequency coils were employed for transmission and reception of MR signals at the Larmor frequency of ^{129}Xe-MRI (Clinical MR Solutions, Brookfield, WI, USA). A 3D balanced steady-state free precession sequence was used (Stewart *et al.*, 2018). The protocol used the following settings: repetition time/echo time of 6.7/2.2 milliseconds, in-plane resolution of ~4x4 mm^2 with a slice thickness of 10 mm. A ~40 cm field of view with a flip angle of 9° or 10° at a bandwidth of ±8kHz was used.

^{1}H-MRI acquisition: The ^{1}H-MRI scans were acquired with a quadrature transmit–receive body coil in the coronal plane (Stewart *et al.*, 2018). A 3D spoiled gradient-recalled sequence was used with the following settings: repetition time/echo time of 1.9/0.6 milliseconds, in-plane resolution ~4x4 mm^2 with a slice thickness of 5 mm. A ~40 cm field of view with a flip angle of 5° at a bandwidth of ±83.3kHz was used. ^{1}H-MRI scans were acquired before and after ^{129}Xe-MRI scans at a similar lung inflation level (i.e., FRC+bag or TLC) and subsequently rigidly registered and resampled to the resolution of ^{129}Xe-MRI, using the ANTs framework implemented in an in-house MATLAB (Mathworks, Nantucket, MA, USA) software (Avants *et al.*, 2014).

6.3.3 Image quality assessment

Testing set scans were classified as either containing, or not containing, an artifact for both the ^{1}H-MRI and ^{129}Xe-MRI scans. An image was classified as containing an artifact if and only if the artifact was inside the lung parenchyma or within the region encompassed by the ribs. This was chosen to focus solely on artifacts that were likely to have a significant impact on DL-based LCE performance. Artifacts were determined by three readers; B.A.T and G.J.C have 10 years and J.R.A has 2 years of experience. B.A.T and G.J.C are both imaging scientists with extensive experience in the pulmonary MRI field and J.R.A is currently undertaking a Ph.D. in lung imaging. Each reader was blinded, and ^{129}Xe-MRI and ^{1}H-MRI scans were assessed over two sessions. ^{1}H-MRI scans were assessed for artifacts initially followed by ^{129}Xe-MRI scans with a washout period of 24 hours for J.R.A and G.J.C; B.A.T performed the analysis similarly but with a 4-hour washout period between sessions. Scans would be classified as containing an artifact if the majority of readers scored the scan as containing an artifact. We determined the SNR for all testing set cases in the ^{129}Xe-MRI and ^{1}H-MRI scans in order to assess the impact of noise on the performance of DL-based LCEs. SNR was calculated as follows:

$$\mathrm{SNR} = \frac{\text{Mean signal itensity}}{\text{Standard deviation of noise}} \tag{6.1}$$

^{129}Xe-MRI SNR analysis: Signal was assessed at a high signal location within the trachea and noise was delineated in two locations for each slice, one under the diaphragm and the other above the apex. It was ensured that both the noise and signal were calculated on regions not containing an artifact as to not conflate the artifact and SNR analyses of DL segmentation performance. Signal and noise were delineated on the central slice of the scan and one slice either side of the central slice, resulting in three consecutive slices being delineated for each participant. Figure 6.1 shows the central slice signal and noise delineations for nine random cases in the testing set.

^{1}H-MRI SNR analysis: Signal was assessed at a location within the shoulder muscle and noise was delineated outside of the chest cavity. It was ensured that both the noise and signal were calculated on regions not containing an artifact as to not conflate these two analyses of DL segmentation performance. Signal and noise were delineated on the central slice of the scan and one slice either side of the central slice, resulting in three consecutive slices being delineated for each participant. Figure 6.2 shows the central slice signal and noise delineations for nine random cases in the testing set.

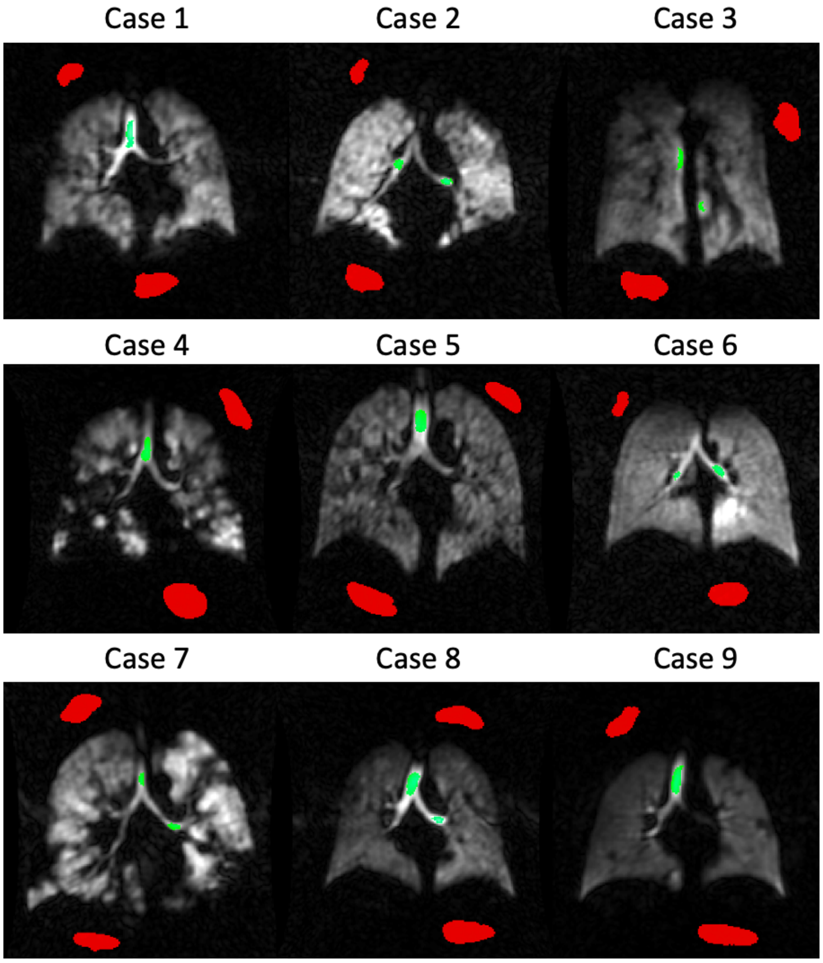


Figure 6.1 ^{129}Xe-MRI signal (green) and background noise (red) delineations.

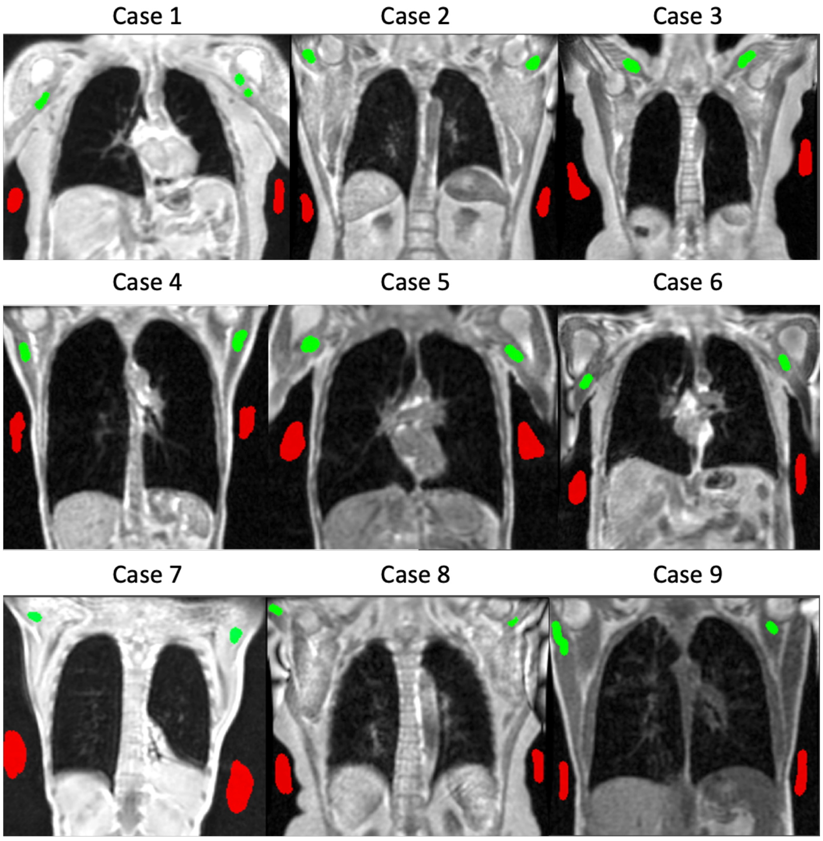


Figure 6.2 ^{1}H-MRI signal (green) and background noise (red) delineations.

6.3.4 Lung cavity estimation segmentations

Figure 6.3 displays fused ^{129}Xe-MRI and ^{1}H-MRI scans after rigid registration, demonstrating the continued misalignment between ventilation and structural scans and thus highlighting the requirement for an LCE. Segmentation of LCEs from ventilation and structural MR image pairs was conducted semi-automatically using paired spatial fuzzy c-means clustering (SFCM) (Biancardi *et al.*, 2018). Images are initially bilaterally filtered to remove noise and maintain edges (Tomasi and Manduchi, 1998). The standard FCM algorithm assigns *N* pixels to *C* clusters via fuzzy memberships with the assumption that pixels in close proximity are highly correlated and hence have similarly high membership to the same cluster (Bezdek *et al.*, 1984). This spatial information will modify the membership value only if, for example, the pixel is noisy and would have been incorrectly classified.

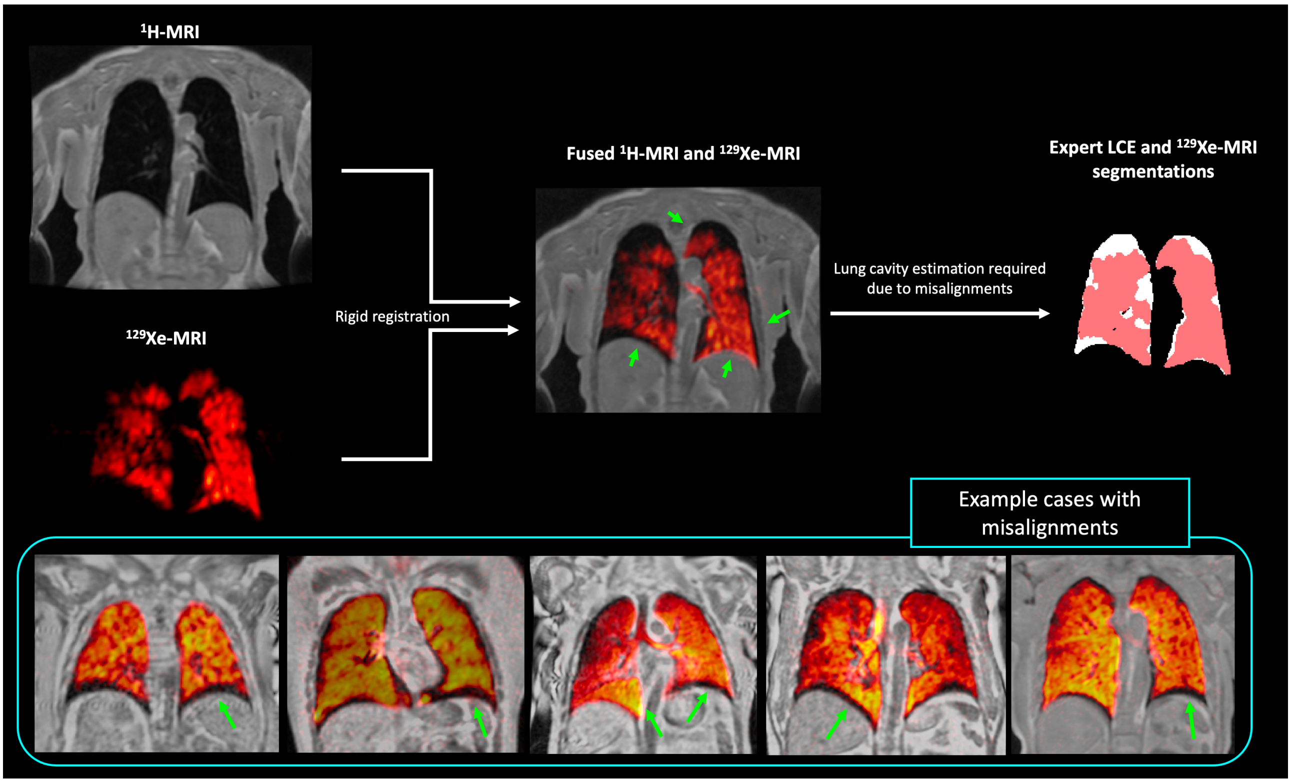

¹H-MRI
¹²⁹Xe-MRI
Rigid registration
Fused ¹H-MRI and ¹²⁹Xe-MRI
Lung cavity estimation required due to misalignments
Expert LCE and ¹²⁹Xe-MRI segmentations
Example cases with misalignments

Figure 6.3 Illustration showing the motivation to generate lung cavity estimations in the spatial domain of ^{129}Xe-MRI due to misalignments in image acquisitions between modalities. Example cases demonstrating misalignments between ^{129}Xe- and ^{1}H-MRI. Misalignments are indicated by green arrows.

The SFCM method makes use of nearby pixels during the iteration process by considering the membership of voxels within a predefined window and will weigh the central pixel depending on the provided weighting variables (Chuang *et al.*, 2006). Heuristic values for the number of clusters and cluster selection threshold for inclusion in the ventilation or structural masks were identified, resulting in the selection of 18 clusters for both masks by

A.M.B who had 3.5 years of experience. For the manual segmentations used in this work, the SFCM clustering was applied to both ^{129}Xe-MRI and ^{1}H-MRI scans in a pair-wise fashion to take advantage of the combined information arising from the co-location of the image pair (Biancardi *et al.*, 2018). LCEs were pooled retrospectively from several studies and, consequently, were subsequently manually reviewed and edited by several experienced observers, where each scan was segmented by a single observer, but the dataset as a whole contained LCEs manually-edited by observers with a range of expertise: H.M had 7 years, G.J.C had 6 years, P.J.C.H had 5 years, A.M.B had 5 years, L.J.S had 3.5 years, J.A.E had 3 years, and J.R.A had 2 years, of experience in editing LCEs**.**

6.3.5 Deep learning frameworks

We assessed three DL methods to generate LCEs by varying the input channels provided to each network. These consisted of single-channel and dual-channel CNN approaches (Figure 6.4) as follows:

1) Ventilation-only (^{129}Xe-MRI)
2) Structural-only (^{1}H-MRI)
3) Dual-channel (^{129}Xe-MRI + ^{1}H-MRI)

All methods used a variation of the common 2D UNet encoder-decoder network architecture; here we used a 3D implementation of the UNet, referred to as the nn-UNet, which has been modified to reduce memory constraints, allowing 30 feature channels (Isensee *et al.*, 2018). Convolution operations varied in kernel size from 3x3x3 to 1x1x1 depending on the layer of the network. The network also made use of instance normalisation. An isotropic spatial window size of 96x96x96 was used. Each network was trained with a parametric rectified linear unit (PReLU) activation function (He *et al.*, 2015), ADAM optimisation (Kingma and Ba, 2015), and cross-entropy loss function. A learning rate of 1x10^{-5} and batch size of 2 were used. A decay of 1x10^{-6} and L2 regularisation were selected to minimise overfitting. Each method was trained for 300 epochs resulting in a model training time of approximately 8 days. All networks were trained using the medical imaging DL framework NiftyNet 0.6.0 (https://github.com/NifTK/NiftyNet) built on top of TensorFlow 1.14 (Gibson *et al.*, 2018b). Training and inference were performed on an NVIDIA Tesla V100 graphical processing unit (GPU) (Nvidia corporation, Santa Clara, CA, USA) with 16 GB of RAM.

Data augmentation: Constrained random rotation and scaling were used for data augmentation before ^{129}Xe-MRI and ^{1}H-MRI scans were fed into the network. The augmentation method used does not increase the total size of the dataset but instead utilises random rotation and scaling factors to modify scans before entering the network. Each time a scan is fed into the network, random rotation and scaling factors with limits -10° to 10° and -10% to 10%, respectively, where different factors at an interval within these limits, were applied.

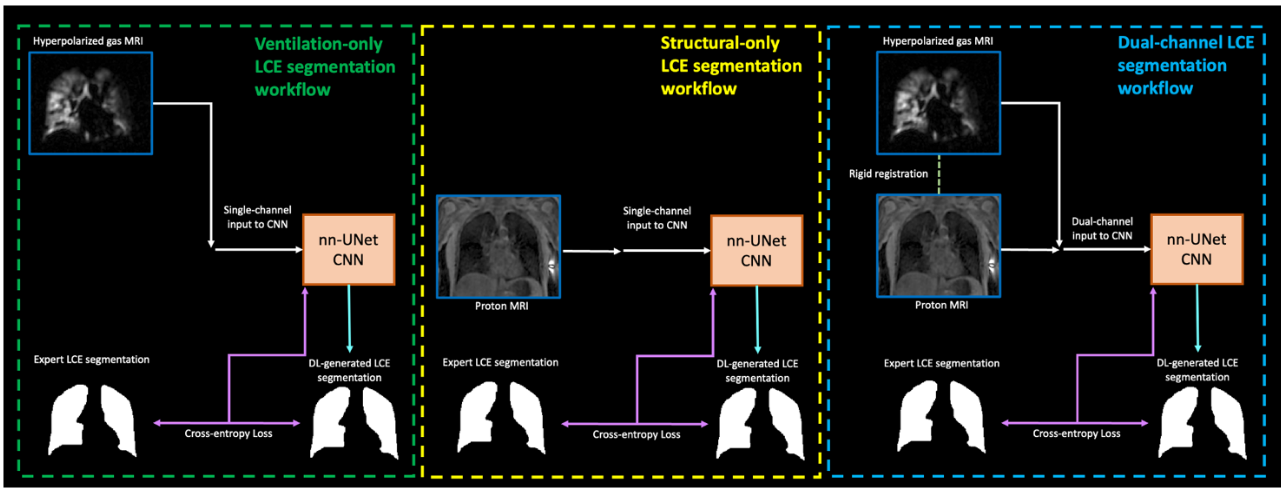

Hyperpolarized gas MRI
Ventilation-only LCE segmentation workflow
Single-channel input to CNN
nn-UNet CNN
Expert LCE segmentation
DL-generated LCE segmentation
Cross-entropy Loss
Structural-only LCE segmentation workflow
Proton MRI
Single-channel input to CNN
nn-UNet CNN
Expert LCE segmentation
DL-generated LCE segmentation
Cross-entropy Loss
Hyperpolarized gas MRI
Dual-channel LCE segmentation workflow
Rigid registration
Proton MRI
Dual-channel input to CNN
nn-UNet CNN
Expert LCE segmentation
DL-generated LCE segmentation
Cross-entropy Loss

Figure 6.4 From left to right: ventilation-only, structural-only, and dual-input deep learning workflows.

Training and testing sets: The dataset was divided into training and testing sets; the data split was conducted at the level of scans whereby 15% of the scans were randomly selected as the testing set. If a participant had multiple repeat or longitudinal scans, one scan out of these was randomly selected and the other scans discarded from the analysis; these removed scans do not appear in the dataset. This was done to ensure that no participant was present in both the training and testing sets and that the testing set contained only one scan from each participant, thereby reducing potential biases in favour of specific participants. Therefore, the training set contained 422 corresponding ^{129}Xe-MRI and ^{1}H-MRI scans from a total of 257 participants and the testing set contained 58 scans from 58 participants, representing 81.6% and 18.4% of the total number of participants, respectively. Even though the testing set allocation was randomly determined, at least one scan from each disease or healthy cohort (described in Table 6.1) was present in the testing set. The training set had the following demographic distributions: median age (range) of 41 (8.9, 83); median VDP (range) 3.23% (0.01, 64.8); sex 44% male, 56% female. The testing set had the following demographic distributions: median age (range) of 53 (6.4, 76); median VDP (range) 5.19% (0.05, 62.0); sex 49% male, 51% female.

6.3.6 Quantitative evaluation

The DL-generated LCEs were quantitatively evaluated using the overlap-based DSC metric that assesses the overlap between the sets of voxels in the ground truth (X) and computationally-generated (Y) segmentations, defined as:

$$\mathrm{DSC} = 2\frac{|\mathrm{Y} \cap \mathrm{X}|}{|\mathrm{Y}| + |\mathrm{X}|} \tag{6.2}$$

Average boundary Hausdorff distance $(\mathrm{Avg\ HD})$ in millimetres (Shapiro and Blaschko, 2004) assesses the conformity of boundaries between the sets of boundary voxels in X and Y defined as $\mathrm{X_B}$ and $\mathrm{Y_B}$, respectively, and is defined as follows:

$$\mathrm{Avg\ HD} = \frac{1}{2}\left(d(\mathrm{Y_B}, \mathrm{X_B}) + d(\mathrm{X_B}, \mathrm{Y_B})\right) \tag{6.3}$$

where $\mathrm{x_B}$ and $\mathrm{y_B}$ represent individual voxels in the set of $\mathrm{X_B}$ and $\mathrm{Y_B}$, respectively, and $d(\mathrm{Y_B}, \mathrm{X_B})$ and $d(\mathrm{X_B}, \mathrm{Y_B})$ represent the directed average Hausdorff distances given by:

$$d(\mathrm{Y_B}, \mathrm{X_B}) = \frac{1}{\mathrm{N}} \sum_{\mathrm{y_B} \in \mathrm{Y_B}} \min_{\mathrm{x_B} \in \mathrm{X_B}} \|\mathrm{y_B} - \mathrm{x_B}\| \tag{6.4}$$

where N is the number of observations in $|\mathrm{Y_B}|$ and $\|\mathrm{y_B} - \mathrm{x_B}\|$ is the Euclidean distance between $\mathrm{y_B}$ and $\mathrm{x_B}$.

A relative error metric (XOR) was used to evaluate segmentation errors as follows:

$$\mathrm{XOR} = \frac{|\mathrm{Y} \cap \mathrm{X}'| + |\mathrm{Y}' \cap \mathrm{X}|}{|\mathrm{X}|} \tag{6.5}$$

where Y' and X' are the complements of Y and X, respectively. The metric was used because it is expected to correlate with the manual editing time required to correct the segmentation outcome (Biancardi and Wild, 2017).

6.3.7 Clinical evaluation

In addition to quantitative evaluation metrics, clinical evaluation metrics were used to assess the lung parenchymal volume defined by the LCE. DL-generated LCE volumes were compared to ground truth LCE volumes to assess LCE accuracy. The VDP has been used as a robust measure of lung function (Woodhouse *et al.*, 2005). VDP was calculated from structural and functional volumes aligned via rigid registration as follows:

$$\mathrm{VDP}\ (\%) = \left(1 - \frac{\text{ventilated lung volume}}{\text{LCE volume}}\right) \times\ 100 \tag{6.6}$$

We assessed the performance of the DL-generated LCEs by computing VDP values for each scan in the testing set. As shown in Equation*(6.6*, in addition to LCE volumes, ventilated lung volumes are required. Thus, we employed a previously trained nn-UNet fully-CNN, developed for automatic hyperpolarised gas MRI ventilated lung segmentation in a large diverse dataset (Astley *et al.*, 2022), which was done to generate accurate DL-based ^{129}Xe-MRI ventilated lung segmentations for the current testing set. The fully-automatic DL-derived VDPs were compared to VDPs derived from manually-edited ventilated and LCE segmentations. Ventilated volumes were initially generated using a binning method (He *et al.*, 2014). ^{129}Xe-MRI scans were normalised by the average value of the ^{129}Xe signal in the lung cavity and ventilation defects were defined as any value below 33% of the mean signal intensity. Thus, the ventilated volume was defined as the complement of the ventilation defect (Collier *et al.*, 2018).

6.3.8 Statistical analysis

All statistical analyses were conducted using GraphPad Prism (version 9.2.0; GraphPad Software, San Diego, CA, USA). Data were tested for normality using Shapiro-Wilk tests. When normality was not satisfied, non-parametric tests were conducted. One-way repeated measures analysis of variance (ANOVA) or Friedman tests were conducted as appropriate with Bonferroni correction for post-hoc multiple comparisons to assess statistical significance of differences between DL ventilation-only, structural-only, and dual-input methods. Pearson or Spearman correlation and Bland-Altman analyses were conducted to compare the volumes of the dual-input DL method and manual LCEs. In addition, paired t-tests and Bland-Altman analyses were used to compare manual and DL-generated VDP values. Independent t-tests with Welch's correction or Mann-Whitney U tests were used as appropriate to assess differences in VDPs between scans containing or not containing artifacts. Relationships between differences in manual and DL-generated VDPs and SNRs were assessed using Pearson or Spearman correlation. A p-value <0.05 was considered statistically significant.

6.4 Results

6.4.1 Quantitative evaluation

Figure 6.5 demonstrates the qualitative and quantitative performance of each DL method comparing the DL-generated LCEs to the manual LCEs for four cases. For all cases, the dual-input method generated realistic LCEs that might accurately mimic manual LCEs.

Quantitative results for each DL method are provided in Figure 6.6a. The results demonstrate that the dual-input method generated the most accurate segmentations across all metrics used. The dual-input method achieved a median (range) DSC, Avg HD, and XOR of 0.967 (0.867, 0.978), 1.68 mm (37.0, 0.778 mm), and 0.066 (0.246, 0.045), respectively. The dual-input method significantly outperformed the single-channel methods. The results for all metrics are displayed graphically in Figure 6.6b. Due to the significant improvements demonstrated by the dual-input DL method across all segmentation metrics, we selected this method for assessment using clinical evaluation metrics.

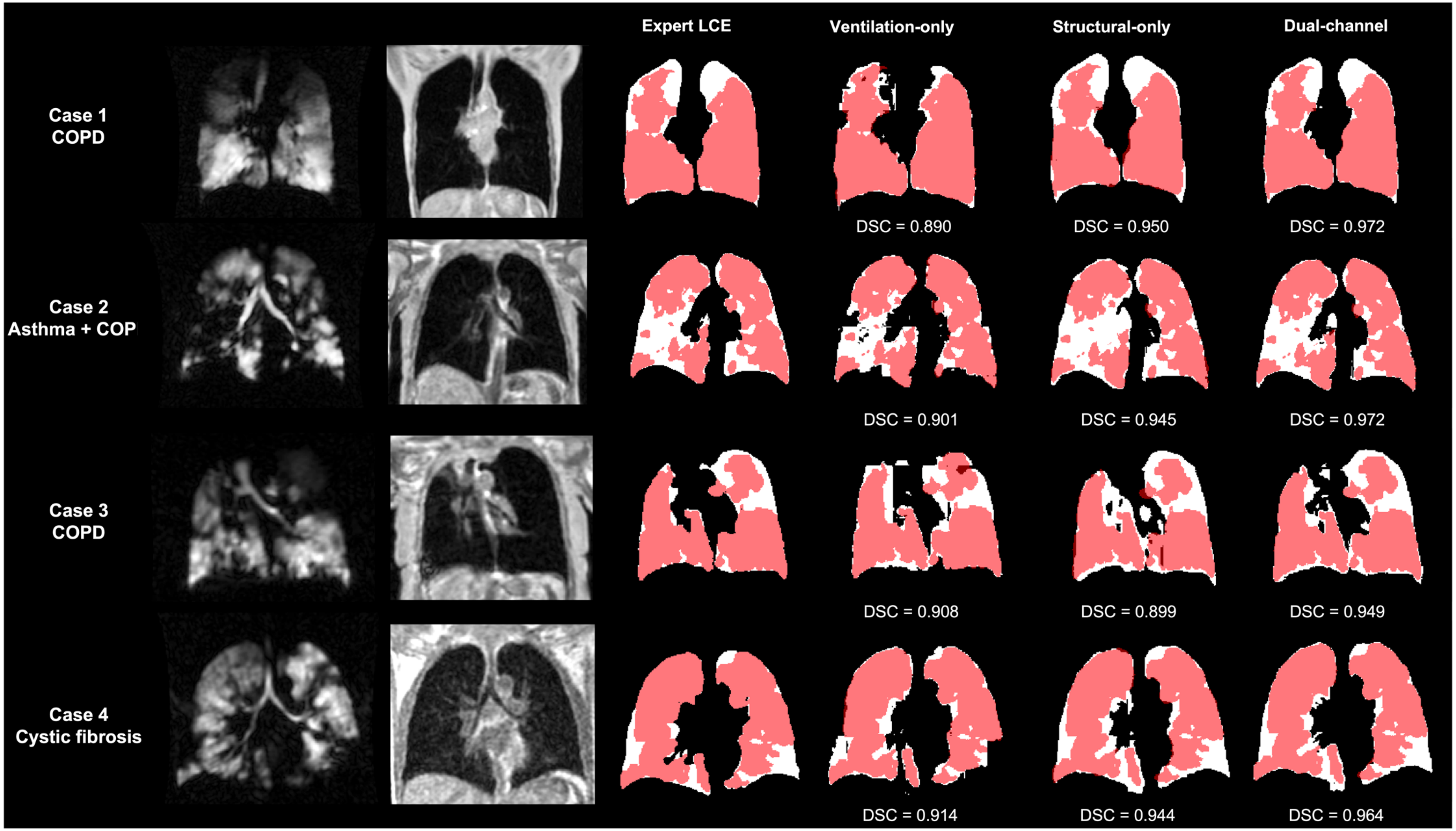


Figure 6.5 Example coronal slices showing the ^{1}H-MRI and corresponding similar-breath hyperpolarised gas MRI scan with the expert LCE and the LCE generated from the three DL methods for four cases within the testing set. Expert hyperpolarised gas MRI segmentations are provided for each case to aid visualisation of alignments. DSC values are provided for each case.

a)

LCE DL methods	DSC	Average HD (mm)	XOR
	Median (range)	Median (range)	Median (range)
Ventilation-only	0.952 (0.719, 0.979)	2.22 (0.762, 66.0)	0.095 (0.043, 0.749)
Structural-only	0.935 (0.797, 0.959)	4.19 (2.13, 11.5)	0.132 (0.082, 0.355)
Dual-channel	**0.967 (0.867, 0.978)**	**1.68 (0.778, 37.0)**	**0.066 (0.045, 0.246)**

b)

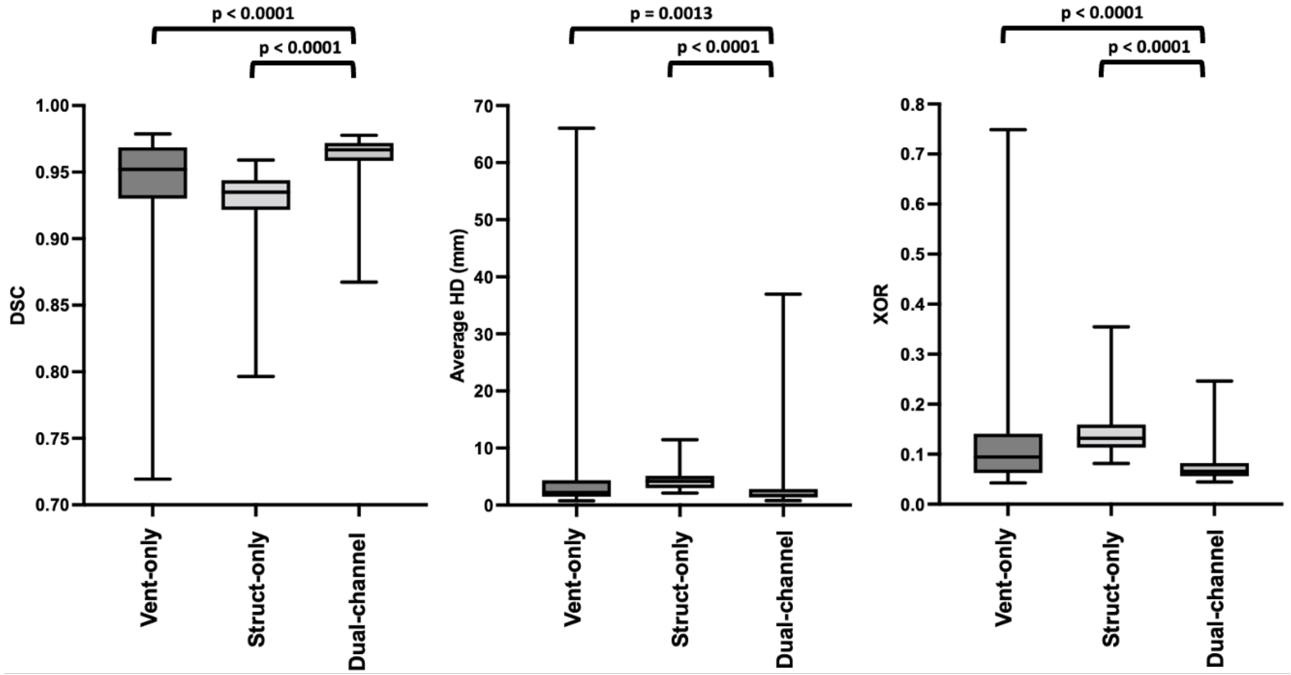


Figure 6.6 a) Quantitative results for the testing set (n=58) using the DSC, Avg HD (mm) and XOR metrics for the ventilation-only, structural-only, and dual-input DL methods. Median (range) values are given with the best values shown in bold. b) Comparison of LCE performance for each of the three DL methods using the DSC (left), Avg HD (centre) and XOR (right) metrics. Significance of differences between DL methods as assessed by Friedman tests with Bonferroni correction for multiple comparisons are displayed for each metric.

6.4.2 Clinical evaluation

Figure 6.7 shows Pearson correlation and Bland-Altman analyses of lung volumes for the dual-input, DL-generated LCEs compared to manual LCEs. The dual-input method exhibited a statistically significant, strong Pearson’s correlation of 0.98 and minimal bias of 0.06±0.26 litres with limits of agreement (LoA) of -0.45 to 0.56 litres. Figure 6.8 shows example coronal slices of the manual LCEs and ventilated lung volumes compared with those generated by the DL methods.

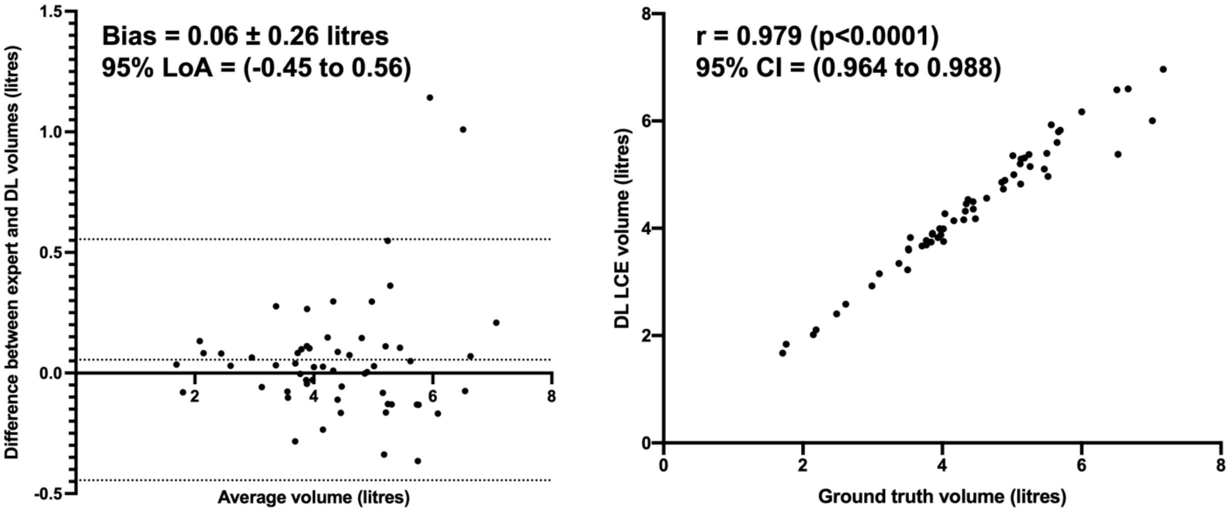


Figure 6.7 Bland-Altman analysis (left) and Pearson correlation (right) of lung volumes for 58 testing set cases comparing the manual LCEs to the dual-input DL-generated LCEs.

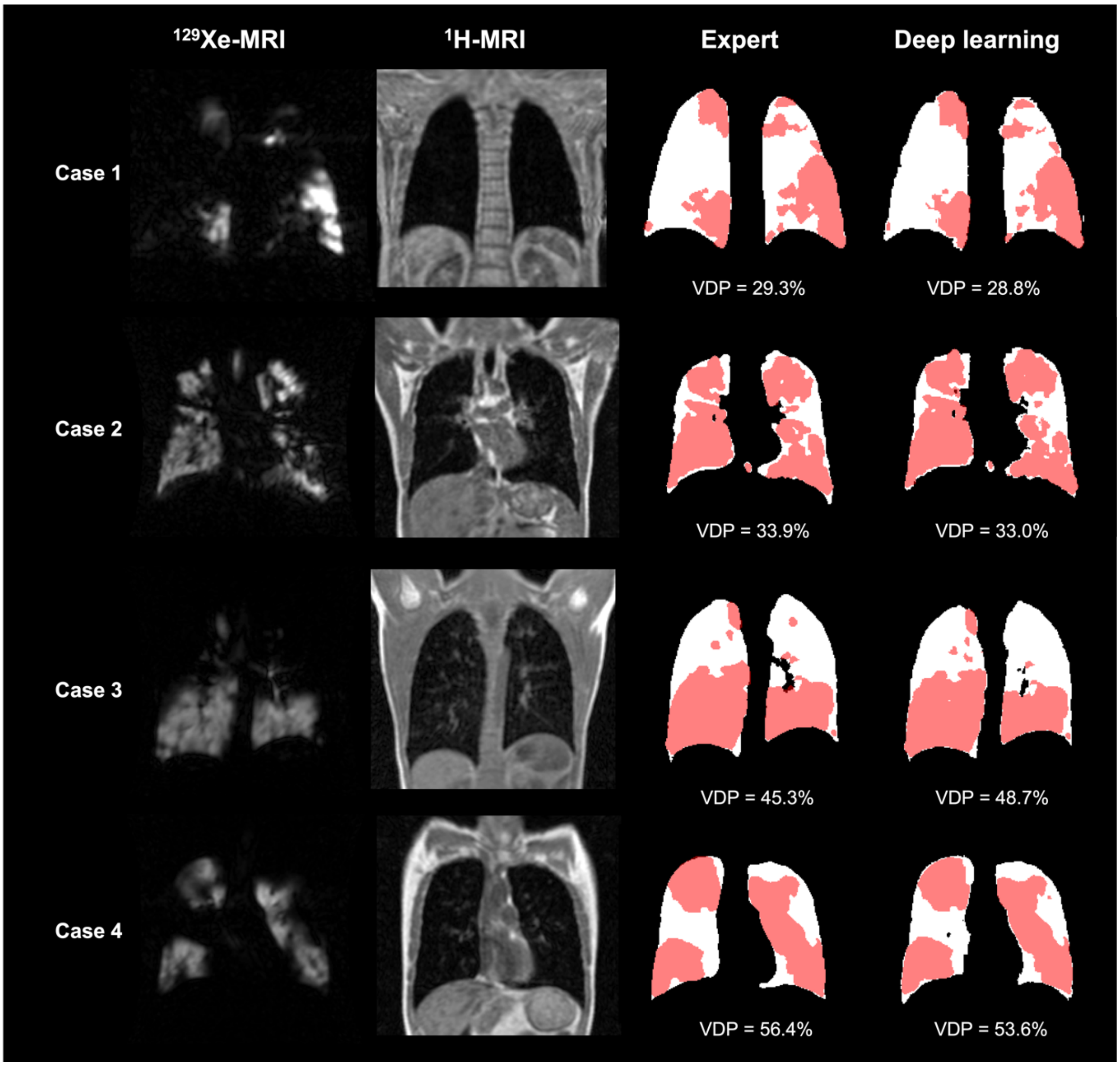


Figure 6.8 Example coronal slices of four cases with ventilation defects showing fused manual LCEs (white) and hyperpolarised gas MRI ventilated lung segmentations (pink) compared to those generated using the dual-input DL method and previously described hyperpolarised gas MRI ventilated lung segmentation method. Manual and DL-generated VDPs are given for each case.

Figure 6.9a contains an estimation plot indicating that there is no significant difference between DL-generated VDPs and manual VDPs (p=0.71). In addition, Bland-Altman analysis of bias using the VDP values resulted in a bias of -0.19% and LoA of -7.73% to 7.35%. A Bland-Altman plot is shown in Figure 6.9b for the VDP generated using the proposed DL workflow compared to VDP values from manual assessment.

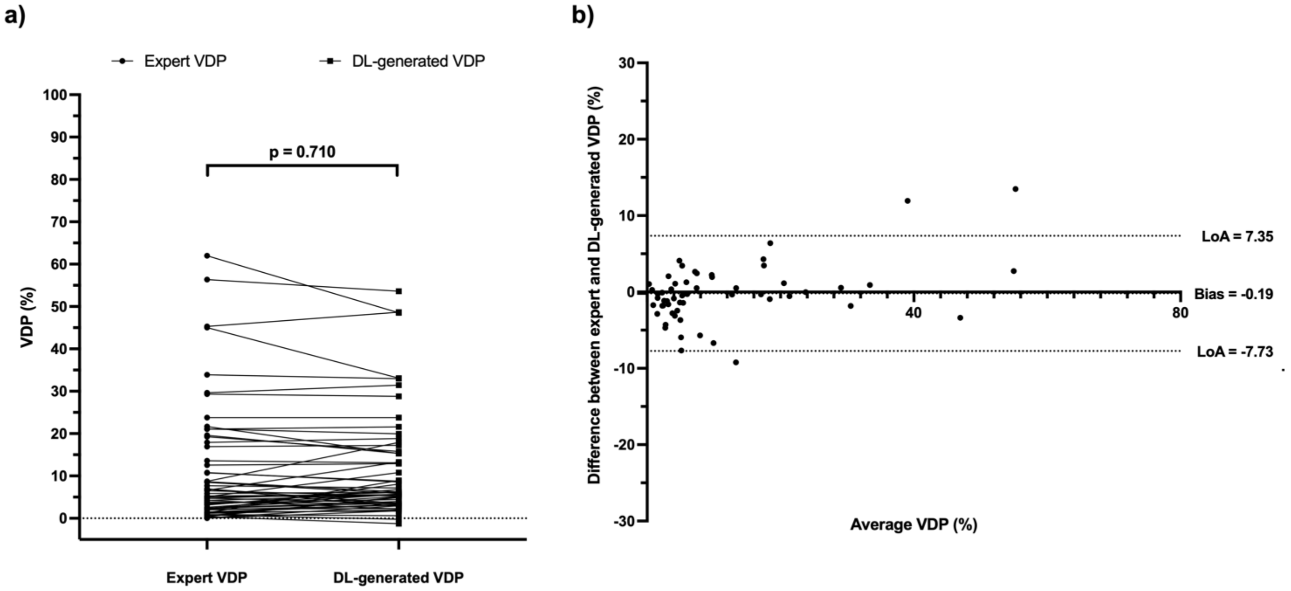


Figure 6.9 a) Estimation plot of manual- and DL-generated VDPs (left) with significance of differences; b) Bland-Altman analysis (right) of VDPs for 58 testing set cases comparing the manual LCE to the dual-input DL-method.

6.4.3 Image quality assessment

For ^{1}H-MRI scans, all three readers agreed on 12 cases and the majority opinion of two readers was used for three cases, resulting in 15 testing set ^{1}H-MRI scans containing an artifact. 9 scans were not included as only one reader identified them as containing an artifact. For ^{129}Xe-MRI scans, all three readers agreed on two cases and the majority opinion of two readers was used for 10 cases, resulting in 12 testing set ^{129}Xe-MRI scans containing an artifact. 13 scans were not included as only one reader identified them as containing an artifact. Five cases within the testing set contained artifacts in both ^{1}H-MRI and ^{129}Xe-MRI scans. Artifacts included zipper, aliasing, signal dropout, motion, wrap-around and image warping. Figure 6.10a concerns the presence of image artifacts identified by the three independent readers in either the ^{1}H-MRI or ^{129}Xe-MRI scans. The differences between the manual and DL-generated VDPs were significantly impacted by the presence of imaging artifacts in ^{129}Xe-MRI scans; similar effects were not exhibited when considering artifacts in ^{1}H-MRI scans (p=0.67). Figure 6.10b plots the Spearman's correlation between the difference in VDP with SNR and shows that there was no significant correlation between the two variables for both ^{1}H-MRI (p=0.22) or ^{129}Xe-MRI (p=0.49) scans.

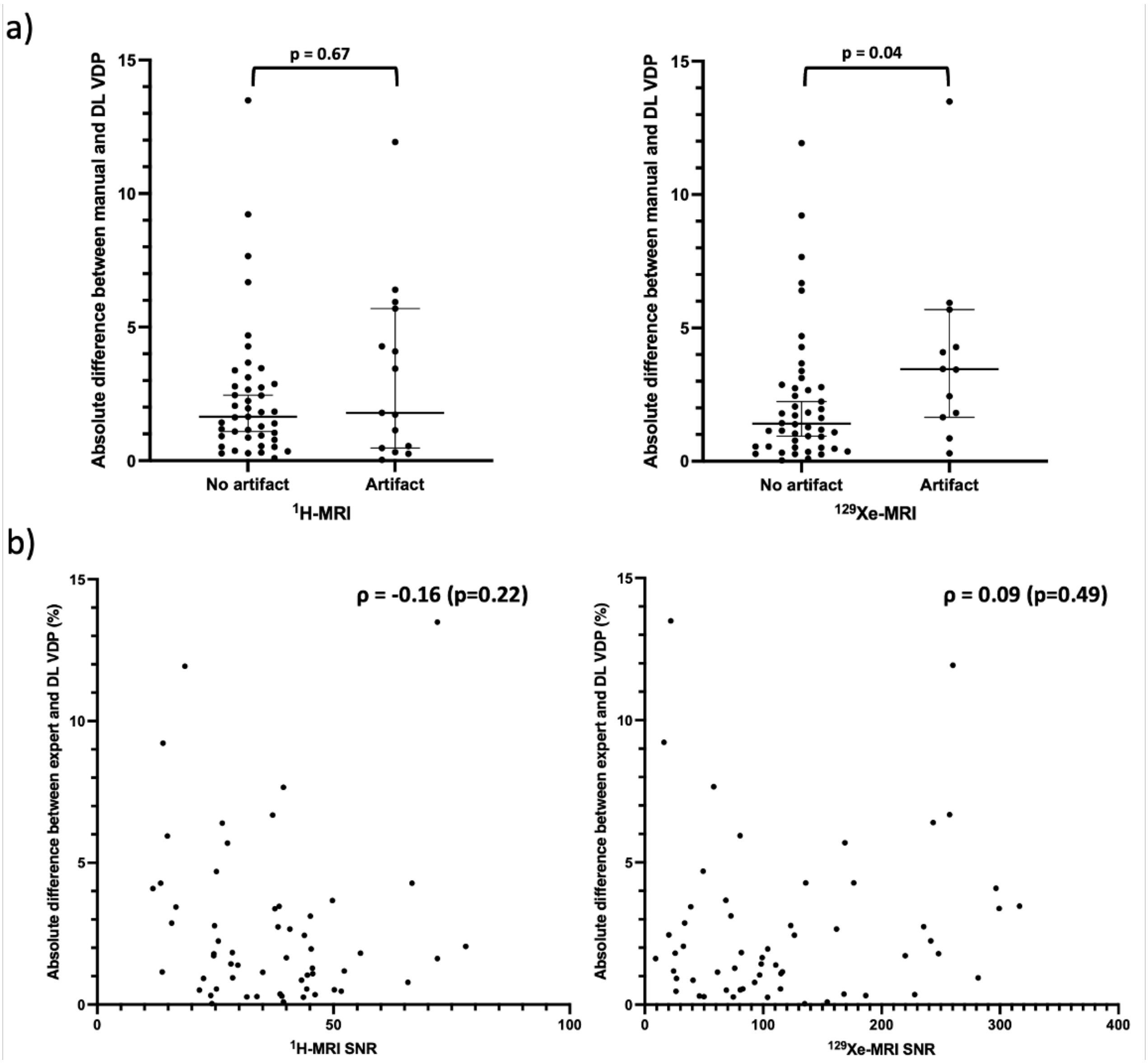


Figure 6.10 a) Absolute differences between manual and DL VDPs stratified by presence (or absence) of image artifacts in the ^{1}H-MRI (left) and ^{129}Xe-MRI (right) scans. Mann-Whitney U tests were conducted, and p values indicated. b) Scatterplot of absolute differences between manual and DL VDPs and SNRs for the ^{1}H-MRI (left) and ^{129}Xe-MRI (right) scans. Spearman's ρ values are provided.

Figure 6.11 displays three failure cases where the differences in VDP between manual and DL-generated VDPs are outside the LoA in the Bland-Altman analysis. Case 1 contained a gas motion artifact on the ^{129}Xe-MRI, leading to an error in the segmentation around this region. Case 2 contained a zipper artifact in the ^{1}H-MRI, which traversed the lung parenchyma, possibly contributing to errors in the DL-generated LCE. Case 3 showed a large degree of noise in the ^{129}Xe-MRI scan.

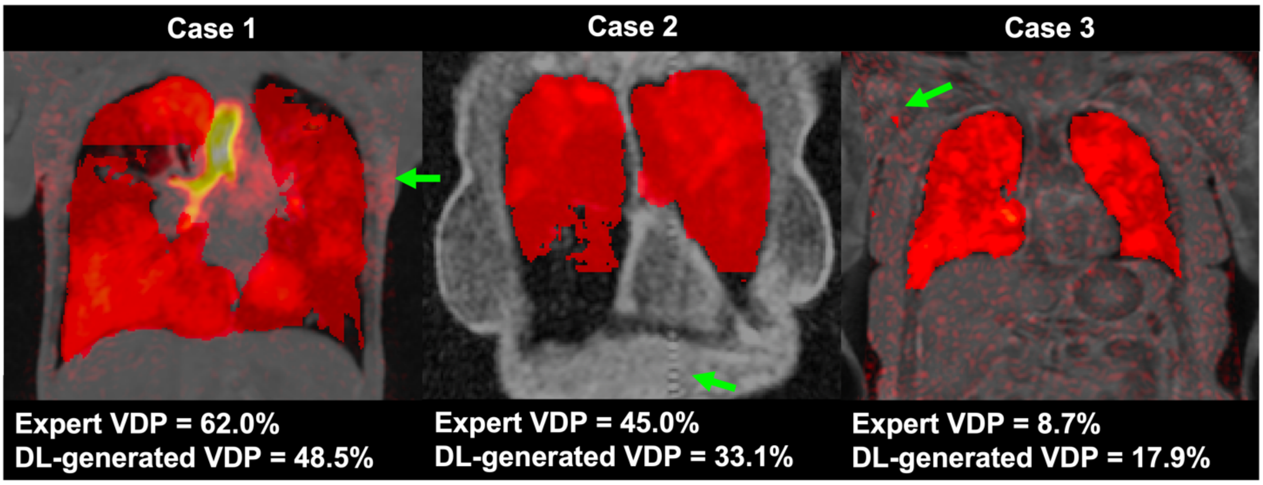


Figure 6.11 Example coronal slices of fused ^{1}H-MRI, ^{129}Xe-MRI and LCEs from three cases in the testing set which, through Bland-Altman analysis, fall outside the limits of agreement. Manual and DL-generated VDPs are given. Case 1 contains a motion artefact on the ^{129}Xe-MRI. Case 2 contains a zipper artefact on the ^{1}H-MRI. Case 3 exhibits a large degree of noise throughout the ^{129}Xe-MRI. Artifacts are indicated with green arrows.

6.5 Discussion

In this study, we propose a dual-channel CNN for LCE that leverages ^{1}H-MRI and ^{129}Xe-MRI scans. Our method significantly outperformed single-channel alternatives that do not integrate both functional and structural lung imaging in a range of diseases for adult and paediatric participants. Furthermore, we combined this dual-channel LCE approach with a DL-based method for hyperpolarised gas MRI ventilated lung segmentation to automatically generate a key clinical biomarker of lung function, namely, the VDP, showing strong agreement with manually-derived VDPs. The proposed method showed no reduction in performance in scans with a large degree of noise; however, it showed decreased performance when artifacts were present in ^{129}Xe-MRI scans.

Qualitative comparison of the various DL methods demonstrated the differences in LCEs due to varying modalities used in the input channels. For the majority of cases, the ventilation-only method was unable to generate realistic LCEs due to the lack of structural features provided to the CNN. Conversely, the structural-only method generated reasonable LCEs; however, in cases where there were misalignments between the ^{129}Xe-MRI and ^{1}H-MRI scans, the structural-only DL method could not account for the inherent registration errors. Misalignments were addressed in the dual-input method using both ventilation and structural features in the input channels, probably providing the network adequate context to accurately generate LCEs that represented structural lung regions in the domain of ^{129}Xe-

MRI. This seems supported by the quantitative results adjusted for multiple comparisons, indicating that the dual-input method significantly outperformed single-channel methods across all metrics tested.

The nn-UNet employed is specifically designed to reduce memory constraints during network training, a requirement that benefits the dual-input method, facilitating the use of larger batch and patch sizes (Isensee *et al.*, 2018). Previous studies have described DL-based approaches to segment the lung parenchyma on ^{1}H-MR images; however, these approaches have conducted the segmentation using single-channel networks (Tustison *et al.*, 2019; Zha *et al.*, 2019; Astley *et al.*, 2021). The inclusion of functional features present in the hyperpolarised gas MRI scans may provide the network context with which to adapt the structural LCE to account for inherent registration errors between the ^{1}H-MRI and ^{129}Xe-MRI acquisitions. Previous work by Tustison et al. utilised separate networks for segmenting ^{1}H-MRI and hyperpolarised gas MRI (Tustison *et al.*, 2019); however, due to several factors, including inherent registration errors and differences in inflation levels, a network that generated a structural segmentation purely using ^{1}H-MRI seems inadequate.

Although same-breath acquisition of helium-3 (^{3}He) and ^{1}H-MRI has been leveraged in previous studies (Wild *et al.*, 2011; Tahir *et al.*, 2018; Tahir *et al.*, 2014; Horn *et al.*, 2014; Stewart *et al.*, 2018), due to the lower bandwidths and longer repetition times required for ^{129}Xe-MRI, owing to its lower intrinsic signal intensity compared to ^{3}He, longer acquisition times and thus longer breath-holds are inevitable. These are prohibitively long for many patients who are unable to maintain lengthy breath-holds, inducing movement, particularly at the diaphragm. In this study, ^{129}Xe-MRI was acquired in approximately 10 seconds; ^{129}Xe-/^{1}H-MRI back-to-back acquisition times would be approximately 19 seconds. Our recent work with compressed sensing has enabled us to reduce this time to 15 seconds (Collier *et al.*, 2019); however, although the shorter breath-hold is more feasible for patients, the likelihood of changes in lung posture during back-to-back scanning persist. As such, a lung cavity estimation will still be required for many patients.

Tustison et al. used a 3D UNet CNN to generate ^{1}H-MRI lung segmentations (Tustison *et al.*, 2019). However, the authors noted that this limits the batch size due to computational constraints; the nn-UNet used here may overcome these challenges (Isensee *et al.*, 2018). Additionally, the authors generated ventilated lung segmentations of hyperpolarised gas MRI using a 2D CNN (Tustison *et al.*, 2019). Conversely, both the dual-channel DL approach to LCE generation and the single-channel DL approach to hyperpolarised gas MRI ventilated lung segmentation used here employed 3D CNNs. These 3D CNNs can process images in a fully volumetric fashion. LCEs represent volumetric lung parenchymal regions that are located across multiple slices in the scan; consequently, the network's ability to process scans in three dimensions potentially enhances the delineation of lung parenchymal volumes compared to 2D alternatives, which do not allow the network to learn inter-slice features of the scan that occur in a volumetric fashion; this has been demonstrated previously in the segmentation of adipose tissue in cardiac MRI (Kulasekara *et al.*, 2022). We used a large, diverse training set comprising patients with numerous pulmonary pathologies and used a testing set that contains only one scan from each participant. This resulted in a robust dual-channel CNN, which may be demonstrated by the limited bias in the Bland-Altman analysis that showed that the accuracy of the LCEs does not diminish with changing volumes.

Furthermore, evaluation of VDP may demonstrate the ability of DL to both produce accurate LCEs and ventilated lung segmentations. The VDPs generated using the DL workflow exhibited no statistically significant differences with manual VDPs. In addition, the Bland-Altman analysis of VDP showed a bias of only -0.19%. This may indicate that the DL-generated workflow can provide statistically indistinguishable VDPs without subsequent editing. Removing the editing step could allow for a more streamlined workflow to generate automatic VDP values. This, in turn, leads to a vast reduction in the time taken to generate VDP values. Previous approaches to edit segmentations generated by semi-automatic segmentation methods could take ~1.5 hours per scan. The automatic DL-based approach proposed here may eliminate this editing time or could at least drastically reduce it. In addition, inference using the dual-input method could yield accurate LCEs in ~30 seconds using a single GPU, further facilitating the computation of rapid and robust VDPs, leading to potentially higher clinical throughput.

For all testing set cases, we assessed the impact of SNR and imaging artifacts on DL-generated VDPs and observed that our approach is potentially invariant to SNR. No significant impact on VDP accuracy was observed due to the presence of at least one artifact (n=15) on the ^{1}H-MRI scans (p=0.67). In contrast, for ^{129}Xe-MRI, there was significantly reduced VDP accuracy for images containing at least one imaging artifact (n=12). This may indicate that the presence of imaging artifacts in ^{129}Xe-MRI scans has the potential to produce inaccurate DL-generated VDPs, representing a challenge for this approach. The prevalence of imaging artifacts in the training set was not assessed and therefore it cannot be concluded whether the network was exposed to these features previously. In addition, there was less agreement between readers for ^{129}Xe-MRI artifacts, reducing the generalisability of this evaluation.

6.5.1 Limitations

The large dataset used for this study contains participants with numerous pulmonary pathologies; however, each scan in the dataset is acquired with the same acquisition protocol. This reduces the generalisability of the model as performance has not been demonstrated on scans acquired at a different centre, using a different scanner manufacturer, with different field strengths or MRI sequences. Therefore, the proposed DL model is potentially limited in its application to scans acquired with different acquisition protocols. Future investigations will aim to validate approaches on a wider range of scan acquisitions, facilitating inter-centre deployment of the proposed DL approach. Nonetheless, we have made our trained model publicly available which will enable other centres to tailor the model to their unique datasets via the use of fine-tuning and transfer learning.

Whilst there are multiple examples of good segmentation performances on ^{1}H-MR images with imaging artifacts, the clinical implications of reduced performance on some of these scans is a limitation of our study. Future investigations could employ multiple strategies to reduce the impact of imaging artifacts on DL performance; this could be done by implementing specialised data augmentation techniques such as increasing the proportion of images containing each specific artefact, boosting their prevalence during network training, or by artificially augmenting scans with plausible, synthetic noise. In addition, it may be feasible to build a secondary network to identify the presence of imaging artifacts, hence triggering a manual review; however, there is unlikely to be a sufficiently large dataset to build an effective model for this purpose.

In future work, it may be possible to generate both ventilated and structural lung segmentations within a single model using a dual-class segmentation network. This approach would have the inherent benefit of co-location, thereby potentially further dealing with misalignments between imaging modalities. However, the DL-generated hyperpolarised gas MRI segmentation method used in this work utilised a dataset comprising 759 scans, significantly larger than the dataset used here for LCE; hence, generating ventilated lung segmentations in a dual-class model would reduce the size of the training set, and consequently likely reduce segmentation performance.

6.6 Conclusion

We used a dual-channel 3D CNN approach for LCE and compared it to single-channel DL methods. We demonstrated that the dual-channel approach, leveraging both hyperpolarised gas and ^{1}H-MRI as inputs, may yield improved LCEs. In addition, we used this approach in conjunction with a DL-based hyperpolarised gas MRI segmentation method to automatically generate VDPs, which did not significantly differ from manual VDPs.

Chapter 7
A hybrid model- and deep learning-based framework for functional lung image synthesis from multi-inflation CT and hyperpolarised gas MRI

Background: Hyperpolarised gas MRI is a functional lung imaging modality capable of visualising regional lung ventilation with exceptional detail within a single breath. However, this modality requires specialised equipment and exogenous contrast, which limits widespread clinical adoption. CT ventilation imaging employs various metrics to model regional ventilation from non-contrast CT scans acquired at multiple inflation levels and has demonstrated moderate spatial correlation with hyperpolarised gas MRI. Recently, deep learning (DL)-based methods, utilising convolutional neural networks (CNNs), have been leveraged for image synthesis applications. Hybrid approaches integrating computational modelling and data-driven methods have been utilised in cases where datasets are limited with the added benefit of maintaining physiological plausibility.

Purpose: To develop and evaluate a multi-channel DL-based method that combines modelling and data-driven approaches to synthesise hyperpolarised gas MRI lung ventilation scans from multi-inflation, non-contrast CT and quantitatively compare these synthetic ventilation scans to conventional CT ventilation modelling.

Methods: In this study, we propose a hybrid DL configuration that integrates model- and data-driven methods to synthesise hyperpolarised gas MRI lung ventilation scans from a combination of non-contrast, multi-inflation CT and CT ventilation modelling. We used a diverse dataset comprising paired inspiratory and expiratory CT and helium-3 hyperpolarised gas MRI for 47 participants with a range of pulmonary pathologies. We performed 6-fold cross-validation on the dataset and evaluated the spatial correlation between the synthetic ventilation and real hyperpolarised gas MRI scans; the proposed hybrid framework was compared to conventional CT ventilation modelling and other non-hybrid DL configurations. Synthetic ventilation scans were evaluated using voxel-wise evaluation metrics such as Spearman's correlation and mean square error (MSE), in addition to clinical biomarkers of lung function such as the ventilated lung percentage (VLP). Furthermore, regional localisation of ventilated and defect lung regions was assessed via the Dice similarity coefficient (DSC).

Results: We showed that the proposed hybrid framework is capable of accurately replicating ventilation defects seen in the real hyperpolarised gas MRI scans, achieving a voxel-wise Spearman's correlation of 0.57±0.17 and an MSE of 0.017±0.01. The hybrid framework significantly outperformed CT ventilation modelling alone and all other DL configurations using Spearman's correlation. The proposed framework can generate clinically relevant metrics such as the VLP without manual intervention, resulting in a Bland-Altman bias of 3.04%, significantly outperforming CT ventilation modelling. Relative to CT ventilation modeling, the hybrid framework yielded significantly more accurate delineations of ventilated and defect lung regions, achieving a DSC of 0.95 and 0.48 for ventilated and defect regions, respectively.

Conclusions: The ability to generate realistic synthetic ventilation scans from CT has implications for several clinical applications, including functional lung avoidance radiotherapy and treatment response mapping. CT is an integral part of almost every clinical lung imaging workflow and hence is readily available for most patients; therefore, synthetic ventilation from non-contrast CT can provide patients with wider access to ventilation imaging worldwide.

7.1 Preface

Work contained within this chapter has been submitted to the *Journal of Medical Physics*:

> **Astley J.R.,** Biancardi A.M., Marshall H., Hughes P.J.C., Collier G.J., Hatton M.Q., Wild J.M. and Tahir B.A. (2022). A hybrid model- and deep learning-based framework for functional lung image synthesis from multi-inflation CT and hyperpolarized gas MRI. *Medical Physics.* [in press].

The work contained within this chapter has also been published as conference proceedings at the following conferences:

> **Astley J.R.**, Biancardi A.M., Marshall H., Collier G.J., Hughes P.J.C., Wild J.M. and Tahir B.A. (2021). A hybrid model- and deep learning-based framework for functional lung image synthesis from non-contrast multi-inflation CT. Medical imaging and deep learning (MIDL) 2021. *Online.*

> **Astley J.R.**, Biancardi A.M., Walker M., Hughes P.J.C., Marshall H., Collier G.J., Hatton M.Q., Wild J.M. and Tahir B.A. (2021). Hyperpolarized gas MRI ventilation synthesis from CT: comparison of conventional and deep learning methods. American association of physicists in medicine (AAPM) 2021. *Online.*

Additional material that could not be included within the journal article or within conference proceedings is also contained within this chapter.

7.1.1 Author contributions

J.R.A. and B.A.T. made substantial contributions to the conceptualisation of the work. A.M.B., P.J.C.H., H.M. L.J.S., G.J.C., J.A.E., N.D.W., M.Q.H., J.M.W. and B.A.T. were involved with patient recruitment, image acquisition and/or analysis. J.R.A. performed the deep learning experiments, interpreted data, and conducted statistical analyses. J.R.A. drafted the manuscript. B.A.T. substantively revised the manuscript. All authors reviewed and approved the submitted manuscript.

7.2 Introduction

Lung diseases represent significant global health challenges (Vos *et al.*, 2017; Torre *et al.*, 2015). Imaging of the lungs constitutes a key component of clinical care, providing both anatomical and functional information for a wide range of lung pathologies. Functional lung imaging modalities such as single-photon emission computed tomography (SPECT), positron emission tomography (PET) and hyperpolarised gas magnetic resonance imaging (MRI) have shown efficacy in several applications such as early diagnosis, functional lung avoidance radiotherapy and treatment response evaluation (Tahir *et al.*, 2017; Ireland *et al.*, 2016; Horn *et al.*, 2017). Hyperpolarised gas MRI is a functional lung imaging modality capable of visualising regional lung ventilation with exceptional detail within a single breath (Fain *et al.*, 2007). Quantitative biomarkers derived from this modality, including the ventilated lung percentage (VLP), provide further insights into regional ventilation (Woodhouse *et al.*, 2005). However, this modality requires specialised equipment, including a laser polariser, and inhaled contrast agents such as helium-3 (^{3}He) or xenon-129 (^{129}Xe) noble gases, which currently limits widespread clinical adoption (Stewart *et al.*, 2021).

Computed tomography (CT) is the most widely used anatomical imaging modality and is an integral part of clinical care for most patients with lung pathologies. CT ventilation imaging (CTVI) aims to model regional ventilation from non-contrast CT scans acquired at multiple inflation levels, either during tidal breathing or breath-hold (Guerrero *et al.*, 2005; Reinhardt *et al.*, 2008). CTVI assumes that changes in regional lung volume and/or lung density between inflation levels is representative of lung ventilation (Ding *et al.*, 2012). Several metrics have been proposed to generate synthetic ventilation maps from multi-inflation CT, such as those that map changes in Hounsfield units (CT^{HU}) or the determinant of the Jacobian (CT^{JAC}) (Guerrero *et al.*, 2005; Reinhardt *et al.*, 2008). The CT^{HU} metric is based on differences in HU intensities between inflation levels whereas the CT^{JAC} metric is a measure of volume expansion computed directly on the deformation vector field between inflations. Previous validation of CTVI methods included assessing Spearman's correlation with well-established lung function measures, such as spirometry, resulting in moderate correlations, ranging from 0.38 to 0.73 for both the CT^{HU} and CT^{JAC} methods (Brennan *et al.*, 2015; Yamamoto *et al.*, 2014). CTVI models have also been validated against nuclear medicine imaging modalities, exhibiting moderate correlation with SPECT and PET imaging (Castillo *et al.*, 2010; Kipritidis *et al.*, 2014); however, these studies report highly variable results and often use small numbers of patients (Yamamoto *et al.*, 2010). Furthermore, nuclear medicine imaging has a relatively poor spatial and temporal resolution and a susceptibility to aerosol deposition artifacts, particularly within defect regions (Jögi *et al.*, 2010; Magnant *et al.*, 2006). In addition, the requirement of radioactive contrast agents makes nuclear medicine imaging unattainable for some patient groups e.g., paediatrics. By using hyperpolarised gas MRI for validation, Tahir *et* al. (2018) showed moderate Spearman's correlations of several CTVI metrics.

Recently, deep learning (DL)-based methods utilising convolutional neural networks (CNNs) have become widespread in numerous lung imaging applications, including image synthesis (Astley *et al.*, 2020b). Zhong *et* al. (2019a) used a CNN to synthesise CT-based ventilation surrogates from 4DCT, reporting a mean square error (MSE) of 7.6%. However, a limitation of this approach is that CT ventilation images, used as the ground truth ventilation, are in themselves the subject of intense validation efforts (Kipritidis *et al.*, 2019). Ren *et* al. (2022) have shown the capability of deriving synthetic perfusion maps from CT using SPECT perfusion as ground truth; a 3D UNet CNN was used, achieving an average Spearman's correlation of 0.81 using 3-fold cross-validation. A Dice similarity coefficient (DSC) value of 0.81 was achieved for both high-functional and low-functional lung regions. Furthermore, Liu *et* al. (2020) proposed a CNN-based approach to synthesise Technegas SPECT ventilation images from non-contrast 4DCT. They demonstrated, after post-processing, Spearman's correlations of 0.73 and 0.71 for 10-phase and 2-phase 4DCT, respectively. 10-fold cross-validation was used, achieving an average DSC across all folds of 0.83 for high-functional lung regions, 0.61 for medium-functional lung regions and 0.73 for low-functional lung regions. Subsequently, Grover *et* al. investigated the utility of CNNs for synthesising Galligas PET ventilation images, demonstrating a mean Spearman's correlation of 0.58 and a mean DSC for high, medium, and low functional regions of 0.55 (Grover *et al.*, 2022). However, SPECT and PET have significantly longer acquisition time, on the order of 30-45 minutes, compared to CT imaging which facilitates acquisition within a single breath. This leads to the possibility of time-delayed ventilation filling (Marshall *et al.*, 2012), reducing the relationship between structural and functional imaging modalities. Conversely, hyperpolarised gas MRI ventilation has an acquisition time spanning a single breath, similar to that of CT, leading to a potentially more accurate representation of ventilation at a specific point in time. Capaldi *et* al. (2020) has recently used a 2D UNet CNN to map free-breathing proton MRI to ^{3}He hyperpolarised gas MRI, achieving a Pearson correlation of 0.87 and a mean DSC of 0.90 and 0.37 for ventilated and defect lung regions, respectively. However, synthesising hyperpolarised gas MRI directly from multi-inflation CT has not yet been demonstrated.

Despite promising results achieved by DL synthesis techniques in multiple domains, there has been a lack of widespread adoption due to an inability to produce physiologically consistent results. Additionally, there is often a shortage of available data representative of a diverse population; to this end, several researchers have proposed the use of hybrid approaches that leverage computational modelling alongside data-driven approaches, such as deep learning (Long *et al.*, 2018), precluding the requirement for large datasets. For example, hybrid physics- and model-based approaches have been used in weather forecasting (Grover *et al.*, 2015), earth surface modelling (Goldstein *et al.*, 2014) and spatiotemporal dynamic systems evolution in robotics (Hamilton *et al.*, 2017). Hybrid approaches have also been used for data generation in situations where there is limited data available (Willard *et al.*, 2020).

We hypothesised that a hybrid framework that integrates physiological-based multi-inflation level CT ventilation modelling and CNN-based DL may generate accurate surrogate ventilation maps. Accordingly, we propose a hybrid model- and DL-based framework, where conventional CT^{HU} models are used alongside structural inspiratory and expiratory CT scans as inputs to a CNN for functional lung image synthesis. In addition, we propose an automatic pipeline for predicting VLPs from the DL-generated synthetic ventilation scans using CNN-based segmentation. Due to the relatively small dataset, data-driven approaches alone are unlikely to generate accurate synthetic ventilation images, especially in patients with significant ventilation defects. Therefore, the combination of data-driven and physiological modelling approaches utilises both methods' benefits to produce physiologically consistent results, whilst also allowing features to be learnt from underlying patterns in the available data.

7.3 Materials and Methods

7.3.1 Dataset

The dataset comprised paired inspiratory and expiratory CT and hyperpolarised ^{3}He MRI scans for 47 patients originating from three clinical observational studies that were approved by the National Research Ethics Committee (REC). Lung cancer (n=16) data was collected between 2015 and 2017 (REC:14/LO/0481) (Tahir *et al.*, 2018). Asthma (n=12) data was collected between 2012 and 2013 (REC:11/EM/0402) (Tahir *et al.*, 2016). Cystic fibrosis (n=19) data was collected between 2013 and 2014 (REC:12/YH/0343) (Marshall *et al.*, 2017).

7.3.2 Image acquisition

Image acquisition details for CT and ^{3}He MRI across the three studies are provided in Table 7.1. Additional image acquisition details are given in the subsequent sections.

CT acquisition

Study 1 (Tahir *et al.*, 2018): comprised 16 lung cancer participants. All participants underwent radiotherapy planning breath-hold CT on a 16-slice Lightspeed scanner (GE Healthcare, Princeton, NJ, USA); each acquisition was acquired within 15-20 seconds.

Study 2 (Tahir *et al.*, 2016): comprised 12 asthma participants. All participants underwent high resolution breath-hold CT with a Sensation 16 CT scanner (Siemens, Forchheim, Germany).

Study 3 (Marshall *et al.*, 2017): comprised 19 cystic fibrosis participants. All cystic fibrosis participants underwent low dose inspiratory and ultra-low dose expiratory non-contrast CT imaging, following the protocol of Loeve *et* al.(2009), on a GE Lightspeed VCT 64 CT scanner (GE Healthcare, Milwaukee, WI, USA). The CT scanner tube voltage was 80 kV for children weighing < 35 kg and 100 kV for those weighing 35 kg and above. Inspiratory scans were performed with a modulating tube current (max 150mA) and expiratory scans were performed at a fixed current of 25 mA; as a result, expiratory scans were lower dose.

MRI acquisition

All subjects underwent 3D volumetric ^{3}He hyperpolarised gas MRI in the coronal plane at FRC+1L with full lung coverage at 1.5T on a HDx scanner (GE Healthcare, Milwaukee, WI, USA). Helium was polarised on-site to around 25% polarisation (GE Healthcare, Amersham, UK). Flexible quadrature radiofrequency coils were employed for transmission and reception of MR signals at the Larmor frequency of ^{3}He (Clinical MR Solutions, Brookfield, WI, USA). An anatomical proton (^{1}H) MRI in the same breath as ^{3}He MRI was acquired for each patient. Details of this acquisition for each study are provided below:

Study 1 (Tahir *et al.*, 2018): Same-breath ^{1}H MRI scans were acquired at the same resolution as ^{3}He MRI using the scanner's inbuilt body coil with a 3D spoiled gradient-echo sequence. Repetition time/echo time were equal to 1.9/0.6 milliseconds with a flip angle of 5° and ±83.3kHz bandwidth.

Study 2 (Tahir *et al.*, 2016): Same-breath ^{1}H MRI scans were acquired at the same slice thickness as ^{3}He MRI with an in-plane resolution of 3x6mm^2 using the scanner's inbuilt body coil with a 2D steady-state free-precision sequence. Repetition time/echo time was equal to 2.4/0.7 milliseconds with a flip angle of 50° and ±167kHz bandwidth.

Study 3 (Marshall *et al.*, 2017): Same-breath ^{1}H MRI scans were acquired at the same resolution as ^{3}He MRI using an 8-element chest receiver array with a 2D steady-state free-precession sequence. Repetition time/echo time was equal to 2.9/0.9 milliseconds with a flip angle of 50° and ±250kHz bandwidth.

Table 7.1 CT and hyperpolarised gas MRI acquisition details.

		Study 1	Study 2	Study 3
	Disease:	Lung cancer	Asthma	Cystic fibrosis
	Total subjects:	16	12	19
CT scans:	Acquisition orientation:	Axial	Axial	Axial
	Dose mode:	Radiotherapy planning	High resolution	Ultra-low dose (expiration) & low dose (inhalation)
	Breathing inflation:	FRC & FRC+1L	FRC & TLC	Inspiratory & expiratory breath-hold
	Slice thickness:	2.5 mm	~ 2.1 mm	2.5 mm
	In-plane resolution:	~ 0.98 x 0.98 mm^2	~ 0.8 x 0.8 mm^2	~ 0.6 x 0.6 mm^2
	Tube voltage / Current:	120kV / 315mA	120kV / 120mA	80-100kV / 25-150mA
Hyperpolarised gas MRI scans:	Hyperpolarised gas:	^{3}He	^{3}He	^{3}He
	Dimension:	3D	2D	2D
	Sequence:	Balanced steady-state free precession	Spoiled gradient echo	Spoiled gradient echo
	Acquisition orientation:	Coronal	Coronal	Coronal
	Breathing inflation:	FRC+1L	FRC+1L	FRC+1L
	Slice thickness:	5 mm	10 mm	10 mm
	In-plane resolution:	~ 4 x 4 mm^2	~ 3 x 3 mm^2	~ 3 x 3 mm^2
	TR / TE:	1.9 / 0.6 msec	3.6 / 1.1 msec	3.6 / 1.1 msec
	Field of view:	40cm	38.4cm	30-40cm
	Flip angle:	10°	8°	8°
	Bandwidth:	±166.6kHz	±63kHz	±63kHz
	Time-difference:	Same day	< 4 days	Same day

Abbreviations: 2D, 2-dimensional; 3D, 3-dimensional; FRC, functional residual capacity; 1L, 1 litre; SD, standard deviation; ^{3}He, helium-3; TR, repetition time; TE, echo time; CT, computed tomography; MRI, magnetic resonance imaging.

7.3.3 Image segmentation

The Chest imaging platform (CIP) (San Jose Estepar *et al.*, 2015) (Harvard, Massachusetts, USA) was used to generate segmentations of the lung parenchyma on inspiratory and expiratory CT scans. These segmentations were subsequently reviewed and manually edited by multiple experienced observers, specifically, B.A.T and J.R.A. Segmentation of the lung parenchyma from ^{1}H MRI scans was conducted using spatial fuzzy c-means clustering (Hughes *et al.*, 2018). ^{1}H MRI segmentations were subsequently manually edited by two experienced observers, namely, B.A.T and P.J.H. Both observers have a PhD in respiratory imaging.

7.3.4 Image registration

Inspiratory and expiratory CT scans were aligned using deformable image registration and subsequently registered to the spatial domain of ^{3}He MRI via a corresponding anatomical ^{1}H MRI scan. The empirically optimised script used identical rigid and affine parameters as the EMPIRE10_BSplineSyn script as previously described (Tahir *et al.*, 2014; Tahir *et al.*, 2018). Registration pipelines consisted of rigid, affine and diffeomorphic stages. For the deformable registration stage, explicit B-spline regularisation was applied to the resulting transform of the affine pipeline. A knot spacing for the update field of 65 mm provided optimal results. Additionally, a 5-level multi-resolution pyramid was used (instead of 4 levels) with down-sampling factors of 10 × 6 × 4 × 2 × 1 and corresponding smoothing Gaussian sigmas of 5 × 3 × 2 × 1 × 0 mm and the normalised correlation coefficient similarity metric with a radius of 2 voxels instead of 4. A step size of 0.2 was selected for the gradient descent optimisation algorithm. All registrations were conducted using the advanced normalisation tools (ANTs) registration framework (Avants *et al.*, 2008) based on parameters provided previously (Tahir *et al.*, 2019). For each patient, two registrations were performed:

1) Inspiratory CT to expiratory CT
2) Expiratory CT to ^{1}H MRI (same-breath as ^{3}He MRI)

Figure 7.1 shows example unregistered inspiratory and expiratory CT images with the corresponding warped CT images in the domain of ^{3}He MRI. Registrations were quantitatively assessed for overlap using the Dice similarity coefficient (DSC) (Dice, 1945).

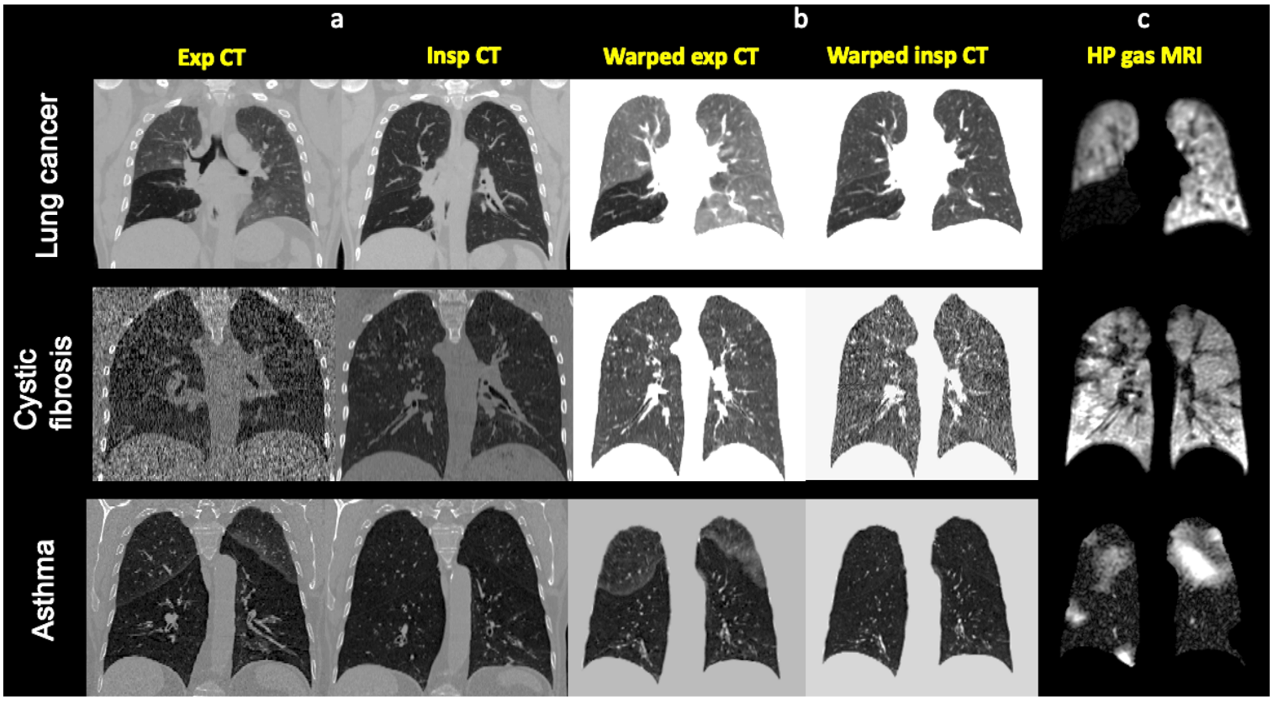


Figure 7.1 Example coronal slices for three patients with lung cancer, cystic fibrosis, or asthma of inspiratory and expiratory CT scans (a) before and (b) after deformable registration to the spatial domain of (c) hyperpolarised gas MRI.

7.3.5 CT ventilation modelling

CT-based surrogate ventilation images were computed using the CT$^{\mathrm{HU}}$ model-based metric originating from theory proposed by Simon *et* al. (2004). CTVI scans were generated at expiratory geometry and computed using voxel-wise intensity differences in Hounsfield unit (HU) values based on the formulation by Guerrero *et* al. (2005) shown below:

$$\mathrm{CT^{HU}} = 1000\frac{\mathrm{HU_{insp}} - \mathrm{HU_{exp}}}{\mathrm{HU_{exp}}\ (1000 + \mathrm{HU_{insp}})} \tag{7.1}$$

where $\mathrm{HU_{insp}}$ represents the HU of voxels in the warped inspiratory scan which spatially correspond to voxels in the expiratory scan and $\mathrm{HU_{exp}}$ represent the HU of inspiratory and expiratory voxels. CT$^{\mathrm{HU}}$ aims to measure the change in the fractional content of air, in a voxel-wise manner, between expiratory and inspiratory phases (Simon *et al.*, 2012). The method assumes that there is uniform air distribution in a given parenchymal voxel and that the observed change in lung density between respiratory phases is attributable solely to changes in ventilation. Several CTVI works have employed various degrees of filtering to account for image noise and possible registration errors (Tahir *et al.*, 2018; Castillo *et al.*, 2010; Castillo *et al.*, 2012; Kipritidis *et al.*, 2014). This has previously been used for post-processing of CT^{HU} ventilation methods in the range of 1x1x1 to 7x7x7 median filtering (Tahir *et al.*, 2018); after a series of investigations, median filtering to CT^{HU} ventilation images across the whole lung region with kernel size 6x6x1 achieved the best results. An anisotropic kernel was used due to the anisotropic resolution of ^{3}He MRI.

7.4 Deep learning experiments and evaluation

7.4.1 CNN architecture configurations

We evaluated four CNN configurations using either single-channel or multi-channel inputs as follows:

1) Expiratory CT
2) Inspiratory CT
3) Expiratory CT + inspiratory CT
4) Inspiratory CT + expiratory CT + CT^{HU} model

For each configuration, input feature maps constituting patches of 128x128x48 voxels were used due to memory constraints. Patches were fed into a 3D fully-convolutional neural network with VNet architecture (Milletari *et al.*, 2016). The network consisted of convolutional steps containing between one and three convolutional layers with subsequent deconvolutional steps, enforcing the original input resolution. As demonstrated by Milletari *et* al. (2016), each step is designed to learn residual functions by initially processing the first convolutional layer using a non-linear activation function and subsequently replicating this output to the last convolutional layer within the step (Milletari *et al.*, 2016). Convolutional operations in the initial input block used two convolutional layers with 5x5x5 kernels and a stride of 1 followed by 2x2x2 kernels with a stride of 2 to reduce image dimensionality. For the multi-channel configurations 3) and 4), we concatenated network blocks, combining the feature maps from spatially aligned inspiratory CT, expiratory CT and CT^{HU} modelling. This allowed the network to make use of concordant features represented across multiple inflation levels and modalities (Berger *et al.*, 2018). The rest of the network consisted of four convolutional blocks that contained a varying number of convolutional layers with either 5x5x5 kernels with a stride of 1 or 2x2x2 kernels with a stride of 2, resulting in a maximum of 248 channels. Each convolutional operation employed a PReLU non-linear activation function with valid padding. Subsequent deconvolutional blocks, with the same structure as the convolutional blocks, were used to reduce the number of channels. Fine-grained feature forwarding introduced residual functions to corresponding convolution and deconvolution steps. The final output block made use of a 1x1x1 convolutional layer.

7.4.2 CNN training parameters

All hyperpolarised gas MRI, CT and CT^{HU} ventilation scans were masked by their respective lung parenchymal segmentations, thereby eliminating the effect of background voxels and allowing the network to focus on features within the lung parenchyma. All hyperpolarised gas MRI scans used in the dataset underwent pre-processing to normalise image intensities to values between 0-1. Training data was augmented to reduce overfitting whilst still maintaining physiological plausibility. To do this, we employed constrained random rotations with limits -10° to 10° and scaling of -10% to 10%, where a different random rotation or scaling for each axis was applied at an interval within the defined limits above. The data augmentation method used does not increase the overall number of scans in the dataset; instead, each scan is given random scaling and rotation factors before being fed into the network. Therefore, the number of epochs can be increased as each time a scan is passed through the network, it is plausibly augmented by a different random factor at each epoch. Batch normalisation was applied for each pass using a mini-batch size of 2 with the aim of reducing covariate shift between network layers (Ioffe and Szegedy, 2015). The weights of the network were trained from scratch and initialised using Xavier initialisation, representing a Gaussian distribution with mean of 0 and variance of 1/N, where N represents the number of weights and biases. A root mean square error (RMSE) loss was used to optimise the network employing ADAM (Kingma and Ba, 2015) optimisation with an initial learning rate of 1x10^{-5}, reducing by a factor of 10 after 1500 epochs and trained for a total of 2150 epochs. L2 regularisation with a decay of 0.00001 was used to penalise large network weights and minimise potential overfitting. Training and testing were performed using TensorFlow (Abadi *et al.*, 2016) 1.15 and Python 3.6 (Gibson *et al.*, 2018b). Training was parallelised across four NVIDIA Tesla V100 GPUs each with 16GB of RAM.

Due to the somewhat limited size of the dataset, we employed 6-fold cross-validation, generating six separately trained models tested on a random subset of 7 or 8 patients as shown in Figure 7.2. The use of cross-validation to increase the size of the testing set allowed for inferential statistical analyses to be conducted. Each model was stopped at 2150 epochs to constrain model training, mitigating overfitting. All DL configuration outputs were subsequently median filtered with a kernel size of 6x6x1 in line with the filtering applied to CT^{HU} ventilation images.

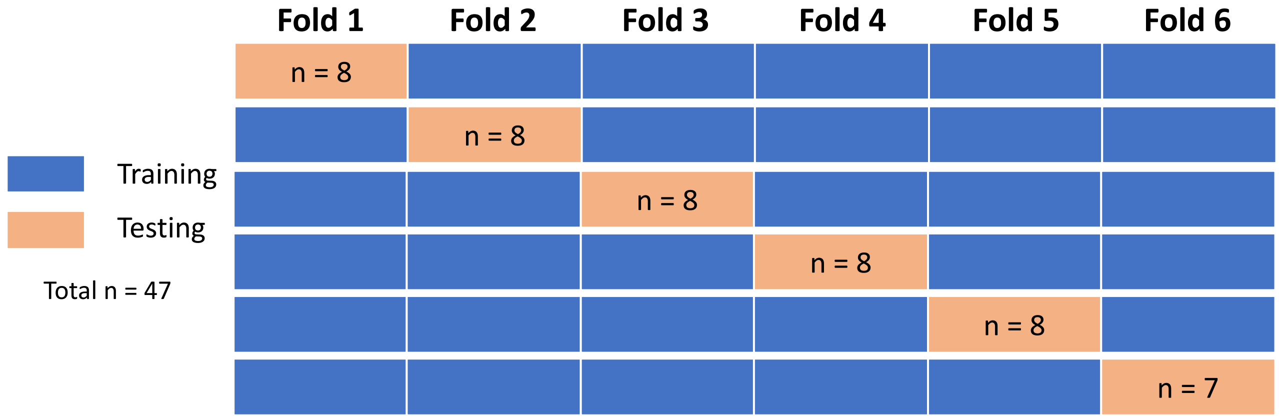


Figure 7.2 Breakdown of cross-validation strategy used for training and testing across 47 patients.

7.4.3 *Quantitative evaluation*

Synthetic ventilation images generated via the CT^{HU} method and DL approaches were quantitatively evaluated using both voxel-wise and clinical metrics. Following previous works in the CTVI field and the VAMPIRE grand challenge, Spearman's correlation was selected as the primary evaluation metric (Kipritidis *et al.*, 2019). DL-based methods were additionally assessed using the MSE metric. Further, based on voxel-wise evaluation metrics, the clinical metric of VLP was computed on the best performing approach. Furthermore, regional localisation of ventilated and defect lung regions was assessed via the DSC.

The spatial correlations of the DL-generated synthetic ventilation images and the CT^{HU} model with corresponding ^{3}He MRI scans were assessed at full resolution using Spearman's rho (ρ) on all voxels within the lung region, defined by the same-breath ^{1}H MRI lung segmentation. Spearman's ρ quantifies the degree of monotonicity between any two ventilation images. It takes a range between -1 and 1 where 1 represents a perfect positive correlation and -1 represents a perfect negative correlation. Consequently, a Spearman's ρ of 0 represents no correlation.

Quantitative performance was further evaluated for all DL-based approaches using the voxel-wise MSE metric. The MSE represents the mean square difference between estimated values and actual values across all voxels within the lung region. MSE is derived from the square of errors and, therefore, always takes a positive value with the MSE approaching 0 as the error concordantly decreases.

The quantitative biomarker of the VLP has been used extensively in the hyperpolarised gas MRI literature as a robust measure of lung function. VLP is calculated by comparing structural and ventilated lung segmentations to generate a percentage value of ventilated lung volume as follows:

$$\text{VLP (\%)} = \left(\frac{\text{ventilated lung volume}}{\text{total lung volume}}\right) \times\ 100 \tag{7.2}$$

In our clinical lung image analysis workflow, VLP values are derived from expert segmentations of hyperpolarised gas MRI for ventilated lung volumes and ^{1}H MRI for total lung volume (Stewart *et al.*, 2021). In this study, we compared these expert VLP values to VLP values derived using the same ^{1}H MRI expert segmentations for total lung volume and DL-based ventilated lung segmentations. We used a previously validated nn-UNet CNN developed for automatic hyperpolarised gas MRI segmentation (Astley *et al.*, 2022) to segment synthetic ventilated lung regions. These segmentations were used to calculate VLP automatically without manual editing. Figure 7.3 depicts a high-level description of the hybrid model/DL workflow and the automatic calculation of VLP values using DL and expert approaches. In addition to VLP values, DSC overlap values were computed between DL-generated or CT^{HU}-generated ventilated lung segmentations and expert ^{1}H MRI segmentations to define both ventilated and defect regions.

7.4.4 Statistical analysis

Statistical analysis was performed using GraphPad Prism 9 (GraphPad, San Diego, CA). In this work, a p-value <0.05 was considered statistically significant. A one-way repeated measures analysis of variance (ANOVA) test for multiple comparisons was used to determine differences between DL configurations for both voxel-wise Spearman's ρ and MSE. Post-hoc paired *t*-tests were used to assess differences in Spearman's ρ between the CT^{HU} ventilation model and the four DL configurations compared to the reference ^{3}He MRI ventilation scans. Kruskal-Wallis tests were used to assess differences in Spearman's ρ between the three studies contained within the dataset. Bland-Altman analyses of bias were used to compare expert VLP values to DL-derived and CT^{HU}-derived VLP values for the best performing DL-based configuration. Paired *t*-tests were used to assess differences in overlap of ventilated and defect lung regions for the best performing DL configuration and the CT^{HU} ventilation model.

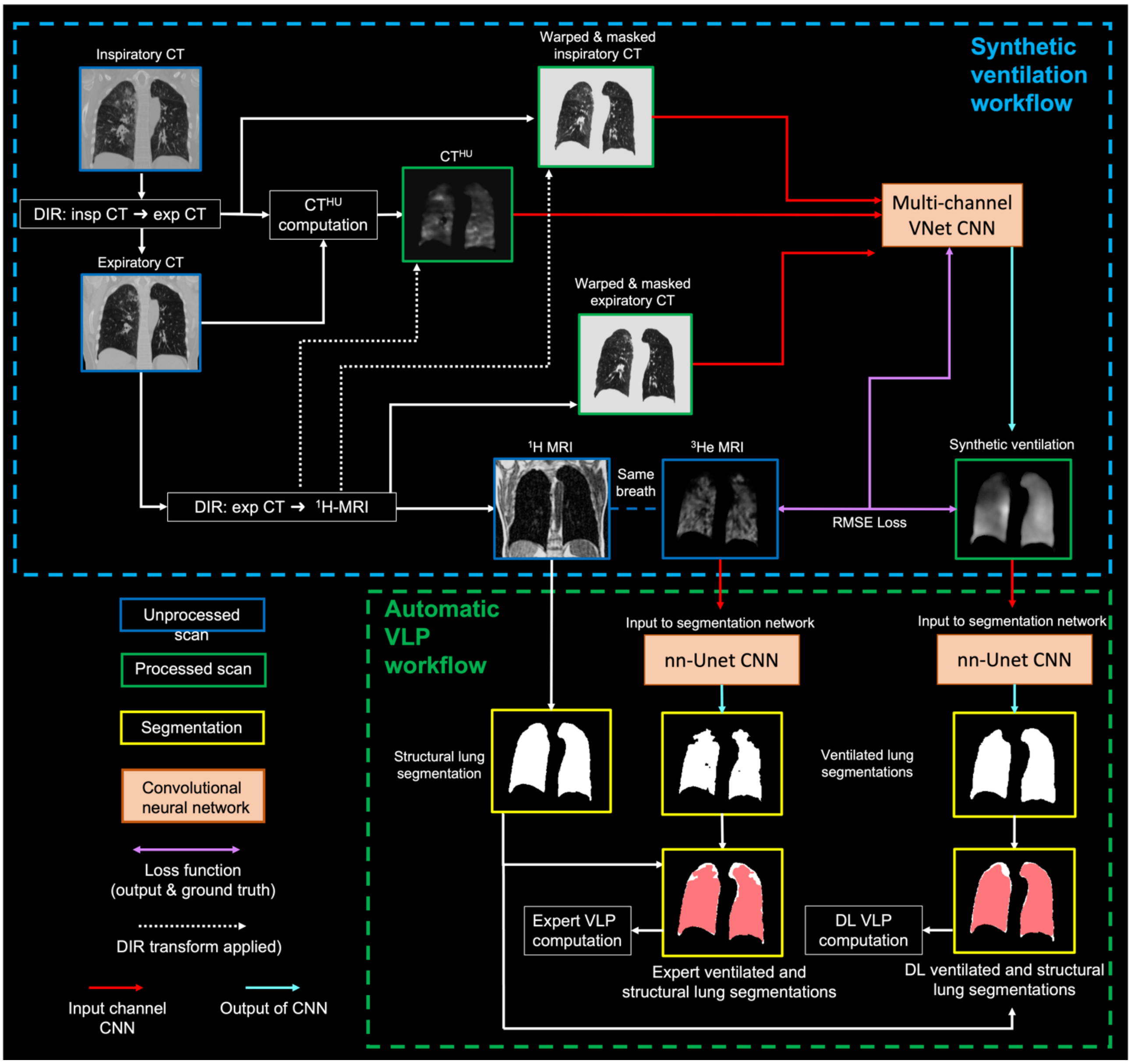

Synthetic ventilation workflow
Inspiratory CT
Warped & masked inspiratory CT
CT^HU
DIR: insp CT ➔ exp CT
CT^HU computation
Multi-channel VNet CNN
Expiratory CT
Warped & masked expiratory CT
¹H MRI
³He MRI
Synthetic ventilation
Same breath
DIR: exp CT ➔ ¹H-MRI
RMSE Loss
Unprocessed scan
Processed scan
Segmentation
Convolutional neural network
Loss function (output & ground truth)
DIR transform applied)
Input channel CNN
Output of CNN
Automatic VLP workflow
Input to segmentation network
nn-Unet CNN
Input to segmentation network
nn-Unet CNN
Structural lung segmentation
Ventilated lung segmentations
Expert VLP computation
DL VLP computation
Expert ventilated and structural lung segmentations
DL ventilated and structural lung segmentations

Figure 7.3 Hybrid model- and DL-based synthetic ventilation workflow and accompanying automatic calculation of VLP.

7.5 Results

7.5.1 Image registration

Registrations between inspiratory CT and expiratory CT, and expiratory CT and ^{1}H MRI were evaluated using the DSC metric. All studies generated a median (range) DSC value exceeding 0.98 for inspiratory and expiratory CT, and 0.91 for expiratory CT and ^{1}H MRI (see Table 7.2).

Table 7.2 Evaluation of overlap between inspiratory and expiratory CT after two-step registration process. Median DSC and range are given for the three studies comprising the data used in this work.

Study	Insp CT to Exp CT Median DSC (range)	Exp CT to ^{1}H MRI Median DSC (range)
Lung cancer	0.984 (0.969, 0.989)	0.963 (0.942, 0.973)
Asthma	0.986 (0.977, 0.988)	0.948 (0.930, 0.960)
Cystic fibrosis	0.983 (0.970, 0.990)	0.919 (0.864, 0.955)

Figure 7.4 shows the alignment of internal lung structures between inspiration and expiration CT scans. The qualitative results indicate that major structures within the lungs, such as bifurcations and vessels, are accurately aligned between inflation levels.

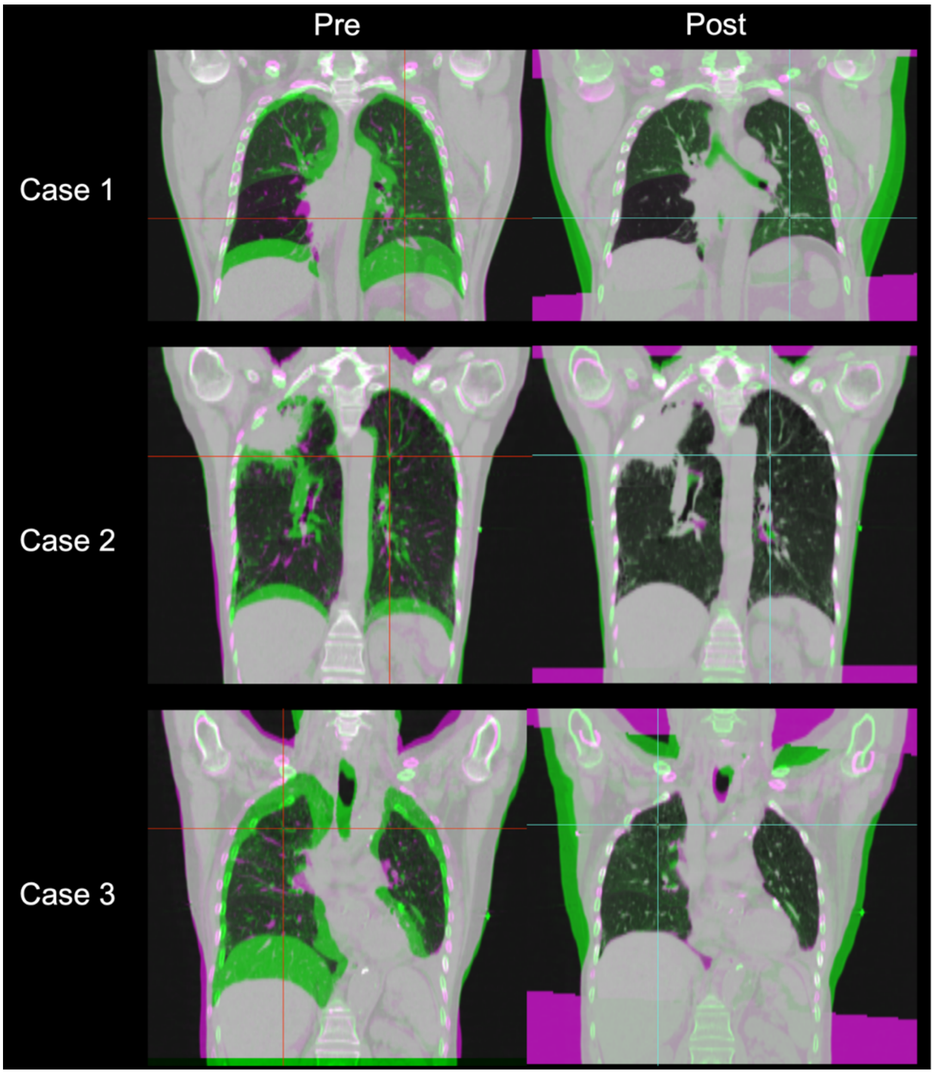


Figure 7.4 Alignment of internal lung structures between inspiration and expiration CT. Bifurcations are marked with crosshairs as landmarks.

7.5.2 Qualitative and quantitative evaluation

Qualitatively, there are numerous examples of the hybrid DL-generated synthetic ventilation images accurately replicating gross ventilation defects in the ground-truth hyperpolarised gas MRI scans. Figure 7.5 shows qualitative spatial agreement between ^{3}He MRI and synthetic ventilation approaches for three example cases. For the three cases displayed, the hybrid DL method, with inspiratory CT, expiratory CT and the CT^{HU} model as inputs, generated the highest Spearman's ρ compared to the CT^{HU} model and all other DL configurations. For Case 1, the differences in performance between DL configurations demonstrate that when a singular structural image is used as an input, the resulting synthesised ventilation scan is unable to capture gross ventilation defects in the left lung; however, when the hybrid DL configuration is utilised, the resulting synthetic scan accurately captures gross ventilation defects which mirror defects observed in the hyperpolarised gas MRI scan.

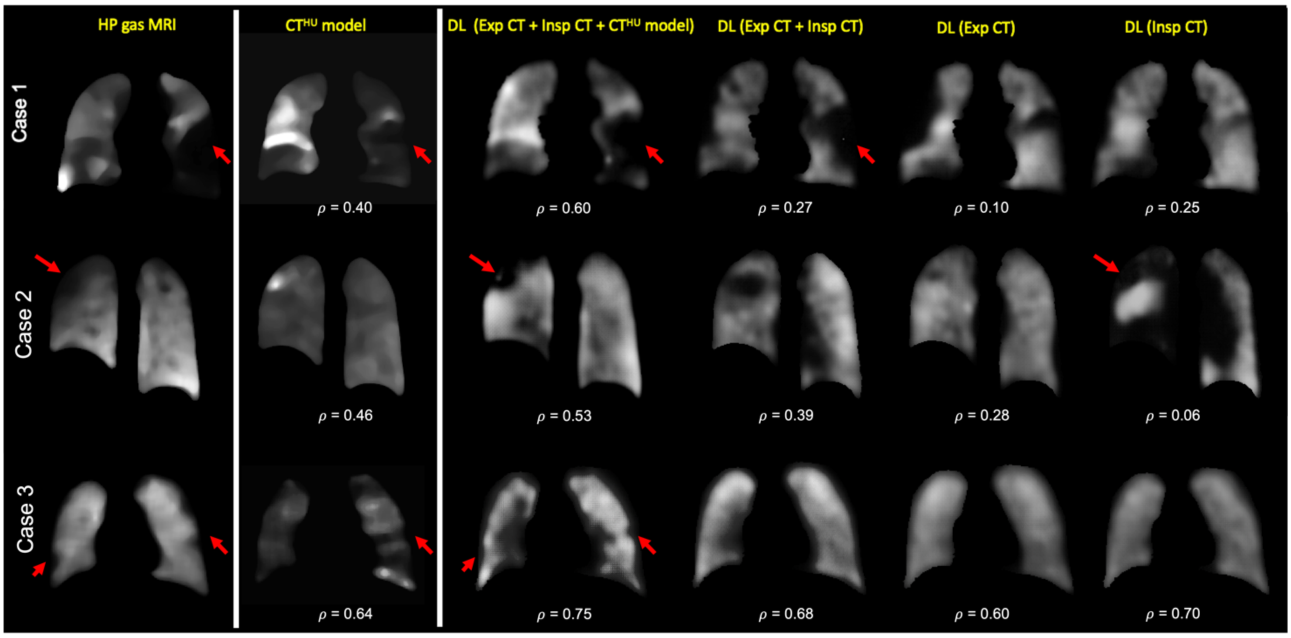


Figure 7.5 Example coronal slices from the CT^{HU} model and the four DL frameworks for three cases compared to ^{3}He MRI. Spearman's ρ values between each method and ^{3}He MRI are provided. Red arrows demonstrate examples of accurately replicated defects.

Significant differences between methods were determined by a one-way ANOVA test ($p<0.05$). The hybrid method yielded statistically significant improvements in Spearman's ρ compared to the CT^{HU} model with mean±SD ρ of 0.57±0.17 vs 0.51±0.22 ($p=0.003$). Furthermore, this approach significantly outperformed all other DL approaches which did not employ the CT^{HU} model as an input ($p<0.05$). DL-based approaches were additionally assessed using voxel-wise MSE; the hybrid approach generated the lowest MSE based on descriptive statistics. No significant differences were observed between the three best performing DL-methods using the MSE metric for synthetic ventilation scans. Table 7.3 summarises the descriptive statistics for all methods across 47 patients via 6-fold cross-validation.

Table 7.3 Descriptive statistics for the CT^{HU} model and DL methods after combining the testing set performance via 6-fold cross-validation. Mean±SD Spearman's ρ for the DL methods and the CT^{HU} model are shown. Additionally, mean±SD MSE are given for the DL methods. The best ρ and MSE values are shown in bold.

Synthetic ventilation generation methods	Filtered Spearman's ρ	MSE
	Mean ± SD	Mean ± SD
CT^{HU} model	0.51 ± 0.22	N/A
DL (expiration CT)	0.52 ± 0.20	0.024 ± 0.01
DL (inspiration CT)	0.47 ± 0.21	0.020 ± 0.01
DL (expiration CT + inspiration CT)	0.52 ± 0.19	0.020 ± 0.01
DL (expiration CT + inspiration CT + CT^{HU} model)	**0.57 ± 0.17**	**0.017 ± 0.01**

Figure 7.6 shows Spearman's correlations between ^{3}He hyperpolarised gas MRI for both the CT^{HU} ventilation model and DL-based configurations; the proposed hybrid framework demonstrated significantly greater Spearman's correlations when compared to all other DL configurations and the CT^{HU} ventilation model. Additionally, MSEs between ^{3}He hyperpolarised gas MRI and DL configurations are displayed, indicating minimal significant differences between DL configurations.

The dataset contains scans from three independent research studies with varying acquisition protocols from participants with varying pulmonary pathologies. No significant difference in Spearman's ρ between datasets was observed using the CT^{HU} ventilation model. A significant difference was observed between the Spearman's *ρ* of Study 1 and Study 3 using the hybrid DL configuration (p=0.03); no other significant differences were observed (Study 1 vs Study 2, p=0.93; Study 2 vs Study 3, p=0.51).

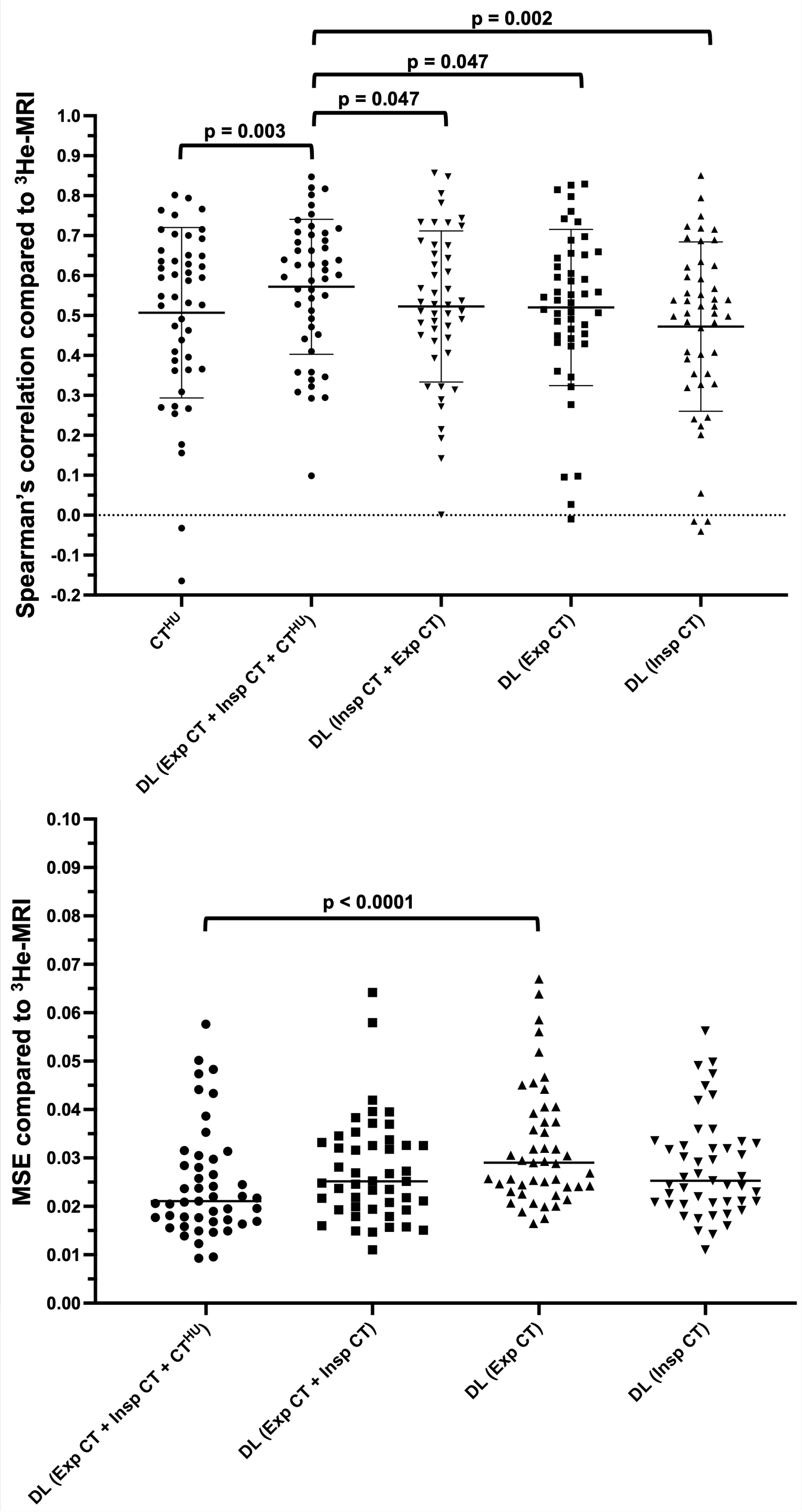


Figure 7.6 (Top) Spearman's ρ values for synthetic ventilation scans derived from the CT^{HU} model and DL configurations. A paired *t*-test compared CT^{HU} with the hybrid DL configuration. One-way ANOVA tests compared Spearman's ρ values for DL configurations. (Bottom) MSE values for synthetic ventilation scans derived from DL configurations. One-way ANOVA tests compared MSE values for DL configurations. Only significant p-values are provided.

7.5.3 VLP evaluation

The hybrid model/DL configuration exhibited significant improvements in Spearman's ρ when compared to all other methods investigated. Therefore, we further investigated this configuration using a clinical metric, namely, VLP. Using the workflow defined in Figure 7.3, we compared expert VLP values to those computed from synthetic ventilation scans generated by the hybrid configuration. Figure 7.7 shows fused structural and functional images with corresponding VLP values for four cases in the dataset. Cases with significant ventilation defects were chosen to illustrate the hybrid framework's ability to replicate gross defects. For example, Case 2 shows almost no ventilation signal in the left lung of the hyperpolarised gas MRI scan which is largely replicated in the output of the hybrid configuration. We used Bland-Altman analyses of bias to compare VLP values derived from hyperpolarised gas MRI versus VLP values derived using the hybrid DL configuration and the CT^{HU} ventilation model as shown in Figure 7.8. The hybrid DL synthetic ventilation surrogates resulted in a bias of only 3.04% with limits of agreement (LoA) of -15.45% to 21.53% compared to the CT^{HU} ventilation model which produced a bias of -10.74% with LoA of -47.55% to 26.07%.

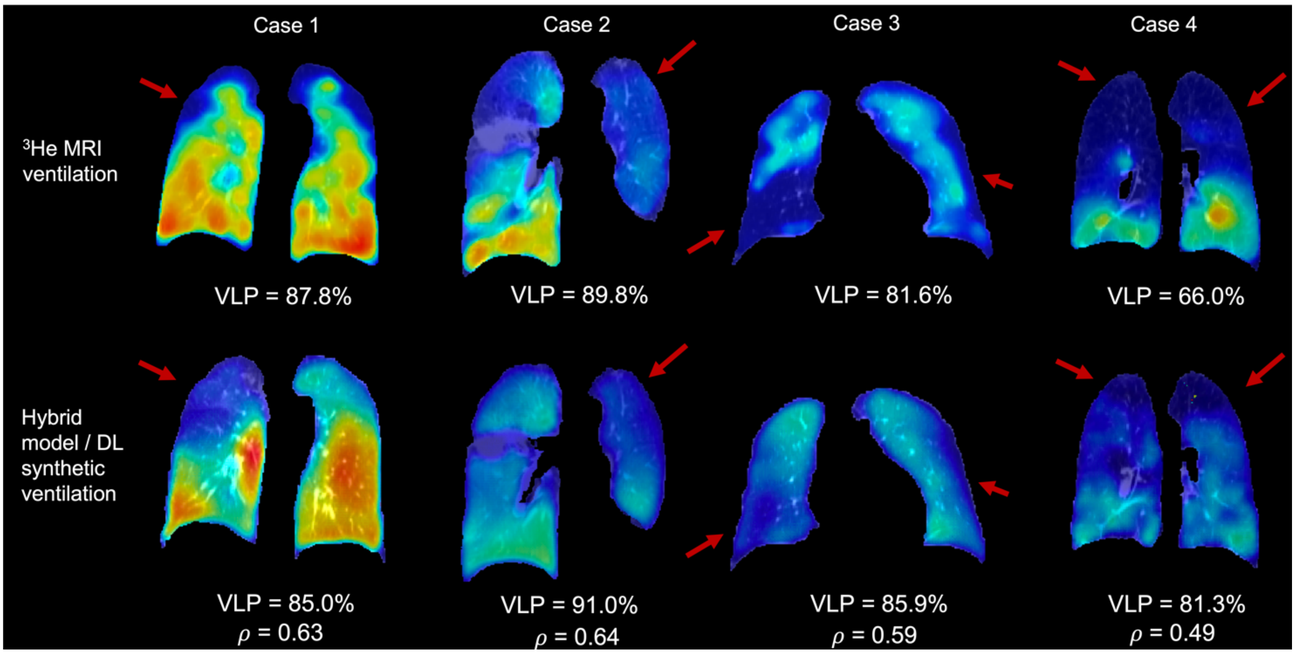


Figure 7.7 Fused ventilation (jet colormap showing minimum to maximum ventilation) and structural scans (grayscale) from four patients derived from either ^{3}He MRI and warped expiratory CT (top) or synthetic ventilation generated using the proposed hybrid model/DL approach and warped expiratory CT (bottom). Red arrows indicate defects replicated in synthetic ventilation scans. VLP values are given.

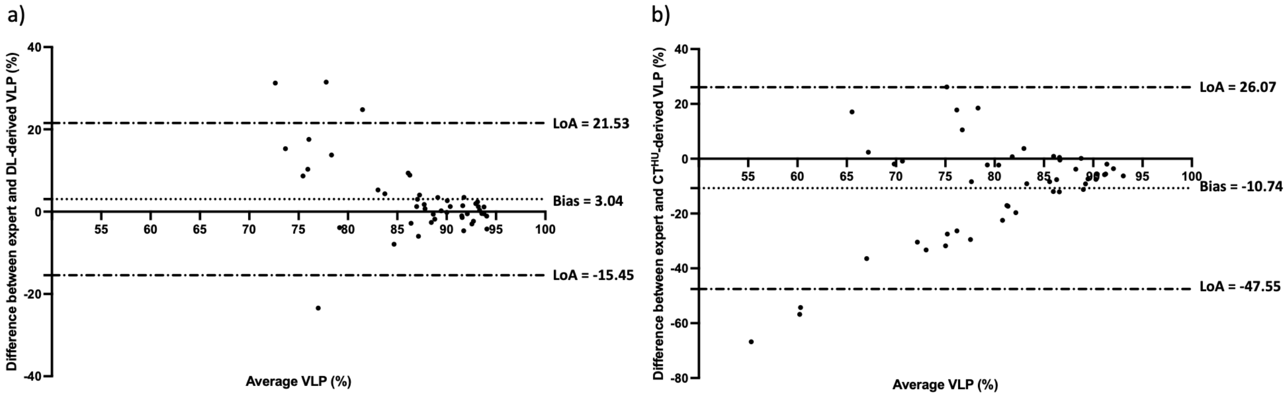


Figure 7.8 Comparison of VLPs derived from hyperpolarised gas MRI versus a) the hybrid model/DL configuration and b) the CT^{HU} ventilation model using Bland-Altman analysis.

DSC values of ventilated and defect lung regions for the hybrid DL and CT^{HU} model are compared to expert ventilated and defect lung regions computed using hyperpolarised gas MRI (see Table 7.4). The hybrid DL configuration produced significantly greater DSC values for both the ventilated and defect lung regions, achieving a median (range) DSC of 0.946 (0.715, 0.977) and 0.483 (0.288, 0.743) for ventilated and defect lung regions, respectively.

Table 7.4 Median (range) DSC of the hybrid DL configuration and CT^{HU} ventilation model for ventilated and defect lung regions. The best DSC values are shown in bold.

Region	Hybrid DL	CT^{HU}
Ventilated lung	**0.946 (0.715, 0.977)**	0.903 (0.046, 0.956)
Defect lung	**0.483 (0.288, 0.743)**	0.426 (0.049, 0.730)

7.6 Discussion

In this work, we proposed a hybrid model- and DL-based framework, integrating CT^{HU} models of lung ventilation and structural, multi-inflation CT as inputs to a VNet CNN capable of producing synthetic ventilation scans that correlated well with corresponding ground-truth ^{3}He MRI ventilation scans. To the best of our knowledge, this work represents the first use of DL to predict hyperpolarised gas MRI ventilation directly from multi-inflation CT. As shown in Figure 7.5 and Figure 7.7, the synthetic ventilation scans generated using the hybrid framework mimic moderate-to-large defects present in the corresponding ^{3}He MRI scans. This has the potential to produce DL-based synthetic ventilation scans from routinely acquired CT scans without exogenous contrast. Compared with conventional CT^{HU} modelling, the hybrid framework yields a statistically significant improvement in spatial correlation. The comparison with CT^{HU} ventilation surrogates is somewhat limited due to the inclusion of pulmonary vessels in CT^{HU} images. Commonly, vessels are excluded from CT^{HU}

images; however, this adds a significant time-consuming manual intervention step. The hybrid configuration developed here can potentially learn to accommodate pulmonary vessels without manual intervention through learning mechanisms. ^{1}H MRI scans used in this study were acquired using spoiled-gradient echo sequences; pulmonary vessels are significantly more challenging to identify using these sequences compared to balanced steady-state free-precession MRI (Wild *et al.*, 2012) or CT; hence delineating corresponding vessels in imaging modalities is a significant challenge. In addition to outperforming conventional CT^{HU} modelling, the hybrid configuration significantly outperformed all other DL configurations using Spearman's correlation, indicating the significant benefit of leveraging classical modelling and data-driven approaches. The hybrid configuration's performance is further enhanced by harnessing a combination of structural and functional modalities. Functional CT^{HU} ventilation images have demonstrated moderate correlation with hyperpolarised gas MRI previously (Tahir *et al.*, 2018); however, differences remain. By combining structural CT images at multiple inflations with CT^{HU} images, additional information contained within the structural images can be utilised to modify the predicted ventilation image via a deep learning approach. The measured Spearman's correlations of the CT^{HU} model and the hybrid configuration demonstrate some correlation with each other, but, crucially the inclusion of structural CT images in combination with the CT^{HU} images as inputs generated a significant improvement in Spearman's ρ compared to the conventional CT^{HU} method or the DL configurations not integrating CTVI modelling. Although Spearman's ρ was utilised as the primary evaluation metric, performance was also evaluated using the MSE. The MSE was not calculated for the CT^{HU} ventilation model as this model is directly derived from HU values which have physiological meaning, limiting a direct quantitative comparison with hyperpolarised gas MRI where specific voxel intensity values are arbitrary and not consistent between scans. The MSE was calculated for all DL configurations, indicating minimal significant differences between DL configurations; this is potentially due to the MSE assessing specific values of intensity, compared to correlations between corresponding voxel intensities, and may be less important than correlated regions of low intensity.

We evaluated the hybrid framework on a diverse and challenging dataset using 6-fold cross-validation. The dataset contained scans of patients with one of three lung pathologies, namely, lung cancer, moderate-to-severe asthma or mild cystic fibrosis. The scans were pooled from three separate clinical studies, resulting in a wide range of acquisition protocols in the dataset: high-dose and low-dose CT; different CT scanner types, settings and breathing manoeuvres; 2D vs 3D ^{3}He MRI; differences in in-plane resolutions and slice thicknesses. The proposed hybrid framework exhibited some differences between studies present in the dataset i.e., between Study 1 and Study 3; however, it cannot be determined whether this variation in performance is due to differences in participant disease or the image acquisition parameters used. The lack of differences when comparing performance of the remaining study combinations indicates a level of robustness and generalisability to both disease and acquisition parameters. 6-fold cross-validation was employed, resulting in six separately trained models. This expanded the number of scans available for evaluation; however, the dataset remains relatively limited in size, containing only 47 patients. Future work will aim to expand the dataset further and investigate novel data augmentation techniques, including synthetic data generation.

Some differences between hyperpolarised gas MRI scans, the hybrid approach and the CT^{HU} model are observed for the majority of example cases. In general, synthetic ventilation scans are less detailed than their corresponding hyperpolarised gas MRI scans in terms of minor ventilation defects within the lung border. The CT^{HU} model performs well in some instances but poorly in others; this is potentially due to registration errors, but the accuracy of the CT^{HU} model is also limited by the signal intensity model utilised. It is hypothesised that lung ventilation cannot be solely captured by differences in signal intensities as other components likely contribute to ventilation. Thus, the hybrid approach showed improved performance when compared to the CT^{HU} model is certain cases. It was also observed that in cases where the CT^{HU} model performed poorly, the hybrid approach also exhibited worse than average performance. Nevertheless, the hybrid approach's ability to modify CTVI ventilation surrogates, in combination with information from multi-inflation structural imaging, leads to improved performance in all cases within the dataset. When comparing the hybrid approach to the DL configuration which only utilises multi-inflation CT, there are some instances where the CT^{HU} model underperformed and, consequently, the hybrid approach also underperformed. However, as demonstrated, the hybrid approach produced significantly more accurate synthetic ventilation scans in terms of Spearman's correlation on average. There is a potential that, as the amount of available representative scans increases, configurations excluding CTVI modelling may generate synthetic ventilation images that are more correlated with hyperpolarised gas MRI scans. In future work, if the dataset is expanded, we can assess whether the inclusion of the CTVI modelling still provides significant performance benefits.

The VNet CNN architecture was used due to its fully-convolutional nature. Fully-convolutional networks contain no fully connected layers and hence contain significantly fewer parameters than conventional networks with fully connected layers; this minimises the network's ability to simply memorise scans within the training set, referred to as overfitting. The fully-convolutional VNet not only reduces the overall number of parameters but also makes the number of parameters independent of image matrix size. Therefore, the network was trained and tested on scans with different matrices and acquisition protocols using fixed-size patches of 128x128x48 voxels. We further reduced the possibility of overfitting using L2 weight regularisation with a decay of 0.00001 to penalise large network weights.

Segmentations of ventilated lung volumes derived from hyperpolarised gas MRI and thoracic cavity volumes derived from structural ^{1}H MRI segmentations have been extensively used in the literature to generate VLPs, an established biomarker of regional lung function (Woodhouse *et al.*, 2005). We demonstrated that VLPs derived from the proposed hybrid framework are comparable with ground truth VLPs from ^{3}He MRI, producing a significantly reduced bias compared to the CT^{HU} method. Bland-Altman analysis of bias, however, indicated that there was reduced accuracy in patients with more significant ventilation defects, resulting in higher predicted VLP values than the corresponding expert values. In addition to VLP analysis, synthetic ventilation scans were segmented using a DL-based segmentation algorithm (Astley *et al.*, 2022) to provide regional localised comparisons of ventilated and defect lung regions. The hybrid DL configuration generated a median DSC of 0.95 for ventilated regions and 0.48 for defect regions, significantly outperforming the DSC achieved by the CT^{HU} method. Both VLP values and regional overlap values require the segmentation of synthetic ventilation scans and are, therefore, susceptible to biases in the segmentation algorithm used; the automatic segmentation method used here was trained to segment hyperpolarised gas MRI and not synthetic ventilation scans (Astley *et al.*, 2022). There is limited consensus on the appropriate segmentation schema required for the delineation of ventilated and defect regions, resulting in an inability to produce accurate comparisons between research studies. It is possible that ventilated lung regions were overestimated during segmentation due to the less pronounced changes in ventilation heterogeneity. Further investigation to improve automatic segmentation of synthetic ventilation scans generated by the hybrid configuration could reduce these biases.

As previously demonstrated by Levin *et* al. (2017), the minimum resolution of functional lung images need not be higher than the smallest pulmonary gas exchange unit, namely, the acinus, which has been estimated to be on the order of 10x10x10mm^3 in adult humans. They further indicate that resolutions of 20x20x20mm^3 may be appropriate due to the spatial clustering of most ventilation defects (Levin *et al.*, 2017).

Our study only investigates one CTVI modelling method, namely, CT^{HU}; however, several other CTVI methods have been used in the literature. Subsequent research will aim to assess the differences in performance of the hybrid approach using classical CTVI metrics, such as CT^{JAC}, and emerging metrics with more robust formulations (Reinhardt *et al.*, 2008; Castillo *et al.*, 2019). One key consideration is the requirement of accurate registration between multi-inflation CT and hyperpolarised gas MRI. Building a network capable of synthesising ventilations scans independent of image registration would reduce the computational costs and time taken to generate synthetic images. Both the CT^{HU} metric and the proposed hybrid model rely on accurate registrations and, consequently, are susceptible to errors in cases where the registration is suboptimal. Removing this requirement would eliminate biases due to errors in registration.

A previous approach by Westcott *et* al. (2019) utilised texture analysis, feature selection and classical machine learning methods to generate synthetic lung ventilation maps from thoracic CT in COPD patients. They evaluated the synthetic ventilation maps using whole-lung metrics; however, more accurate voxel-wise evaluation metrics were not reported.

The ability to generate synthetic ventilation scans from CT has implications for several clinical applications, including functional lung avoidance radiotherapy (Tahir *et al.*, 2017; Ireland *et al.*, 2016) and treatment response mapping (Horn *et al.*, 2017). Kida *et* al. (2016) has previously demonstrated that a Spearman's ρ of ~0.4 between CT^{HU} and SPECT images produces clinically indistinguishable radiotherapy plans. In this study, we observed correlations of ~0.6 between the hybrid DL configuration and hyperpolarised gas MRI, indicating the former's potential clinical utility in functional lung avoidance radiotherapy. Synthesising hyperpolarised gas MRI in comparison to other functional lung imaging modalities such as SPECT has several advantages, including enhanced spatial and temporal resolution and the lack of aerosol deposition artifacts or time-delayed ventilation filling effects. CT is an integral part of almost every clinical lung imaging workflow and hence is readily available for most patients; therefore, synthetic ventilation from non-contrast CT can provide patients with wider access to ventilation imaging worldwide.

7.7 Conclusion

We propose a hybrid model/DL framework to synthesise ventilation scans from routinely acquired non-contrast multi-inflation CT and classical CTVI modelling. We show that a synergy between model-based CTVI and CNN-based learning yields statistically significant improvements in performance compared with conventional CTVI modelling alone and other DL configurations that do not integrate modelling.

Chapter 8
PhysVENeT: A physics-informed deep learning-based framework for the synthesis of 3D hyperpolarised gas MRI ventilation

Functional lung imaging modalities such as hyperpolarised gas MRI ventilation enable visualisation and quantification of regional lung ventilation; however, these techniques require specialised equipment and exogenous contrast, limiting clinical adoption. Physics-based, computational modelling techniques to generate proton (^{1}H)-MRI ventilation surrogates have been proposed. These approaches have demonstrated moderate correlation with hyperpolarised gas MRI. Recently, deep learning (DL) has been used for image synthesis applications, including functional lung image synthesis. Here, we propose a 3D multi-channel convolutional neural network that employs physics-based ventilation modelling and multi-inflation structural ^{1}H-MRI to synthesise 3D synthetic ventilation surrogates (PhysVENeT). The dataset comprised paired inspiratory and expiratory ^{1}H-MRI scans and corresponding hyperpolarised gas MRI scans from 170 participants with various pulmonary pathologies. We performed 5-fold cross-validation on 150 of these participants and used 20 participants with a previously unseen disease (post COVID-19) for external validation. Synthetic ventilation surrogates were evaluated using voxel-wise correlation and structural similarity metrics; the proposed PhysVENeT framework significantly outperformed computational ^{1}H-MRI ventilation modelling and other DL approaches which did not utilise structural imaging and physics-informed modelling. PhysVENeT can accurately reflect ventilation defects and exhibits minimal overfitting on external validation data compared to DL approaches that do not integrate physics-informed modelling.

8.1 Preface

Work contained within this chapter has been submitted to *IEEE Transactions on Medical Imaging* as a journal article:

> **Astley J.R.**, Biancardi A.M., Marshall H., Smith L.J., Hughes P.J.C., Collier G.J., Saunders L.C., Tofan M., Hatton M.Q., Hughes R., Wild J.M. and Tahir B.A. PhysVENeT: A physics-informed deep learning-based framework for the synthesis of 3D hyperpolarized gas MRI ventilation. *Scientific reports* [accepted pending revisions].

The work contained within this chapter has also been published as conference proceedings at the following conferences:

> **Astley J.R.**, Biancardi A.M., Marshall H., Smith L.J., Hughes P.J.C., Collier G.J., Hatton M.Q., Wild J.M. and Tahir B.A. (2022). Deep learning–based synthesis of hyperpolarized gas MRI ventilation from 3D multi-inflation proton MRI. Medical imaging and deep learning (MIDL) 2022. *Zurich, Switzerland.*

> **Astley J.R.**, Biancardi A.M., Marshall H., Tofan M.M, Smith L.J., Hughes P.J.C., Collier G.J., Hatton M.Q., Blè F.X, Hughes R., Wild J.M. and Tahir B.A. (2022). Deep learning-based synthesis of hyperpolarized gas MRI ventilation from 3D multi-inflation proton MRI. The international society for magnetic resonance in medicine (ISMRM) 2022. *London, UK.*

Additional material that could not be included within the journal article or within conference proceedings is also contained within this chapter.

8.1.1 Author contributions

J.R.A., J.M.W. and B.A.T. made substantial contributions to the conceptualisation of the work. A.M.B., P.J.C.H., H.M. L.J.S., G.J.C., J.A.E., N.D.W., M.Q.H., J.M.W. and B.A.T. were involved with patient recruitment, image acquisition and/or analysis. J.R.A. performed the deep learning experiments, interpreted data, and conducted statistical analyses. J.R.A. drafted the manuscript. B.A.T. substantively revised the manuscript. All authors reviewed and approved the submitted manuscript.

8.2 Introduction

The global prevalence of pulmonary diseases constitutes a substantial health challenge (Vos *et al.*, 2017; Torre *et al.*, 2015). Although respiratory diseases remain widespread in developed nations, they are significantly more prevalent in developing nations possibly attributable to factors such as poorer air quality (Portney and Mullahy, 1990), in-home woodburning (Torres-Duque *et al.*, 2008) and tobacco consumption (Mackay and Crofton, 1996) which pose significant respiratory challenges.

Pulmonary imaging constitutes a primary component of the clinical workflow of patients with respiratory diseases; various modalities can provide anatomical or functional information that aids in their diagnosis, monitoring, and treatment. Thoracic computed tomography (CT) and proton MRI (^{1}H-MRI) are used to ascertain anatomical lung information. However, the relationship between parenchymal destruction and regional function is only somewhat understood. Therefore, functional lung imaging modalities such as single-photon emission CT (SPECT), positron emission tomography (PET) and hyperpolarised gas MRI can be used to glean functional insights. These techniques have shown efficacy in several lung disease applications, including diagnosis, treatment planning and treatment response mapping (Tahir *et al.*, 2017; Ireland *et al.*, 2016; Horn *et al.*, 2017). Hyperpolarised gas MRI is a specialised functional lung imaging modality which has excellent sensitivity to abnormal lung function and allows for the visualisation of regional ventilation (Woodhouse *et al.*, 2005; Marshall *et al.*, 2017). However, to acquire hyperpolarised gas MRI ventilation images, specialised equipment such as a gas polariser is required, that can limit widespread clinical uptake (Stewart *et al.*, 2021). Surrogates of regional ventilation computed from structural images acquired at different lung inflation levels have been proposed. CT ventilation imaging (CTVI) models regional ventilation from multi-inflation CT by assessing changes in regional lung density (Guerrero *et al.*, 2005) or lung volume (Reinhardt *et al.*, 2008). CTVI methods are the subject of intense validation efforts (Kipritidis *et al.*, 2019). However, CT imaging is ionising and thus impractical for repeat scanning or scanning of paediatric patients. Analogous to CTVI, structural ^{1}H-MRI has also been used to derive ^{1}H-MRI-based regional ventilation surrogates (Zapke *et al.*, 2006; Bauman *et al.*, 2009; Voskrebenzev *et al.*, 2017). ^{1}H-MRI ventilation models are derived from differences in signal intensities of co-registered voxels in multi-inflation ^{1}H-MRI. The model assumes that these changes reflect naturally occurring density variations in the lungs during breathing (Kjørstad *et al.*, 2017). These computational approaches have shown moderate correlation with hyperpolarised gas MRI

(Capaldi *et al.*, 2018; Tahir *et al.*, 2021). Structural ^{1}H-MRI can be acquired without contrast and is non-ionising, which allows it to be used in paediatric patients and longitudinal applications.

In recent years, deep learning (DL) has been applied to several pulmonary image analysis applications, including image synthesis (Astley *et al.*, 2020b). Ren *et* al. used a pre-trained convolutional neural network (CNN) to synthesise SPECT perfusion maps from CT (Ren *et al.*, 2022); they employed a dataset comprising 33 lung cancer patients and 137 non-lung cancer patients where the proposed approach generated a voxel-wise Spearman's correlation of 0.64 averaged across all lobes. Similarly, Liu *et* al. proposed a CNN-based method to synthesise Technegas SPECT ventilation maps from non-contrast 4DCT using a dataset of 50 participants (Liu *et al.*, 2020). They indicate that, after median filtering, the proposed approach achieved a Spearman's correlation of 0.73 for 10-phase, and 0.71 for 2-phase, 4DCT. Furthermore, Zhong *et al.* (2019a) leveraged a CNN to synthesise CTVI surrogates from 4DCT; they reported a mean±SD structural similarity index measure (SSIM) of 0.88±0.04 (Zhong *et al.*, 2019a). Capaldi *et al.* (2020) used structural free-breathing ^{1}H-MRI to synthesise ventilation MRI surrogates for a single 2D coronal section (Capaldi *et al.*, 2020); a 2D UNet CNN with a mean absolute error (MAE) loss function was used. These ventilation surrogates were correlated with ^{3}He hyperpolarised gas MRI, achieving a Pearson correlation of 0.87 after six-fold cross-validation on a dataset of 114 participants (Capaldi *et al.*, 2020).

Whilst these approaches have demonstrated the efficacy of CNN-based methods for pulmonary image synthesis, the robustness of these approaches and the inability to produce physiologically consistent results limit clinical applicability. In addition, medical imaging datasets are often limited in size and unrepresentative of a diverse population, limiting the effectiveness of DL techniques. Researchers have proposed the use of hybrid networks which combine computational modelling and DL (Long *et al.*, 2018). Specifically, physics-informed DL frameworks have been used in weather forecasting (Grover *et al.*, 2015) and earth surface modelling (Goldstein *et al.*, 2014). Networks integrating computational modelling and DL have also been used for data generation in situations where there is limited data available (Willard *et al.*, 2020). Within the medical imaging domain, Poirot *et al.* (2019) have utilised a physics-informed DL approach for dual-energy CT image enhancement.

Here, we propose a physics-informed DL framework for the synthesis of fully-volumetric 3D lung ventilation maps, leveraging physics-based specific ventilation modelling and structural multi-inflation ^{1}H-MRI in a multi-channel CNN configuration. We compare the proposed framework to DL approaches that do not integrate ventilation modelling or structural ^{1}H-MRI and evaluate the quality of synthetic ventilation scans using voxel-wise metrics.

8.3 Materials and methods

8.3.1 Dataset

The dataset comprised 3D isotropic ^{1}H-MRI scans acquired at approximately total lung capacity (TLC) and residual volume (RV), and hyperpolarised ^{129}Xe-MRI ventilation scans acquired at functional residual capacity (FRC) + bag (for any given participant, the bag volume was titrated based on standing height with a range of 400mL-1L) from 170 healthy participants or patients with various pulmonary pathologies. A summary of participant demographics, stratified by pathology, is provided in Table 8.1. Imaging data was collected retrospectively from several prospective clinical studies and patients referred for clinical imaging. Data use was approved by the Institutional Review Boards at the University of Sheffield and the National Research Ethics Committee. All data was anonymised and all investigations were conducted in accordance with the relevant guidelines and regulations with participants (or their guardians) providing informed written consent. Appropriate consent and permissions were granted by the Sponsors to utilise this data for retrospective purposes.

Table 8.1 Summary of patient demographic data.

Disease	Number of	Age	Sex	VDP
	subjects / scans	Median (range)	Frequency (%)	Median (range)
Asthma	64	53 (13, 74)	30M (47%), 34F (53%)	2.4 (0.07, 30.9)
Asthma + COPD	23	59 (33, 71)	15M (65%), 8F (35%)	7.0 (1.3, 29.3)
COPD	17	65 (48, 73)	6M (35%), 11F (65%)	18.6 (6.2, 64.8)
Cystic fibrosis	31	18 (9, 48)	16M (52%), 15F (48%)	7.4 (0.42, 56.4)
Healthy	6	38 (26, 71)	3M (50%), 3F (50%)	0.23 (0.03, 0.62)
Possible airways disease	4	46 (41, 64)	0M (0%), 4F (100%)	6.6 (1.3, 35.0)
Lung cancer	5	73 (68, 79)	4M (80%), 1F (20%)	52.6 (44.9, 69.0)
Post COVID-19	20	58 (25, 73)	18M (90%), 2F (10%)	1.36 (0.55, 5.17)
Total	170	53 (9, 79)	92M (54%), 78F (46%)	3.80 (0.03, 69.0)

8.3.2 Image acquisition

All participants underwent 3D volumetric ^{129}Xe-MRI and ^{1}H-MRI in the coronal plane with full lung coverage on a 1.5T GE HDx scanner (GE Healthcare, Milwaukee, WI, USA). ^{1}H-MRI scans were acquired with an 8-element cardiac coil (Stewart *et al.*, 2018) using a 3D spoiled gradient-recalled sequence with a repetition time/echo time of 1.8/0.7 milliseconds, in-plane resolution of ~3x3 mm^{2} and a slice thickness of 3 mm. A ~35-48cm field of view with a flip angle of 3° at a bandwidth of 166.6kHz was used. Hyperpolarised gas MRI scans were acquired using ^{129}Xe that was polarised on site to ~25% with an in-house developed rubidium spin-exchange polariser (Norquay *et al.*, 2018b). A flexible quadrature radiofrequency coil was employed for transmission/reception of MR signals at the Larmor frequency of ^{129}Xe-MRI (Clinical MR Solutions, Brookfield, WI, USA). A 3D balanced steady-state free precession sequence was used (Stewart *et al.*, 2018) with a repetition/echo time of 6.7/2.2 milliseconds, an in-plane resolution of ~4x4 mm^{2} and slice thickness of 10 mm. A ~38-40 cm field of view with a flip angle of 9° or 10° and a bandwidth of 16kHz was used.

8.3.3 Image segmentation

To facilitate ^{1}H-MRI registration, lung cavity segmentation is required to provide a region over which to perform registrations, thereby producing more accurate deformations compared to employing registration algorithms over the whole scan. ^{1}H-MRI TLC and RV scans were segmented using a CNN-based implementable ^{1}H-MRI lung segmentation network as previously developed (Astley *et al.*, 2021). Segmentations were then manually corrected by several expert observers with the following experience: B.A.T had 10 years, H.M had 7 years, P.J.C.H had 5 years, A.M.B had 5 years and J.R.A had 3 years.

8.3.4 Image registration

RV and TLC ^{1}H-MRI scans were aligned using deformable image registration and subsequently registered to the spatial domain and resolution of ^{129}Xe-MRI via a corresponding anatomical ^{1}H-MRI scan acquired at a similar inflation as ^{129}Xe-MRI (Tahir *et al.*, 2014; Tahir *et al.*, 2018). The empirically optimised script used identical rigid and affine parameters as the EMPIRE10_BSplineSyn script as previously described (Tahir *et al.*, 2014; Tahir *et al.*, 2018). Registration pipelines consisted of rigid, affine and diffeomorphic stages. For the deformable registration stage, explicit B-spline regularisation was applied to the resulting transform of the affine pipeline. A knot spacing for the update field of 65 mm provided optimal results. Additionally, a 5-level multi-resolution pyramid was used (instead of 4 levels) with down-sampling factors of 10 × 6 × 4 × 2 × 1 and corresponding smoothing Gaussian sigmas of 5 × 3 × 2 × 1 × 0 mm and the normalised correlation coefficient similarity metric with a radius of 2 voxels instead of 4. A step size of 0.2 was selected for the gradient descent optimisation algorithm. All registrations were performed using the advanced normalisation tools (ANTs) registration framework (Avants *et al.*, 2008). The registration pipeline is further described in Tahir *et al.*, (2019).

8.3.5 ^{1}H-MRI ventilation modelling

Model-based ^{1}H-MRI ventilation surrogates were computed from the aligned TLC and RV ^{1}H-MRI scans. The ^{1}H-MRI ventilation model assumes that differences in signal intensities of co-registered voxels reflect naturally occurring density variations in the lungs during breathing (Kjørstad *et al.*, 2017). Specific ventilation (SV) is a unitless quantity that models the proportion of inhaled air entering the lungs during normal breathing (Capaldi *et al.*, 2018) and is calculated from deformably registered ^{1}H-MRI RV and TLC scans as follows:

$$\mathrm{SV} = \frac{\Delta \mathrm{V}}{\mathrm{V}_{\mathrm{RV}}} \approx \frac{\mathrm{F}_{\mathrm{TLC}} - \mathrm{F}_{\mathrm{RV}}}{\mathrm{F}_{\mathrm{RV}}} \tag{8.1}$$

where $\mathrm{F}_{\mathrm{TLC}}$ and F_{RV} denote the air volume fractions at inspiration and expiration, respectively, and $\mathrm{F}_{\mathrm{AIR}}$ denotes the air volume fraction at an arbitrary inflation level. The MRI signal intensity (SI) is known to be approximately inversely proportional to the volume of air in the lung (Zapke *et al.*, 2006).

$$\mathrm{SI} \ \overset{\sim}{\propto} \ \frac{1}{\mathrm{F}_{\mathrm{AIR}}} \tag{8.2}$$

Substituting Equation (*8.2*) into Equation (*8.1*) allows the specific ventilation to be computed as follows:

$$\mathrm{SV} \approx \left(\frac{\mathrm{SI}_{\mathrm{RV}} - \mathrm{SI}_{\mathrm{TLC}}}{\mathrm{SI}_{\mathrm{TLC}}}\right) \tag{8.3}$$

where $\mathrm{SI}_{\mathrm{RV}}$ and $\mathrm{SI}_{\mathrm{TLC}}$ are voxel-wise signal intensities at RV and TLC, respectively. A median filter of 3x3x1 voxels was subsequently applied to ^{1}H-MRI ventilation maps to account for noise and registration errors, resulting in a filtered resolution of 12x12x10 mm^3 which approximately corresponds to the size of the acinus indicated by Levin *et al.* (2017).

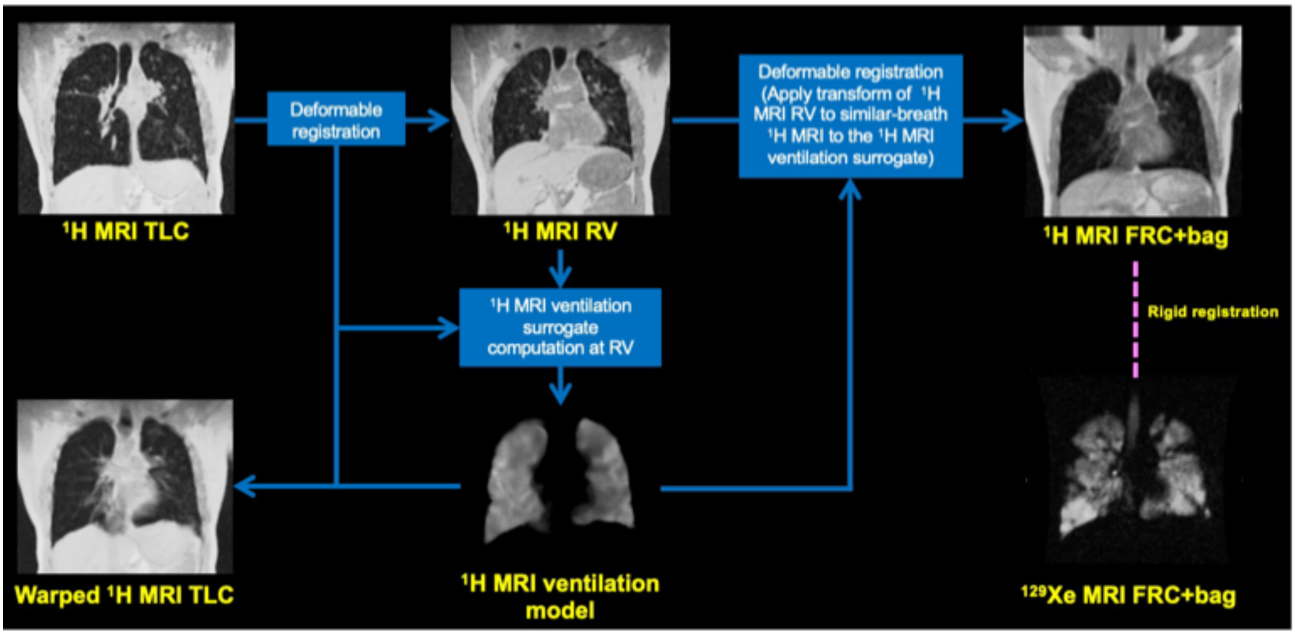

^{1}H MRI TLC
Deformable registration
^{1}H MRI RV
Deformable registration (Apply transform of ^{1}H MRI RV to similar-breath ^{1}H MRI to the ^{1}H MRI ventilation surrogate)
^{1}H MRI FRC+bag
^{1}H MRI ventilation surrogate computation at RV
Rigid registration
Warped ^{1}H MRI TLC
^{1}H MRI ventilation model
^{129}Xe MRI FRC+bag

Figure 8.1 Registration workflow for modelling ^{1}H-MRI ventilation.

8.4 Deep learning experiments and evaluation

8.4.1 CNN architecture configurations

We developed and compared three DL approaches to generate synthetic ^{129}Xe-MRI ventilation maps by varying the input images provided to the CNN. These approaches are referred to below:

1) DL (INSP + EXP + SV Model): *PhysVENeT*
2) DL (INSP + EXP)
3) DL (SV Model)

We assessed the effect of providing a physics-based ^{1}H-MRI ventilation model, alongside structural inspiratory (INSP) and expiratory (EXP) ^{1}H-MRI, as inputs to a CNN (approach 1). This approach that we call "PhysVENeT" is compared to a network which is not physics-informed (approach 2) and a network which does not integrate structural multi-inflation ^{1}H-MRI (approach 3).

For each configuration, input scans with varying dimensions were read by the network using patch-based sampling with patches of 192x192x48 voxels (Gibson *et al.*, 2018b). The VNet CNN allows for non-isotropic patch sizes in-line with the anisotropic nature of ^{129}Xe-MRI. We modified the VNet CNN architecture (Milletari *et al.*, 2016) to learn functional representations from 3D input images by outputting a 3D continuous map of regional ventilation. The CNN contained feature channels of 16, 32, 64, 128 and 256, where convolution operations are employed at each layer to both learn residual features and to reduce the resolution of the feature stack, analogous to commonly employed pooling operations. The input layer employs a convolution operation with a 5x5x5 kernel and stride of 1; two identical convolutions are employed at the second layer and three at the subsequent layers. After each 5x5x5 convolution, a subsequent 2x2x2 kernel with stride of 2 was utilised, generating non-overlapping patches, hence the resolution of the image is divided by two. This is repeated at each layer, resulting in a minimum resolution of 12x12x3 in the final convolution step. The structure of the network is replicated in deconvolution steps bar the output layer. Each convolution operation employed a PReLU non-linear activation function with valid padding. As indicated by Milletari *et al.* (2016), the CNN learns residual fine-grained features at each step which informs corresponding deconvolution operations in the upsampling side of the network (Milletari *et al.*, 2016). The VNet CNN architecture is modified to contain a regression output layer, allowing the network to generate continuous intensity maps in three dimensions. Furthermore, we employ a Huber loss function where the Huber loss ($\mathrm{H}_{\mathrm{Loss}}$) is defined as:

$$\mathrm{H}_{\mathrm{Loss}}(a) = \begin{cases} \frac{1}{2}a^2 & \text{for } |a| \leq \delta \\ \delta \cdot \left(|a| - \frac{1}{2}\delta\right) & \text{for } |a| > \delta \end{cases} \tag{8.4}$$

where a represents the difference between given co-registered voxels in the ground truth and predicted outputs and δ is defined as 0.1. The Huber loss function is expressed as a representation of either the mean square error (MSE) or the absolute value function at δ. The Huber loss has the benefit of combining the minimum-variance estimator of the MSE loss and the median-unbiased estimator of the absolute value loss to produce a loss function that alternatively provides the sensitivity and robustness of the MSE and absolute loss, respectively. This loss was utilised for synthetic ventilation generation to minimise the impact of outliers in the first stages of training and improve sensitivity once the loss has significantly reduced. For DL approaches 1 and 2, which utilise multiple input images, weight sharing was not employed, resulting in input dimensions of 192x192x48x3 or 192x192x48x2 for the PhysVENeT and other DL configurations, respectively, similar to Kläser *et al.* (2021) and Jahangir *et al.* (2022). This method combines the feature maps from spatially aligned inspiratory and expiratory ^{1}H-MRI alongside the ^{1}H-MRI ventilation model. Therefore, the network can leverage concurrent information distributed across multiple input feature maps (Berger *et al.*, 2018). The PhysVENeT architecture (approach 1) is detailed in Figure 8.2.

8.4.2 CNN training parameters

All warped and masked RV and TLC ^{1}H-MRI scans and ^{129}Xe-MRI ventilation scans underwent pre-processing before they were fed into the network; scans were normalised with image intensities between [0, 1]. Training data was augmented to reduce overfitting whilst still maintaining physiological plausibility. We used an augmentation method where the number of scans in the training set remained consistent; however, each set of input images is deformed using a random rotation and scaling factor between [-10°, 10°] and [-10%, 10%], respectively. Different rotation and scaling factors are randomly selected within these limits when the feature map is provided to the network. Thus, the network can be trained for an increased number of epochs as it is highly unlikely to be exposed to the exact same deformations in each epoch. Consequently, we train our network for 900 epochs. Batch normalisation was applied at each layer using a mini-batch size of 2 to reduce covariate shift between network layers during training (Ioffe and Szegedy, 2015). Network weights were trained from scratch and initialised using Xavier initialisation, representing a Gaussian distribution with a mean of 0 and a variance of 1/*N*, where *N* represents the number of weights and biases. The network employs Adam (Kingma and Ba, 2015) optimisation with a learning rate of $1x10^{-5}$. L2 regularisation and a decay of $1x10^{-4}$ were used to minimise overfitting. The network is trained and tested using the open source medical imaging framework NiftyNet (Gibson *et al.*, 2018b) built on top of TensorFlow 1.1.4 (Abadi *et al.*, 2016). An NVIDIA Tesla V100 GPU with 24GB of RAM was required for network training. Post-processing was conducted to account for noise and registration errors in synthetic ventilation maps; ^{1}H-MRI ventilation scans and DL-generated synthetic ventilation scans were normalised with signal intensities between [0, 1] and median filtered with a radius of 3x3x1 voxels.

8.4.3 Data split

The dataset contained scans from 170 participants. 150 participants were used for five-fold cross-validation, resulting in randomly selected training and testing sets of 120 and 30 participants, respectively, for each fold. The remaining 20 participants were used for external validation; these scans were from participants who had previously been hospitalised for COVID-19, a disease not contained within the cross-validation dataset.

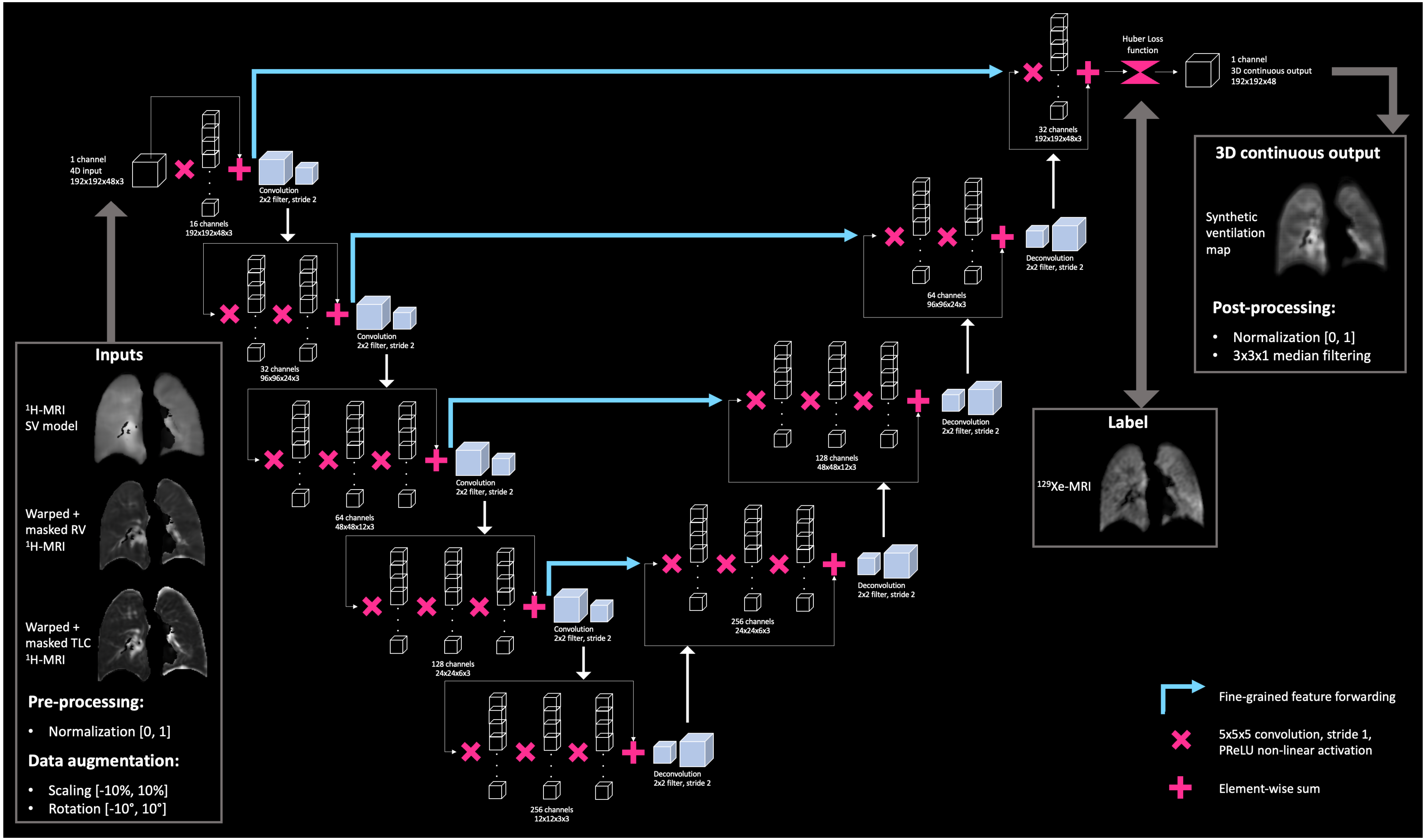


Figure 8.2 PhysVENeT architecture and training strategy.

8.4.4 Quantitative evaluation

Surrogates of ventilation were quantitatively evaluated using two common voxel-wise image synthesis metrics, namely, the voxel-wise Spearman's correlation (*rs*) and SSIM. The Spearman's *rs* was the primary evaluation metric in the CT ventilation imaging grand challenge, VAMPIRE (Kipritidis *et al.*, 2019). In a recent review of DL in pulmonary imaging, SSIM was used for evaluation in several image synthesis investigations (Astley *et al.*, 2020b). Further details of Spearman's *rs* and SSIM calculations are given in the following sections.

Spatial correlation between synthetic ventilation surrogates and corresponding ^{129}Xe-MRI scans was assessed at full resolution using Spearman's *rs*. The correlation was calculated on all voxels within the lung cavity region as defined by the lung volume in a ^{1}H-MRI scan acquired at the same inflation as ^{129}Xe-MRI. Spearman's *rs* quantifies the degree of monotonicity between any two ventilation images within a range of [-1, 1] where 1 represents a perfect positive correlation and -1 represents a perfect negative correlation; therefore, an *rs* of 0 represents no correlation.

SSIM is an image quality measure that encompasses similarity information. SSIM is calculated between non-zero voxels in the reference ^{129}Xe-MRI scan (X) and the synthetic ventilation surrogate (Y) within the lung cavity region, as defined by the lung volume in a ^{1}H-MRI scan acquired at the same inflation as ^{129}Xe-MRI, as follows:

$$\mathrm{SSIM} = \frac{(2\mu_X\mu_Y + c_1)(2\sigma_{X,Y} - c_2)}{(\mu_X^2 - \mu_Y^2 + c_1)(\sigma_X^2 - \sigma_Y^2 + c_2)} \tag{8.5}$$

where μ_X and μ_Y are the average intensities of X and Y, respectively, and σ_X and σ_Y are the variances of X and Y, respectively. $\sigma_{X,Y}$ is the covariance of X and Y. c_1 and c_2 are defined as follows:

$$c_1 = (k_1 L)^2,\ \ c_2 = (k_2 L)^2 \tag{8.6}$$

where L is the dynamic range of pixel intensities in X and Y and k_1 and k_2 are the constants 0.01 and 0.03, respectively (Wang *et al.*, 2004).

8.4.5 Statistical analysis

We initially determined whether the data was normally distributed via Shapiro-Wilk tests; if normality was not satisfied, non-parametric tests were conducted. Friedman tests with Bonferroni correction for *post-hoc* multiple comparisons were used to assess significant differences between DL approaches. For each metric, paired *t*-tests were used to assess significant differences between the DL approaches and the ^{1}H-MRI ventilation model. Wilcoxon tests were used to assess differences between folds on external validation data and differences in performance between the ^{1}H-MRI ventilation model and each fold on the external validation cohort. Statistical analyses were performed using GraphPad Prism 9 (GraphPad, San Diego, CA). In this work, a p-value of <0.05 was considered statistically significant.

8.5 Results

8.5.1 Qualitative evaluation

Figure 8.3 shows example coronal slices comparing synthetic ventilation maps with ^{129}Xe-MRI ventilation imaging for five cases within the dataset. Voxel-wise Spearman's *rs* and SSIM are given for each case and method. A number of cases show large ventilation defects which are replicated in synthetic ventilation scans generated by the PhysVENeT framework. Case 2 shows an area of consolidation in the apex of the right lung on the ^{1}H-MRI scan which manifests as a ventilation defect in the corresponding hyperpolarised gas MRI. This defect is replicated in all DL configurations and the SV model. Case 3 displays subtle ventilation defects which are somewhat replicated by several synthetic ventilation approaches. Case 4 shows a cystic fibrosis patient with large defects in both lungs which are replicated accurately by the PhysVENeT DL configuration. Case 5 displays a case with large defects at the apex of both lungs in the hyperpolarised gas MRI scan. All DL configurations and the SV model do not accurately replicate these ventilation defects, despite reduced intensity in these regions some ventilation is still present in all approaches tested.

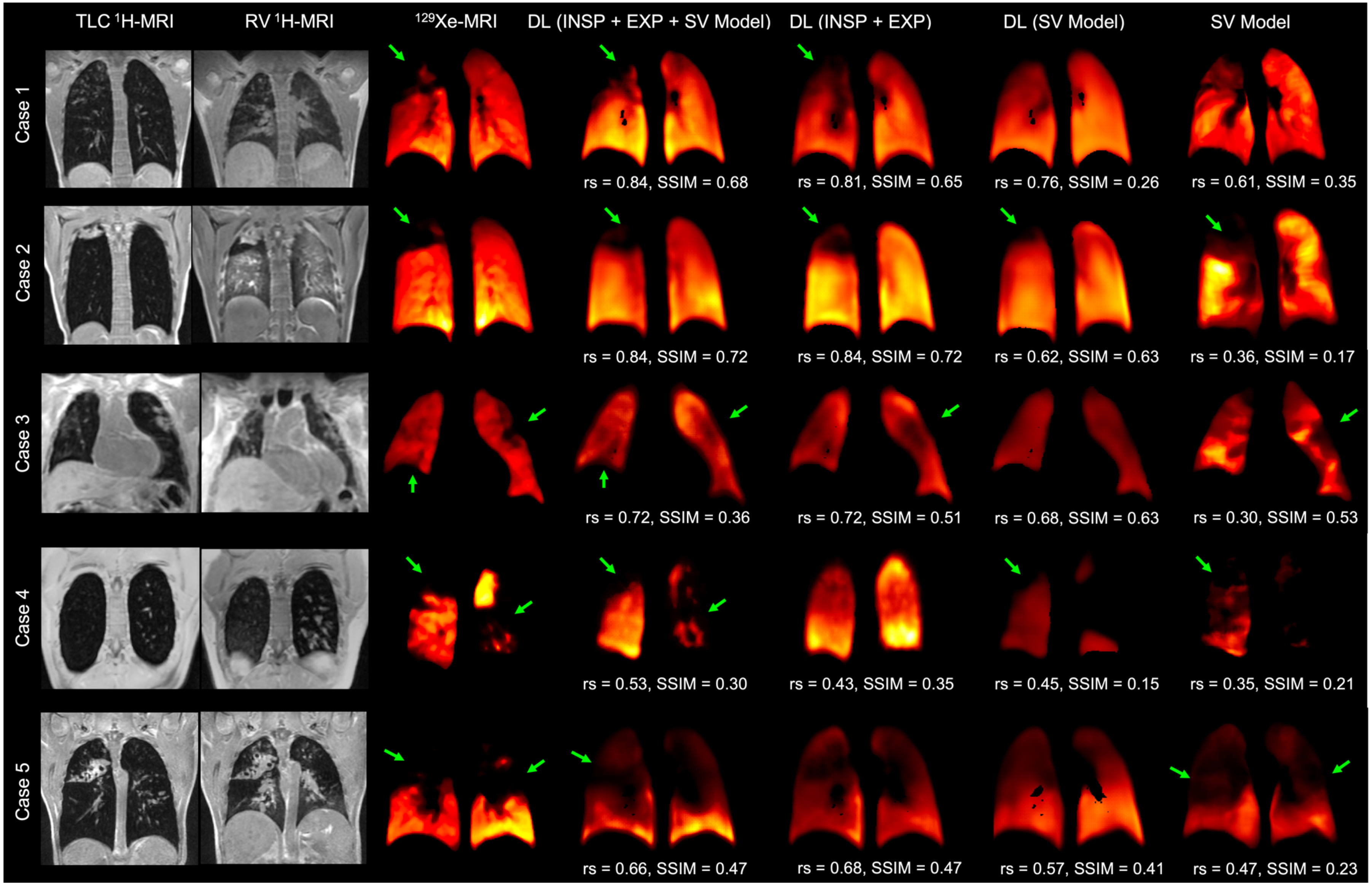


Figure 8.3 Example coronal slices of TLC and RV ^{1}H-MRI, ^{129}Xe-MRI, DL-based synthetic ventilation maps and the ^{1}H-MRI SV model for five participants in the dataset. Voxel-wise Spearman's *rs* and SSIM values are given for each DL approach and the ^{1}H-MRI ventilation model. Green arrows indicate defects which are present in hyperpolarised gas MRI and replicated in synthetic ventilation scans.

8.5.2 Quantitative evaluation

The PhysVENeT framework generated the highest Spearman's *rs*, achieving a median (range) of 0.68 (0.13, 0.85) and the DL (INSP + EXP) approach generated the highest SSIM, achieving a median (range) of 0.58 (0.14, 0.76) when compared to ground-truth ^{129}Xe-MRI ventilation. A full summary of results is provided in Table 8.2. Using inferential statistics adjusted for multiple comparisons, the PhysVENeT significantly outperformed all other DL approaches and ^{1}H-MRI ventilation modelling in terms of Spearman's *rs*. In addition, both the PhysVENeT and DL (INSP + EXP) approaches significantly outperformed the DL (Model) and ^{1}H-MRI ventilation model using the SSIM metric. No significant difference was observed between the PhysVENeT and DL (INSP + EXP) networks using the SSIM (p=0.14). The distribution of Spearman's *rs* and SSIM for each method across all images within the cross-validation dataset is displayed in Figure 8.4; significant p-values are provided. Synthetic ventilation performance for the PhysVENeT approach, stratified by pulmonary pathology, is shown in Figure 8.5.

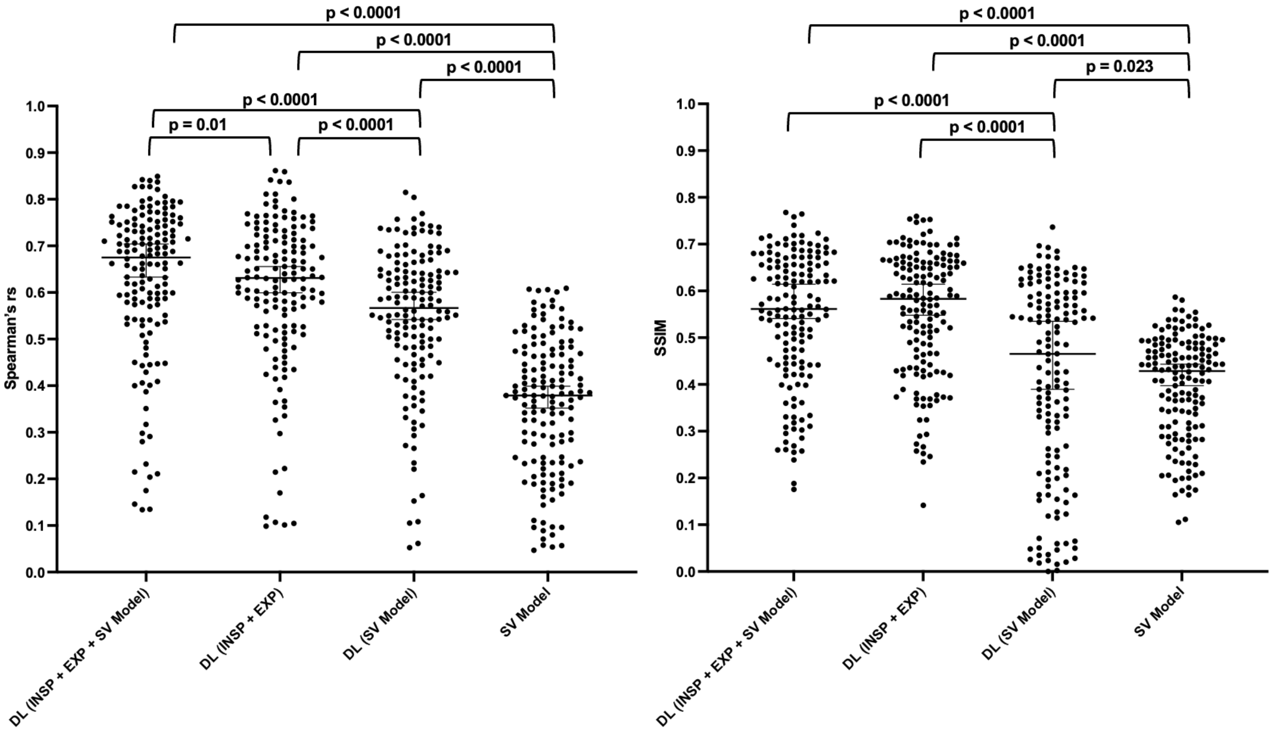


Figure 8.4 Comparison of performance for DL methods and ^{1}H-MRI SV model using the voxel-wise Spearman's *rs* (left) and SSIM (right) metrics. P-values are given for statistically significant comparisons.

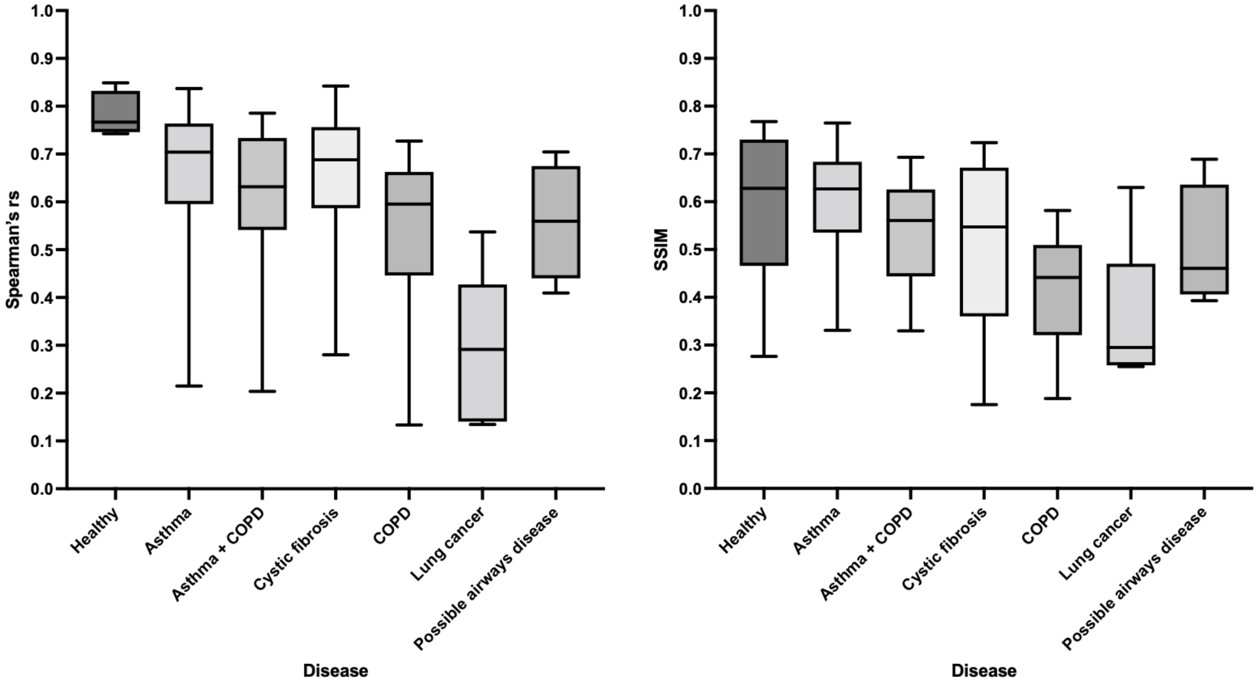


Figure 8.5 Comparison of performance stratified by participant pathology using Spearman's rs (left) and SSIM (right) metrics for the proposed PhysVENeT framework

Table 8.2 Synthetic ventilation results from the ^{1}H-MRI SV model and the three DL approaches compared to hyperpolarised gas MRI ventilation using the Spearman's rs and SSIM metrics. Median (range) is given. Metrics are given for each fold individually and the combined values across all folds.

Cross-validation	**DL (INSP + EXP + SV Model)**		**DL (INSP + EXP)**		**DL (SV Model)**		**SV Model**	
	Spearman's *rs* Median (range)	SSIM Median (range)	Spearman's *rs* Median (range)	SSIM Median (range)	Spearman's *rs* Median (range)	SSIM Median (range)	Spearman's *rs* Median (range)	SSIM Median (range)
Fold 1	**0.68 (0.13, 0.85)**	0.56 (0.19, 0.77)	0.65 (0.11, 0.86)	**0.57 (0.14, 0.76)**	0.58 (0.06, 0.77)	0.50 (0.05, 0.65)	0.37 (0.09, 0.57)	0.39 (0.11, 0.56)
Fold 2	**0.66 (0.18, 0.84)**	0.54 (0.27, 0.72)	0.60 (0.11, 0.81)	**0.55 (0.27, 0.67)**	0.58 (-0.04, 0.82)	0.38 (0.01, 0.74)	0.34 (0.05, 0.61)	0.43 (0.17, 0.56)
Fold 3	**0.67 (0.28, 0.79)**	**0.60 (0.29, 0.72)**	0.65 (0.37, 0.80)	0.59 (0.26, 0.75)	0.54 (0.22, 0.69)	0.30 (0.02, 0.64)	0.39 (0.05, 0.61)	0.43 (0.11, 0.59)
Fold 4	**0.69 (0.14, 0.83)**	0.54 (0.19, 0.70)	0.64 (0.10, 0.84)	**0.59 (0.29, 0.71)**	0.54 (0.05, 0.73)	0.55 (0.04, 0.64)	0.41 (-0.01, 0.60)	0.42 (0.16, 0.52)
Fold 5	**0.66 (0.15, 0.84)**	**0.61 (0.18, 0.76)**	0.63 (0.10, 0.77)	0.59 (0.29, 0.70)	0.64 (0.23, 0.80)	0.54 (0.00, 0.70)	0.38 (0.06, 0.61)	0.45 (0.21, 0.58)
All folds	**0.68 (0.13, 0.85)**	0.56 (0.18, 0.77)	0.63 (0.10, 0.86)	**0.58 (0.14, 0.76)**	0.57 (-0.04, 0.82)	0.47 (0.00, 0.74)	0.38 (-0.01, 0.61)	0.43 (0.11, 0.59)

8.5.3 External validation

An external validation dataset comprising 20 participants who had a pathology not present in the cross-validation dataset were used to assess the generalisability of DL approaches. The PhysVENeT framework achieved the highest Spearman's *rs* and SSIM with a median (range) of 0.62 (0.18, 0.79) and 0.58 (0.05, 0.68), respectively when averaged across all networks trained using each cross-validation fold. The proposed PhysVENeT showed minimal reduction in performance on external validation data, whereas DL approaches that were not physics-informed, or did not integrate structural imaging directly, showed significant reductions in both Spearman's rs and SSIM. Results for DL approaches are given in Table 8.3.

Table 8.3 Synthetic ventilation results on the external validation dataset (n=20) from the three DL approaches compared to ^{129}Xe-MRI ventilation using the Spearman's rs and SSIM metrics. Median (range) is given. Metrics are given for ventilation surrogates generated by each of the five folds during cross-validation and the average values across all folds.

External validation (n=20)	**DL (INSP +EXP + SV Model)**		**DL (INSP + EXP)**		**DL (SV Model)**	
	Spearman's *rs* Median (range)	SSIM Median (range)	Spearman's *rs* Median (range)	SSIM Median (range)	Spearman's *rs* Median (range)	SSIM Median (range)
Fold 1	0.62 (0.28, 0.76)	**0.58 (0.49, 0.66)**	**0.65 (0.29, 0.82)**	0.56 (0.04, 0.68)	0.53 (0.24, 0.74)	0.53 (0.02, 0.64)
Fold 2	**0.63 (0.23, 0.79)**	**0.57 (0.22, 0.65)**	0.55 (0.24, 0.71)	0.25 (0.01, 0.55)	0.56 (0.41, 0.75)	0.55 (0.03, 0.67)
Fold 3	**0.60 (0.31, 0.77)**	**0.60 (0.05, 0.66)**	0.56 (0.26, 0.73)	0.52 (0.03, 0.63)	0.41 (0.13, 0.64)	0.50 (0.01, 0.56)
Fold 4	**0.61 (0.18, 0.74)**	0.58 (0.33, 0.65)	0.58 (0.21, 0.76)	0.55 (0.04, 0.64)	0.50 (0.27, 0.75)	**0.59 (0.07, 0.66)**
Fold 5	**0.63 (0.22, 0.77)**	**0.58 (0.23, 0.68)**	0.54 (0.18, 0.76)	0.51 (0.05, 0.63)	0.60 (0.26, 0.80)	0.54 (0.03, 0.65)
Average across folds	**0.62 (0.18, 0.79)**	**0.58 (0.05, 0.68)**	0.56 (0.18, 0.82)	0.51 (0.01, 0.68)	0.49 (0.13, 0.80)	0.53 (0.01, 0.66)

Significant differences in performance of the PhysVENeT between networks trained on each cross-validation fold and tested on external validation data were observed; however, the ranges of average Spearman's *rs* and SSIM values across all folds were narrower than those of other approaches, with a Spearman's *rs* range of 0.60-0.63 and SSIM range of 0.57-0.60. Significant *p*-values between the five trained models generated by each fold in the cross-validation process are shown in Figure 8.6.

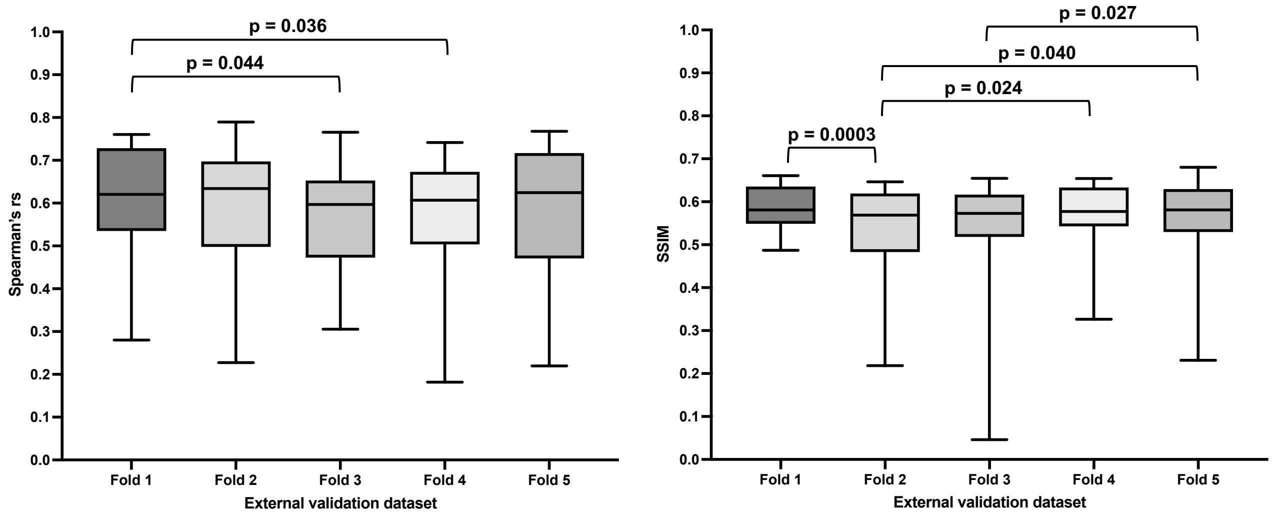


Figure 8.6 Comparison of performance on external validation data using the five trained models generated by the PhysVENeT during cross-validation in terms of Spearman's *rs* (left) and SSIM (right). Significant p-values are given.

8.6 Discussion

In this work, we propose a framework for the generation of synthetic ventilation surrogates from multi-inflation structural ^{1}H-MRI and a physics-based SV model. The PhysVENeT approach integrates computational modelling and DL to produce physics-informed 3D ventilation maps of the lungs. These synthetic ventilation images correlate with ^{129}Xe-MRI in a voxel-wise manner and can mimic gross ventilation defects across a range of pathologies. Generating 3D synthetic ventilation surrogates from structural imaging modalities, without the requirement of specialised equipment or exogenous contrast, can reduce barriers in the widespread adoption of cutting-edge functional lung imaging modalities, such as hyperpolarised gas MRI.

Synthetic ventilation surrogates generated by the PhysVENeT framework significantly outperformed ^{1}H-MRI ventilation maps generated through computational modelling of specific ventilation change. This was demonstrated using the voxel-wise Spearman's *rs* and SSIM metrics calculated across the whole lung region where the PhysVENeT achieved a Spearman's *rs* of 0.68 and an SSIM of 0.56 on the cross- validation dataset. Furthermore, the PhysVENeT significantly outperformed other DL approaches which did not leverage structural ^{1}H-MRI or physics-based ^{1}H-MRI ventilation modelling, using Spearman's *rs*. When inference was conducted on external validation data, the PhysVENeT exhibited increased performance compared to other DL approaches, achieving a Spearman's *rs* of 0.62 and an SSIM of 0.58. The inclusion of both structural ^{1}H-MRI and computational modelling of specific ventilation provides PhysVENeT with the ability to generalise effectively to participants of a previously unseen disease. The increase in generalisability on external validation data, in conjunction with significant increases in correlations on cross validation data, indicates the benefit of using a physics-informed framework.

We used a large dataset that contained 170 participants with numerous pulmonary pathologies and varying degrees of lung function, as measured by the ventilation defect percentage (VDP) (Table 8.1). 150 of these participants were used for five-fold cross validation, leading to five separately trained networks, each tested on 20% of the total cohort. The remaining 20 participants were used for external validation whereby each of the five separately trained networks were used to generate ventilation surrogates for these 20 participants. The physics-informed PhysVENeT framework performed similarly on both the cross-validation and external validation datasets. In addition, the range of SSIM and Spearman's *rs* metrics on the external validation data is much narrower than the other DL approaches. Therefore, by leveraging structural ^{1}H-MRI and physics-based modelling, the PhysVENeT framework exhibits minimal overfitting and is largely generalisable to scans outside the cross-validation dataset.

The framework uses a VNet CNN backbone previously developed for 3D segmentation tasks (Milletari *et al.*, 2016). We adapted the VNet with a Huber loss function to output 3D continuous intensity maps and through the integration of a multi-channel input configuration. The CNN architecture makes use of additional convolution operations to reduce the dimensionality of the image instead of traditional pooling methods. This limits the footprint of the network, reducing the memory consumption (Springenberg *et al.*, 2014). In turn, this facilitates the use of large anisotropic 3D patch sizes. An additional feature of the network architecture is the ability to use anisotropic input dimensions; ^{129}Xe-MRI scans have an anisotropic resolution with an in-plane resolution of ~4x4 mm^{2} and a slice thickness of 10 mm. Thus, we make use of the anisotropic input capabilities of the VNet architecture in contrast to other architectures which require isotropic spatial windowing, such as the nn-UNet (Isensee *et al.*, 2018).

Relevant differences between ^{129}Xe-MRI scans, the PhysVENeT and the SV model are observed for several example cases. In general, synthetic ventilation scans show reduced contrast between ventilated and non-ventilated regions than their corresponding hyperpolarised gas MRI scans where synthetic ventilation scans often contain some areas of signal even within regions of ventilation defects. The SV model performs well in some instances but poorly in others; this is potentially due to registration errors, but the accuracy of the SV model is also limited by the signal intensity model utilised. It is hypothesised that lung ventilation cannot be solely captured by differences in signal intensities as other components likely contribute to ventilation. Therefore, in cases where the SV model performed poorly, the PhysVENeT also exhibited worse than average performance. Regardless, the PhysVENeT's ability to modify ventilation surrogates, in combination with information from multi-inflation structural imaging, leads to improved performance in most cases within the dataset. When comparing the PhysVENeT to the DL configuration which only utilises multi-inflation ^{1}H-MRI, there are some instances where the SV model underperformed and, consequently, the PhysVENeT also underperformed. However, as demonstrated, the PhysVENeT produced significantly more accurate synthetic ventilation scans, particularly on external validation scans. It is hypothesised that the inclusion of the SV model as an input increased the generalisability of the PhysVENeT to pulmonary pathologies previously unseen by the dataset compared to the configuration which utilises only multi-inflation ^{1}H-MRI as an input.

Previous approaches have utilised DL to generate synthetic ventilation maps in 2D. Capaldi *et al.* (2020) used a 2D UNet CNN with a MAE loss function to generate ventilation maps of a single 2D coronal section from free-breathing ^{1}H-MRI, limiting volumetric coverage (Capaldi *et al.*, 2020). Moreover, the 2D intensity maps cannot contextualise the volumetric nature and spatial clustering of ventilation defects (Donovan and Kritter, 2015). This can lead to discontinuities between slices which reduces the plausibility of ventilation defect patterns in DL-based ventilation surrogates. Here, we generate fully-volumetric synthetic ventilation surrogates in three dimensions which allows the proposed CNN to learn features which occur over multiple slices.

Levin *et al.* (2017) has indicated that the resolution of functional lung images need not be higher than the smallest pulmonary gas exchange unit, namely, the acinus. The acinus is approximately 10x10x10 mm^{3} in adult humans. They also report that the sufficient resolution of ventilation images can be as low as 20x20x20 mm^{3} due to the spatial clustering of many ventilation defects (Levin *et al.*, 2017). Consequently, we apply 3x3x1 median filtering as a post-processing step to ^{129}Xe-MRI, ^{1}H-MRI ventilation surrogates, and DL-based synthetic ventilation images before evaluation. This increases the resolution to 12x12x10 mm^{3}, in-line with appropriate resolutions proposed by Levin *et al.* (2017).

Contrast-based functional lung imaging modalities such as hyperpolarised gas MRI require specialised equipment and exogenous contrast, which limit clinical adoption. In addition, non-contrast functional lung imaging techniques such as CTVI require exposure to ionising radiation and have demonstrated large variability in performance. Therefore, the ability to generate 3D ventilation scans from structural non-contrast ^{1}H-MRI scans has wide-reaching implications for functional lung imaging, including the potential to be used for functional lung avoidance radiotherapy and in a triaging capacity for instances where contrast-based functional lung imaging is unavailable.

8.6.1 Limitations

Despite significant improvements in Spearman's *rs* and SSIM when compared to ^{1}H-MRI ventilation modelling, the PhysVENeT framework generated only moderate correlations with ^{129}Xe-MRI. Synthetic ventilation surrogates were unable to accurately replicate all subtle ventilation defects, and, in some cases, they exhibit minimal correlation. As ^{129}Xe-MRI is a direct measure of gas distribution, it can accurately quantify regional ventilation; this ability is diminished in synthetic ventilation surrogates where the ability to accurately discern between ventilated and non-ventilated lung regions is reduced. In addition, accurate registration is also required for the generation of ventilation surrogates and, therefore, the quality of these registrations significantly impacts the performance of the proposed approach. In future work, an approach independent of registration could be considered. Other DL approaches that utilise generative adversarial networks (GAN) or vision transformers (ViT) have been used for image synthesis applications (Goodfellow *et al.*, 2014; Shamshad *et al.*, 2022). The proposed framework used a fully convolutional network that lacks the unsupervised learning benefits of GANs and the long-range feature extraction of ViTs. Future investigations could indicate that utilising these methods over traditional CNNs leads to improved performance.

The dataset used in this work, whilst varied in pathologies and demographics, is limited in MRI acquisition parameters; all scans were acquired on the same scanner at the same field strength from a single centre. Thus, the conclusions of this work cannot be appropriately extended to a dataset of differing sequence or field strength without further investigation. A further limitation related to the dataset is the uneven distribution of pathologies; for example, the dataset contains only five participants with lung cancer. Consequently, when stratified by disease, the Spearman's *rs* and SSIM are lowest for lung cancer patients. This may also be due to the large VDP values present in this cohort, which often lead to increased domain-shift between structural and functional imaging (Tahir *et al.*, 2014). Nevertheless, further expansions of the dataset should focus on increasing the number of participants with pathologies that are underrepresented in the current dataset and the inclusion of a diverse range of MRI acquisition parameters to increase generalisability.

8.7 Conclusion

In this study, we propose a multi-channel CNN to synthesise 3D pulmonary ventilation maps from multi-inflation ^{1}H-MRI. These structural scans are combined with a physics-informed computational model of specific ventilation to enhance the physiological plausibility of the synthetic ventilation surrogates. The PhysVENeT framework produces ventilation maps which correlate with ^{129}Xe-MRI, reflecting ventilation defects observed in the real scans.

Chapter 9
Novel contributions and future research directions

Novel contributions of this work are summarised in section 9.1. Potential clinical applications are discussed in section 9.2. Future research directions are discussed in section 9.3.

9.1 Novel contributions

The aim of this thesis was to explore the role of deep learning (DL) in structural and functional lung imaging. This involved the application of DL approaches to image segmentation and image synthesis, outlined in **Chapters 4-8**.

Chapter 4 investigated the automatic segmentation of ventilated lung in hyperpolarised gas MRI scans, comparing several 3D convolutional neural networks (CNNs). The dataset used for this work is the largest reported to date for hyperpolarised gas MRI segmentation; it contains 759 ^{3}He and ^{129}Xe hyperpolarised gas MRI scans from 341 participants with a wide range of pulmonary pathologies. The proposed network showed increased performance when compared to conventional machine learning approaches, such as spatial fuzzy c-means, as well as outperforming other DL-based hyperpolarised gas MRI segmentation algorithms in the literature. This resulted in the publication of a long paper at the MICCAI 2020 thoracic image analysis workshop and an original article in *Nature Scientific Reports*.

Chapter 5 involved the development of a DL-based segmentation algorithm to delineate the lung parenchyma in ^{1}H-MRI, demonstrating the generalisability of a DL approach when a diverse, representative dataset is utilised. A uniquely large, multi-centre, multi-vendor, multi-disease ^{1}H-MRI dataset, which contained 809 scans from 289 participants, was used. The proposed CNN was capable of generating accurate lung segmentations on scans acquired with a wide range of acquisition parameters; furthermore, this network outperformed machine learning approaches on external data from several other centres. A comparison of

2D and 3D CNNs was performed, where the 3D CNN significantly outperformed its 2D analogue; this is the first study to investigate the effect of dimensionality in DL lung MRI applications. The 3D CNN showed no overfitting and exceptional generalisability despite potential domain-shifts in the scans used for external validation, thereby indicating that the network can learn features independent of MRI acquisition parameters or participant disease.

In **Chapter 6**, a dual-channel 3D CNN was developed to estimate the lung cavity in the domain of functional, hyperpolarised gas MRI. To this end, we used a paired dataset of hyperpolarised gas MRI and ^{1}H-MRI, containing a range of pulmonary pathologies. In contrast to other ^{1}H-MRI segmentation works, the inclusion of structural and functional modalities in a dual-channel configuration allows the network to adapt the lung cavity estimation (LCE) with reference to functional features. This is the first study to automatically generate LCEs, demonstrating significant improvements over single-channel alternatives. This resulted in the publication of an original, first-author article in the *Journal of Magnetic Resonance Imaging*.

In **Chapter 7**, a hybrid DL and model-based framework was developed to synthesise hyperpolarised gas MRI from non-contrast multi-inflation CT. The method drew inspiration from works in other fields that combine computational modelling and DL to improve physiological plausibility. The network generates fully-volumetric ventilation maps that accurately reflect gross ventilation defects observed in hyperpolarised gas MRI scans. This study represents the first attempt to synthesise hyperpolarised gas MRI ventilation scans directly from non-contrast, multi-inflation CT.

Analogous to the synthesis of hyperpolarised gas MRI from CT, **Chapter 8** investigated the synthesis of hyperpolarised gas MRI from non-contrast, multi-inflation ^{1}H-MRI. Extensive validation is employed, with 150 participants used for five-fold cross validation and 20 participants for external validation. The proposed framework generates 3D volumetric ventilation surrogates using a physics-informed, multi-channel CNN. A previous work in the literature had synthesised a specific 2D coronal section of hyperpolarised gas MRI; however, this work represents the first study to synthesise hyperpolarised gas MRI fully-volumetrically.

9.2 Potential clinical applications

Potential clinical applications are divided into the two main image analysis tasks investigated in this work, namely, image segmentation (9.2.1) and image synthesis (9.2.2).

9.2.1 Segmentation

Hyperpolarised gas MRI facilitates the visualisation and quantification of regional lung ventilation with high spatial resolution within a single breath. Quantitative biomarkers derived from this modality, including the ventilation defect percentage (VDP) and coefficient of variation (CoV), provide further insights into regional ventilation. To enable the computation of such biomarkers, segmentation of ventilated regions of the lungs is required. DL-based automatic segmentation networks developed in **Chapter 4** have the potential to dramatically reduce the editing time required to correct ventilated lung segmentations; this can allow for greater clinical throughput and wider adoption of hyperpolarised gas MRI as a functional imaging modality. Trained models developed in this work have been used to generate semi-automatic ventilated lung segmentations for patients referred for clinical scans and patients in several prospective studies at The University of Sheffield.

Segmentation of the lungs in ^{1}H-MRI is required to delineate the lung cavity from other nearby features; these segmentations are used for numerous applications, including in disease characterisation, treatment planning and longitudinal assessment. Functional lung imaging modalities such as quantitative dynamic contrast-enhanced perfusion MRI and oxygen-enhanced ventilation MRI also require ^{1}H-MRI lung segmentations. Furthermore, surrogates of ventilation, such as ^{1}H-MRI ventilation modelling, require the segmentation of the lung parenchyma at different inflations. The generalisable ^{1}H-MRI lung segmentation model developed in **Chapter 5** can generate accurate lung segmentations across a range of centres, MRI sequences, field strengths and resolutions. The ability to rapidly produce lung cavity segmentations can greatly reduce the time required for manual editing, leading to a more streamlined lung imaging workflow. Due to the generalisability demonstrated by the proposed model, it can be applied to ^{1}H-MRI scans acquired with various MRI parameters across multiple centres; therefore, the potential clinical applications are vast. To this end, the trained model has been provided to other researchers to facilitate higher clinical throughput, ultimately leading to increased clinical translation. Within the POLARIS group at The University of Sheffield, the proposed generalisable network has been used for the semi-automatic segmentation of the lungs in the quantitative evaluation of biomarkers derived from ^{1}H-MRI Ultra-short echo time (UTE) scans in interstitial lung disease patients, as well as the computation of ^{1}H-MRI ventilation surrogates.

In **Chapter 6**, LCEs were generated automatically using a dual-channel CNN. This region represents the lung cavity in the spatial domain of functional hyperpolarised gas MRI and is required for the accurate computation of functional lung imaging biomarkers such as the VDP and CoV. LCEs often require extensive manual editing; using clustering approaches, LCEs required approximately 1.5 hours per scan to edit. In this work, a dual-channel CNN using hyperpolarised gas MRI and ^{1}H-MRI as inputs was proposed; this dual-channel CNN may eliminate this editing time or could at least drastically reduce it. In addition, the network inferred LCEs in approximately 30 seconds using a single graphical processing unit (GPU). Removing the manual editing step could allow for a more streamlined workflow to generate automatic VDP values, in turn, leading to a vast reduction in the time taken to generate VDP values and, consequently, greater adoption of hyperpolarised gas MRI. The proposed network has been utilised to generate LCEs for asthma and chronic obstructive pulmonary disease (COPD) patients in the NOVELTY study, whereby the segmentation network dramatically reduced the time to calculate clinical biomarkers in a research setting.

9.2.2 Synthesis

Surrogates of ventilation that are computed from multi-inflation CT, known as CT ventilation imaging (CTVI), have been proposed; in **Chapter 7**, a hybrid framework is developed which combines CTVI modelling with non-contrast, multi-inflation CT imaging to generate synthetic hyperpolarised gas MRI scans. As stated previously, hyperpolarised gas MRI requires specialised equipment which limits clinical uptake; the ability to generate synthetic ventilation scans from multi-inflation CT without exogenous contrast or specialised equipment allows for increased clinical adoption of functional lung imaging. CT is an integral part of almost every clinical lung imaging workflow and hence is readily available for most patients; therefore, the ability to generate synthetic ventilation surrogates has implications for several clinical applications, including functional lung avoidance radiotherapy and treatment response mapping.

In a similar vein to the synthesis of regional ventilation using multi-inflation CT, we employed multi-inflation ^{1}H-MR imaging to generate 3D ventilation surrogates, as detailed in **Chapter 8**. These synthetic ventilation images are derived from non-contrast, multi-inflation ^{1}H-MRI without the requirements of specialised equipment and exogenous contrast that is traditionally required for hyperpolarised gas MRI. CT-based ventilation surrogates, such as CTVI, require exposure to ionising radiation, limiting their applicability to longitudinal scanning applications or applications involving paediatric patients. In contrast, synthetic ventilation maps derived from ^{1}H-MRI do not expose participants to ionising radiation, expanding the number of potential applications for the synthetic ventilation modality with wide-reaching implications for functional lung imaging.

It is envisioned that the role of DL-based synthetic ventilation imaging is not to directly replace functional imaging modalities, such as hyperpolarised gas MRI, due to their reduced sensitivity to subtle defects and lack of interpretability. Rather, synthetic ventilation surrogates could, in future applications, be used to triage patients for functional imaging depending on the predicted ventilation derived from synthetic scans. Hyperpolarised gas MRI is limited to only one centre within the UK and, therefore, the demand for scanning is high; triaging patients based on whether the predicted scan shows large ventilation defects has the potential to reduce this demand.

9.3 Future research directions

9.3.1 Multi-centre evaluation and federated learning

A large consideration in DL is the idea of domain shift, closely related to overfitting. This occurs when there are salient differences between training and testing set scans; if the distribution between the two sets does not substantially overlap, performance is frequently reduced when tested on data from a different disease or different centre (Aggarwal *et al.*, 2021). In Chapter 5 a generalisable ^{1}H-MRI lung segmentation network was proposed and subsequently validated on data from two external centres. However, all other investigations in this work use data solely acquired from one imaging centre, reducing the generalisability of the conclusions reached. For several chapters, investigations can be further developed using external validation datasets; evaluation of ongoing clinical research studies acquired at different centres will provide a further enhancement of the LCE work detailed in Chapter 6 presented in a subsequent research article. Furthermore, hyperpolarised gas MRI segmentation algorithms can be deployed on multi-institutional datasets providing further validation. Multi-centre validation of the proposed hyperpolarised gas MRI segmentation and LCE networks is likely to increase clinical translation of segmentation algorithms and consequently, external multi-centre validation should be a priority for future investigations.

Multi-centre validation, or centralised federated learning, has a key drawback whereby there is a requirement for data to be transferred between imaging centres, possibly conflicting with established data agreements or ethical protocols. Therefore, it is pertinent to consider the concept of decentralised federated learning; this approach envisages each centre as a node that coordinates with other nodes to generate a singular model (Adnan *et al.*, 2022). This means that no data is transferred off-centre and that there is no centralised server, thereby eliminating the single point of failure associated with centralised systems. Decentralised federated learning has the potential to expand the impact of DL in lung imaging. For example, each hyperpolarised gas MRI centre worldwide could act as a decentralised node, consistently updating model weights and parameters to build a fully global hyperpolarised gas MRI segmentation network, without the requirement for any individual centre to make their data publicly available.

9.3.2 Unsupervised learning

This thesis primarily explored supervised learning techniques, where a ground-truth delineation or ventilation scan is provided to the network for image segmentation or synthesis tasks, respectively. In contrast, unsupervised learning techniques require no ground-truth label and, instead, encode reoccurring patterns and features without specific direction. In recent years, an unsupervised learning approach called generative adversarial networks (GAN) has gained prominence in the medical imaging domain (Goodfellow *et al.*, 2014; Kazeminia *et al.*, 2018). GANs consist of two neural networks, a generator and a discriminator, whereby the generator produces an image from a noise vector and the discriminator aims to determine whether the output is ‘real’ or produced by the generator. These networks compete against each other in a zero-sum game until the discriminator can no longer accurately predict if the image produced by the generator is real or fake. This means that labels do not need to be provided to the network before training; in the medical imaging field, this eliminates the requirement for large-scale expert labelling which is often expensive and time-consuming. In several applications, unsupervised learning approaches have demonstrated improved performance over their supervised counterparts (Wilmet *et al.*, 2021). In chapters 7 and 8, CNNs were used to generate synthetic ventilation scans; however, this can be formulated as an unsupervised learning problem using a GAN. Conditional GANs (cGANs) proposed by Mizra and Osindero use prior information to constrain the network by providing both random noise and useful priors as inputs (Mirza and Osindero, 2014). cGANs have the potential to be used for image synthesis tasks where, for example, information regarding lung shape and airway structure could be provided as information priors. This may be particularly useful in future structural-to-structural investigations, such as a synthesis network which aims to transform a low-field 0.5T MRI scan to a more routinely acquired 1.5T MRI scan. CycleGANs proposed by Zhu *et* al. aim to discover underlying relationships between image modalities by finding overlapping features between unpaired images (Zhu *et al.*, 2017). In structural-to-functional synthesis investigations such as those developed in Chapters 7 and 8, this approach may increase the robustness of synthesised outputs. In future investigations, this approach may be utilised for regional ventilation synthesis and compared to multi-channel supervised learning methods.

9.3.3 Vision transformers

Originally developed for natural language processing (NLP), transformers implement attention mechanisms that are able to factor in dependencies regardless of the distances between them in the sentence (Vaswani *et al.*, 2017). Traditionally, CNNs have been used for imaging applications due to the overwhelming memory requirements needed for fully connected networks; transformers also suffer from this challenge. CNNs use the convolution operation on a local receptive field but lack the ability to model long-range dependencies. However, recent developments have aimed to combine the benefits of CNNs and transformers into a single model, known as a vision transformer (ViT) (Dosovitskiy *et al.*, 2020). These methods work by splitting images into a series of patches and feeding them in sequence to the ViT, similar to how words are treated in NLP tasks. The attention mechanism of transformers has no assumptions of locality or equivalence so long-range dependencies can be modelled between image patches, unlike with CNNs. Transformer-based architectures, such as the Attention-UNet and the UNETR, have been utilised for various medical imaging applications (Oktay *et al.*, 2018; Hatamizadeh *et al.*, 2022). A recent review demonstrates the breadth of medical imaging applications where transformers have been employed; this includes the use of ViTs for image synthesis (Shamshad *et al.*, 2022). ViTs can be deployed in the majority of chapters within this thesis; this thesis focuses on CNN-based solutions to segmentation and synthesis applications which have demonstrated promising results thus far. In future investigations, ViTs can be compared to the CNN-based solutions proposed here to generate improved ^{1}H-MRI segmentations across various acquisition protocols, for example. For image synthesis problems, ViTs can be compared to GANs and CNNs on a cross-validation dataset, then subsequently validated on an external validation dataset. ViTs also allow for increased explainability, which may be particularly useful in synthesis applications, and will be further discussed in section 9.3.4. Despite their success in recent years, transformers often require large datasets to improve upon state-of-the-art CNN-based methods, something uncommon in the medical imaging field. In Chapter 8, a larger version of the external validation dataset can be employed for network training to increase the size of the dataset, potentially facilitating the use of ViTs, and variations thereof, for the synthesis of hyperpolarised gas MRI scans from multi-inflation ^{1}H-MRI scans. As large public datasets of pulmonary MR images become available, the efficacy of transformers is expected to surpass that of the previously dominant CNN.

9.3.4 Demystifying deep learning

DL maps a multidimensional and immensely complex function between some input and some output domain. DL is optimal for complex functions as it allows for greater degrees of freedom than conventional modelling through an end-to-end learning approach. End-to-end learning approaches limit the need for human selected feature extraction. However, a trade-off between function mapping accuracy and interpretability has now occurred. By removing the feature extraction step and, instead, letting this be part of the function mapped by DL, we have now lost the ability to categorically determine which features are being used to produce the output. This has extensive implications in medical imaging where, understandably, clinicians and patients alike, require robust and interpretable data with which to make decisions. In the case of traditional DL, the clinician would not be able to explain why the CNN has predicted large ventilation defects and rather only that it has predicted such defects.

Due to the aforementioned challenges, the medical imaging community has seen increased development of what are referred to as explainable or interpretable artificial intelligence (AI) techniques. These are somewhat catchall terms that refer to a series of steps taken by researchers to increase the transparency of a DL model (Jin *et al.*, 2022). Explainability is often achieved through visualisation of activation maps at each level of the network (Pennisi *et al.*, 2021). In the case of CNN-based lung segmentation, at each layer of the network, as the dimensionality of the image is reduced, the feature maps from each layer will transition from horizontal lines to more abstract features such as the surface curvature of the lungs. Networks can be developed that extract these activation maps, allowing researchers to visualise the activations of the network. With this information, it is possible to determine which parts of the image are most activated and, therefore, which parts contribute most to the network output. In Chapters 7 and 8, these activation maps could be employed to provide ‘explanations’ as to which part of the structural images are most important for synthesising hyperpolarised gas MRI scans, thereby increasing the interpretability of proposed synthesis workflows. This is particularly important in synthesis applications and, therefore, the incorporation of activation maps in these works represents an ideal application for explainable deep learning techniques in future investigations. Activation maps can be useful in some cases; however, in other cases, usefulness is diminished where high dimensional activations may not be conducive to human intuition. ViTs give high resolution feature maps which are often more explanatory than those generated by CNNs (Shamshad *et al.*, 2022). In future investigations, if ViTs are used for image synthesis applications, these high-resolution feature maps would provide significant insight into the decision making of networks developed for the synthesis of hyperpolarised gas MRI from structural imaging, described in Chapters 7 and 8. Increasing the interpretability of DL segmentation and synthesis algorithms will likely increase the clinical uptake and translation of DL networks proposed in this thesis, providing an understanding of how synthetic scans are generated and the location of potential errors in these scans.

It is important to note that despite the development of ‘explainable’ DL, there is a fundamental paradox between explainability and DL. The reason DL often outperforms conventional modelling is precisely because of its ‘black box’ nature. For example, if we had an accurate model of specific ventilation then there would be no need to employ DL; however, as we do not have a comprehensive understanding of the relationship between parenchymal density and regional lung function, we require DL to model this function in a more complex form. Whilst the explainable AI approaches outlined above aim to interpret how the model generates outputs, the strength of DL lies squarely in its inscrutability.

9.3.5 Ethics of AI: opinions on medical applications

This thesis focuses on the development and evaluation of DL techniques in the domain of pulmonary imaging; however, there are also philosophical and ethical issues that must be considered when deploying DL in a clinical setting. Here, the author poses a series of ethical considerations that must be addressed for AI to reach its full potential in the medical imaging domain.

Firstly, there are concerns over privacy. It is an open question as to who should own the images used in DL investigations. Traditionally, images are owned by the medical institution that acquired the scan and can be used in accordance with the data usage agreement consented to by the patient; however, as personalised medicine becomes more common, there are increasing demands from patients to have greater control over their own data (Wetzels *et al.*, 2018). It is possible that, in the future, patients will be able to control their own data through a mobile phone application and grant permission personally. This will pose challenges for DL as there is a requirement for large publicly available datasets that may be limited if patients have this ability. In addition, there are privacy issues related to the anonymisation of patient images. Previous studies have demonstrated the ability to recreate images by ‘reverse engineering’ trained neural networks (Liu *et al.*, 2021b). This has implications for patient privacy, as, if these scans can be recreated, there is the potential for external adversarial actors to identify patients and use this information for nefarious purposes. As AI becomes more common in medical imaging, these concerns must be addressed to facilitate increased clinical adoption.

Secondly, the medical imaging community must reckon with the prevalence of biases in DL networks. It has been well established that due to the data-driven nature of DL techniques, they are susceptible to biases in the data provided to the network (Seyyed-Kalantari *et al.*, 2021). This can be somewhat mitigated by providing demographic information in clinical research papers, so that potential readers can assess their implications; however, for large-scale applications where DL methods are being deployed across numerous centres, particularly in other countries, these biases must be considered. Research papers have developed DL approaches to classify melanomas as cancerous or not; however, other researchers have warned of disparities, particularly when this was applied to patients with increased melanin performance, leading to serious concerns over bias (Adamson and Smith, 2018). In theory, this can be addressed, but practically this is challenging. This is because humans have inherent biases and these biases will be transferred to the data produced by these humans, meaning that in most cases some biases will remain. Consequently, it behoves the medical imaging community to remain vigilant. Vigilance can take the form of high-quality scientific research papers that are well validated and deployed only within their respective limitations. The author believes that concerns over biases will become one of the main issues that limit clinical translation in the near future.

Lastly, we must consider the potential issues with liability and associated regulation. Who is to blame if a DL network makes an incorrect classification or erroneous segmentation? Does the liability lie with the radiologist, the DL researcher, or the AI algorithm itself? These are questions which remain undecided and require extensive government regulation to answer. The more trust given to such algorithms, the less a specific person will be directly responsible for the outcomes. For example, in Chapters 7 and 8, methods for synthesising hyperpolarised gas MRI ventilation scans were developed; in this case, who would be responsible if the synthetic ventilation scans incorrectly predicted no ventilation defects, thereby affecting the care provided to patients? The author believes that these concerns over liability will limit the adoption of fully automatic DL methods in medical imaging applications. Due to differences in laws and government regulations between countries, the cross-border application of DL techniques will be significantly affected. Furthermore, in more general terms, do we as a populous, trust algorithms to make such decisions? From a utilitarian perspective, if the DL model improves upon human performance, then the answer would surely be to utilise this improvement; however, from a different ethical perspective, it may be justified to have a worse outcome in order to maintain a human in the loop. These are questions which do not fall within the domain of medical imaging research, but, instead, the domains of philosophy and ethics; a concerted effort by experts in these fields is required to provide adequate answers.

Appendix A
Computation of evaluation metrics

This section provides the code used for the computation of image segmentation (B.1) and synthesis (B.2) evaluation metrics. Evaluation metrics were calculated in Python 3 from documented libraries and validated against alternative libraries to calculate the same metric.

A.1 Segmentation evaluation metrics

The script provided below is used to calculate the following segmentation evaluation metrics:

- Directional Average Hausdorff Distance (mm)
- Directional Max Hausdorff Distance (mm)
- Dice Similarity Coefficient
- Jaccard Overlap
- Volume similarity
- Relative error metric (XOR) on region
- Relative error metric (XOR) on boundary
- Directional Boundary % Hausdorff Distance (mm)
- Directional Boundary Average Hausdorff Distance (mm)

The script can be run using the arguments `-gt`, `-p` and `-%` representing the ground truth path, output path and Hausdorff percentage, respectively. The script outputs three files, namely, a list of images (`image_list.csv`), a table of values in the same order as `image_list.csv` (`output.csv`) and a summary of all evaluation metrics (`results_summary.csv`).

A.1.1 Code

```
import numpy as np
import os
import SimpleITK as sitk
import csv
import pandas as pd
import scipy.spatial
import argparse
import pathlib
import sys

from pathlib import Path

"""
This script generates 9 evaluation metrics for segmentation for any images in the
directories provided. The directories should contain the corresponding images to compare
with the same origin and voxel spacing. Images should be .nii.gz or .mha

    - Directional Average Hausdorff Distance (mm)
    - Directional Max Hausdorff Distance (mm)
    - Dice Similarity Coefficient
    - Jaccard Overlap
    - Volume similarity
    - Relative error metric (XOR) on region
    - Relative error metric (XOR) on boundary
    - Directional Boundary % Hausdorff Distance (mm)
    - Directional Boundary Average Hausdorff Distance (mm)
"""

########
parser = argparse.ArgumentParser()
parser.add_argument('-gt','--gt_root', help="ground truth root", type=Path, required=True)
parser.add_argument('-p','--pred_root', help="predicted root", type=Path, required=True)
parser.add_argument('-%','--percents', help="percent hausdorff", default=95)
args = parser.parse_args()
X = str(args.gt_root)
Y = str(args.pred_root)
gtpath = ""+X+"/"
predpath = ""+Y+"/"
percent = float(args.percents)
########

#### Parameters to define ####
#gtpath = path to search for ground truth
#predpath = path to search for segmentation for evaluation
#percent = percent surface hausdorff distance
#### Parameters to define ####

def file_name(file_dir):
   L=[]
   path_list = os.listdir(file_dir)
   path_list.sort() # sort the read path
   good_extensions = [ '.nii.gz', '.mha', '.nii']
   for filename in path_list:
       if any(x in filename for x in good_extensions): #requires this to be changed if not
nifty images
            L.append(os.path.join(filename))
   return L

def computeQualityMeasures(pred,gt):
    quality=dict()
    labelPred=pred
    labelTrue=gt

    #compute hausdorff distance max and average
    hausdorffcomputer=sitk.HausdorffDistanceImageFilter()
    hausdorffcomputer.Execute(labelTrue>0.5,labelPred>0.5)
    quality["Avg HD"]=hausdorffcomputer.GetAverageHausdorffDistance()
    quality["Max HD"]=hausdorffcomputer.GetHausdorffDistance()

    #compute dice values
    dicecomputer=sitk.LabelOverlapMeasuresImageFilter()
    dicecomputer.Execute(labelTrue>0.5,labelPred>0.5)
    quality["DSC"]=dicecomputer.GetDiceCoefficient()

    #compute jaccard metrics
    jaccomputer=sitk.LabelOverlapMeasuresImageFilter()
    jaccomputer.Execute(labelTrue>0.5,labelPred>0.5)
    quality["Jaccard"]=jaccomputer.GetJaccardCoefficient()

    #compute volume similarity
    vol_sim=sitk.LabelOverlapMeasuresImageFilter()
    vol_sim.Execute(labelTrue>0.5,labelPred>0.5)
    quality["Vol Sim"]=vol_sim.GetVolumeSimilarity()

    #compute xor-errors
    statsCmpt = sitk.StatisticsImageFilter()
    xor_img = labelTrue ^ labelPred
    statsCmpt.Execute(xor_img)
    xor_total = statsCmpt.GetSum()
    statsCmpt.Execute(labelTrue)
    mask_norm = statsCmpt.GetSum()
    xor_on_region = xor_total / mask_norm
    quality["xor on region"] = xor_on_region

    bnd_norm = 0
    for slc in range(labelTrue.GetDepth()):
        ctr = sitk.BinaryContour(labelTrue[:,:,slc])
        statsCmpt.Execute( ctr)
        bnd_norm += statsCmpt.GetSum()
    xor_on_bnd = xor_total / bnd_norm
    quality["xor on boundary"] = xor_on_bnd

    #compute surface hausdorff distances
    hd95computer=sitk.StatisticsImageFilter()
    hd95computer.Execute(labelTrue>0.5)
    hd95computer.Execute(labelPred>0.5)
    eTestImage   = sitk.BinaryErode(labelTrue, (1,1,0))
    eResultImage = sitk.BinaryErode(labelPred, (1,1,0))
    hTestImage   = sitk.Subtract(labelTrue, eTestImage)
    hResultImage = sitk.Subtract(labelPred, eResultImage)
    hTestArray   = sitk.GetArrayFromImage(hTestImage)
    hResultArray = sitk.GetArrayFromImage(hResultImage)
    testCoordinates = [gt.TransformIndexToPhysicalPoint(x.tolist()) for x in
np.transpose( np.flipud( np.nonzero(hTestArray) ))]
    resultCoordinates = [gt.TransformIndexToPhysicalPoint(x.tolist()) for x in
np.transpose( np.flipud( np.nonzero(hResultArray) ))]

    def getDistancesFromAtoB(a, b):
        kdTree = scipy.spatial.KDTree(a, leafsize=100)
        return kdTree.query(b, k=1, eps=0, p=2)[0]

    dTestToResult = getDistancesFromAtoB(testCoordinates, resultCoordinates)
    dResultToTest = getDistancesFromAtoB(resultCoordinates, testCoordinates)
    quality["HD95"] = max(np.percentile(dTestToResult, percent),
np.percentile(dResultToTest, percent))
    quality["surface"] = max(np.mean(dTestToResult), np.mean(dResultToTest))

    return quality

gtnames = file_name(gtpath)
prednames = file_name(predpath)

#creates output csv with headers
with open("output.csv", "a", newline="") as outcsv:
    writer = csv.DictWriter(outcsv, fieldnames = ["Average HD (cm)", "Max HD (mm)", "DSC",
"Jaccard", "Volume Similarity", "xor on region", "xor on boundary", "Boundary HD95",
"Boundary Avg HD"])
    writer.writeheader()

#writes image list in order to new csv
rows = zip(prednames)
with open("image_list.csv", "a", newline="") as outcsv1:
    writer = csv.writer(outcsv1)
    for row in rows:
        writer.writerow(row)
```

```
#generate results for images in directory
for i in range(len(gtnames)):
    gt = sitk.ReadImage(gtpath + gtnames[i], sitk.sitkUInt64)
    pred = sitk.ReadImage(predpath + prednames[i], sitk.sitkUInt64)
    result = computeQualityMeasures(pred,gt)
    print(result, file=open("output.csv", "a"))

#remove all text for calculations
text = open("output.csv", "r")
text = ''.join([i for i in text]).replace("{'Avg HD': ", "")
x = open("output.csv","w")
x.writelines(text)
x.close()

text2 = open("output.csv", "r")
text2 = ''.join([i for i in text2]).replace("'Max HD': ", "")
x = open("output.csv","w")
x.writelines(text2)
x.close()

text3 = open("output.csv", "r")
text3 = ''.join([i for i in text3]).replace("'DSC': ", "")
x = open("output.csv","w")
x.writelines(text3)
x.close()

text5 = open("output.csv", "r")
text5 = ''.join([i for i in text5]).replace("'Jaccard': ", "")
x = open("output.csv","w")
x.writelines(text5)
x.close()

text9 = open("output.csv", "r")
text9 = ''.join([i for i in text9]).replace("'Vol Sim': ", "")
x = open("output.csv","w")
x.writelines(text9)
x.close()

text10 = open("output.csv", "r")
text10 = ''.join([i for i in text10]).replace("'xor on region': ", "")
x = open("output.csv","w")
x.writelines(text10)
x.close()

text11 = open("output.csv", "r")
text11 = ''.join([i for i in text11]).replace("'xor on boundary': ", "")
x = open("output.csv","w")
x.writelines(text11)
x.close()

text6 = open("output.csv", "r")
text6 = ''.join([i for i in text6]).replace("'HD95': ", "")
x = open("output.csv","w")
x.writelines(text6)
x.close()

text7 = open("output.csv", "r")
text7 = ''.join([i for i in text7]).replace("'surface': ", "")
x = open("output.csv","w")
x.writelines(text7)
x.close()

text4 = open("output.csv", "r")
text4 = ''.join([i for i in text4]).replace("}", "")
x = open("output.csv","w")
x.writelines(text4)
x.close()

#create summary results csv
with open("results_summary.csv", "a", newline="") as rescsv:
    writers = csv.DictWriter(rescsv, fieldnames = ["stats", "Average HD (mm)", "Max HD
(mm)", "DSC", "Jaccard", "Volume Similarity", "xor on region", "xor on boundary", "Boundary
HD95 (mm)", "Boundary Avg HD (mm)"])
    writers.writeheader()

#summary results calculation
des = pd.read_csv('output.csv')
stats = pd.DataFrame(des)
describe = stats.describe().to_csv("results_summary.csv", mode="a", header=False, sep=",")
```

A.2 Synthesis evaluation metrics

The script provided below is used to calculate the following synthesis evaluation metrics:

- Structural similarity index measure (SSIM)
- Absolute peak signal-to-noise ratio (PSNR)
- Pearson’s Correlation
- Spearman’s Correlation
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)

The script can be run in bash using the arguments `-gt` and `-p` representing the ground truth path and output path, respectively. The python script outputs three files, namely, a list of images (`image_list_reg.csv`), a table of values in the same order as `image_list_reg.csv` (`reg_output.csv`) and a summary of all evaluation metrics (`results_summary_reg.csv`).

A.2.1 Code

```
import numpy as np
import os
import SimpleITK as sitk
import csv
import pandas as pd
import scipy.spatial
import skimage
import argparse
import pathlib

from pathlib import Path
from scipy import stats
from scipy.stats import spearmanr
from scipy.stats import pearsonr
from skimage.metrics import structural_similarity as ssim
from skimage.metrics import peak_signal_noise_ratio as psnr

"""
This script generates 7 evaluation metrics for synthesis for any images in the directories
provided. The directories should contain the corresponding images to compare with the same
origin and voxel spacing. Metrics are calculated over the masked region of the ground truth
in both images (only the areas of interest). Images should be .nii.gz or .mha

    - SSIM
    - Absolute PSNR
    - Pearson Correlation
    - Spearman Correlation
    - Mean Squared Error (MSE)
    - Root Mean Squared Error (RMSE)
    - Mean Absolute Error (MAE)

"""

######
parser = argparse.ArgumentParser()
parser.add_argument('-gt','--gt_root', help="ground truth root", type=Path, required=True)
parser.add_argument('-p','--pred_root', help="predicted root", type=Path, required=True)
args = parser.parse_args()
```

```
X = str(args.gt_root)
Y = str(args.pred_root)
gtpath = ""+X+"/"
predpath = ""+Y+"/"
######

#### Parameters to define ####
#gtpath = path to search for ground truth
#predpath = path to search for synthesised imgage for evaluation
#### Parameters to define ####

def file_name(file_dir):
   L=[]
   path_list = os.listdir(file_dir)
   path_list.sort() # sort the read path
   for filename in path_list:
       if 'nii.gz' or '.mha' in filename: #requires this to be changed if not nifty images
            L.append(os.path.join(filename))
   return L

def computeQualityMeasures(gt,pred):
    quality=dict()
    ground = ground_truth
    predic = prediction

    #read in ground truth image
    gt_arrays = sitk.GetArrayFromImage(ground)
    gt_array = gt_arrays.astype('float')

    #recover mask from ground truth image
    mask = gt_arrays.astype('float')
    mask[mask > 0] = 1

    #read in predicted image
    pred_array = sitk.GetArrayFromImage(predic)
    pred_array = pred_array.astype('float')

    #mask predicted image
    pred_array_masked = np.multiply(mask, pred_array)

    #remove zeroes only for SSIM calculation - calculated over data range max and min of predicted image
    #if the predicted image gives a value of 0 in the region inside the mask this will not be included in the average SSIM
    no_zero_gt = gt_array[gt_array != 0]
    no_zero_pred = pred_array_masked[pred_array_masked != 0]
    struct = ssim(no_zero_gt, no_zero_pred, data_range=no_zero_pred.max() - no_zero_pred.min())
    quality["ssim"] = struct

    #calculate PSNR
    peak = psnr(no_zero_gt, no_zero_pred, data_range=no_zero_pred.max() - no_zero_pred.min())
    quality["psnr"] = np.abs(peak)

    #calculate pearson correlation
    cor, p_val = pearsonr(no_zero_gt, no_zero_pred)
    quality["cor"] = cor
    #quality["p"] = p_val

    #pred_array_masked[pred_array_masked == 0] = 'nan'
    gt_array[gt_array == 0] = 'nan'

    #images converted to 1D
    pred_flatten = pred_array_masked.flatten()
    gt_flatten = gt_array.flatten()

    #calculate spearman correlation
    corr, p_value = spearmanr(gt_flatten, pred_flatten, nan_policy='omit')
    quality["corr"] = corr
    #quality["p_value"] = p_value

    #calculate mean squared error
    square = np.square(gt_flatten - pred_flatten)
    mse = np.nanmean(square)
    quality["mse"] = mse

    #calculate root mean squared error
    rmse = np.sqrt(mse)
    quality["rmse"] = rmse

    #calculate mean absolute error
    mae = np.nanmean(np.abs(gt_flatten - pred_flatten))
    #mae = np.nanmean(abs_error)
    quality["mae"] = mae

    return quality

gtnames = file_name(gtpath)
prednames = file_name(predpath)

#creates output csv with headers
with open("reg_output.csv", "a", newline="") as outcsv:
    writer = csv.DictWriter(outcsv, fieldnames = ["SSIM", "PSNR", "Pearson corr", "Spearman corr", "MSE", "RMSE", "MAE"])
    writer.writeheader()
```

```
#writes image list in order to new csv
rows = zip(prednames)
with open("image_list_reg.csv", "a", newline="") as outcsv1:
    writer = csv.writer(outcsv1)
    for row in rows:
        writer.writerow(row)

#generate results for images in directory
for i in range(len(gtnames)):
    ground_truth = sitk.ReadImage(gtpath + gtnames[i])
    prediction = sitk.ReadImage(predpath + prednames[i])
    result = computeQualityMeasures(ground_truth,prediction)
    print(result, file=open("reg_output.csv", "a"))

#remove all text for summary calculations
text = open("reg_output.csv", "r")
text = ''.join([i for i in text]).replace("{'ssim': ", "")
x = open("reg_output.csv","w")
x.writelines(text)
x.close()

text7 = open("reg_output.csv", "r")
text7 = ''.join([i for i in text7]).replace(" 'psnr': ", "")
x = open("reg_output.csv","w")
x.writelines(text7)
x.close()

text9 = open("reg_output.csv", "r")
text9 = ''.join([i for i in text9]).replace(" 'cor': ", "")
x = open("reg_output.csv","w")
x.writelines(text9)
x.close()

text6 = open("reg_output.csv", "r")
text6 = ''.join([i for i in text6]).replace(" 'corr': ", "")
x = open("reg_output.csv","w")
x.writelines(text6)
x.close()

text2 = open("reg_output.csv", "r")
text2 = ''.join([i for i in text2]).replace(" 'mse': ", "")
x = open("reg_output.csv","w")
x.writelines(text2)
x.close()

text3 = open("reg_output.csv", "r")
text3 = ''.join([i for i in text3]).replace(" 'rmse': ", "")
x = open("reg_output.csv","w")
x.writelines(text3)
x.close()

text5 = open("reg_output.csv", "r")
text5 = ''.join([i for i in text5]).replace(" 'mae': ", "")
x = open("reg_output.csv","w")
x.writelines(text5)
x.close()

text4 = open("reg_output.csv", "r")
text4 = ''.join([i for i in text4]).replace("}", "")
x = open("reg_output.csv","w")
x.writelines(text4)
x.close()

#create summary results csv
with open("results_summary_reg.csv", "a", newline="") as rescsv:
    writers = csv.DictWriter(rescsv, fieldnames = ["stats", "SSIM", "PSNR", "Pearson corr",
"Spearman corr", "MSE", "RMSE", "MAE"])
    writers.writeheader()

#summary results calculation
des = pd.read_csv('reg_output.csv')
stats = pd.DataFrame(des)
describe = stats.describe().to_csv("results_summary_reg.csv", mode="a", header=False,
sep=",")
```

Appendix B
Publications resulting from this thesis

This section details the journal articles, invited review articles, invited talks and conference proceedings resulting from the work produced in this thesis.

B.1 Accepted Journal articles

Astley J.R., Biancardi A.M., Hughes P.J.C., Marshall H., Smith L.J., Collier G.J., Eaden J.A., Weatherley N.D., Hatton M.Q., Wild J.M. and Tahir B.A. (2022). Large-scale investigation of deep learning approaches for ventilated lung segmentation using multi-nuclear hyperpolarized gas MRI. *Scientific Reports* 12, 10566. https://doi.org/10.1038/s41598-022-14672-2.

Astley J.R., Biancardi A.M., Hughes P.J.C., Smith L.J., Marshall H., Eaden J.A., Weatherley N.D., Bray J., Collier G.J., Wild J.M. and Tahir B.A. (2020). 3D Deep Convolutional Neural Network-Based Ventilated Lung Segmentation Using Multi-nuclear Hyperpolarized Gas MRI. MICCAI: In: Petersen J. et al. (eds) Thoracic Image Analysis. TIA 2020. *Lecture Notes in Computer Science*, vol 12502. Springer, Cham. https://doi.org/10.1007/978-3-030-62469-9_3

Astley, J.R., Biancardi, A.M., Marshall, H., Hughes, P.J.C., Collier, G.J., Smith, L.J., Eaden, J.A., Hughes, R., Wild, J.M. and Tahir, B.A. (2023), A Dual-Channel Deep Learning Approach for Lung Cavity Estimation From Hyperpolarized Gas and Proton MRI. *Journal of Magnetic Resonance Imaging*. https://doi.org/10.1002/jmri.28519

Astley J.R., Biancardi A.M., Hughes P.J.C, Marshall H., Collier G.J., Chan H.F., Saunders L.C., Smith L.J., Brook M.L., Thompson R., Rowland-Jones S., Skeoch S., Bianchi S.M., Hatton M.Q., Rahman N.M., Ho L.P., Brightling C.E., Wain L.V., Singapuri A., Evans R.A., Moss A.J., McCann G.P., Neubauer S., Raman B.; C-MORE/PHOSP-COVID Collaborative Group; Wild J.M., Tahir B.A. Implementable Deep Learning for Multi-sequence Proton MRI Lung Segmentation: A Multi-center, Multi-vendor, and Multi-disease Study. *Journal of Magnetic Resonance Imaging*. 2023 Feb 17. doi: 10.1002/jmri.28643. Epub ahead of print. PMID: 36799341.

Astley J.R., Biancardi A.M., Marshall H., Hughes P.J.C., Collier G.J., Hatton M.Q., Wild J.M. and Tahir B.A. (2022). A hybrid model- and deep learning-based framework for functional lung image synthesis from multi-inflation CT and hyperpolarized gas MRI. *Medical Physics*. 2023 Mar 17. doi: 10.1002/mp.16369. Epub ahead of print. PMID: 36932692.

Astley J.R., Biancardi A.M., Marshall H., Smith L.J., Hughes P.J.C., Collier G.J., Saunders L.C., Tofan M., Hatton M.Q., Hughes R., Wild J.M. and Tahir B.A. PhysVENeT: A physics-informed deep learning-based framework for the synthesis of 3D hyperpolarized gas MRI ventilation. *Scientific Reports* [accepted pending revisions].

B.2 Invited review articles

Astley J.R, Wild J.M. and Tahir B.A. (2020). Deep learning in structural and functional lung image analysis. *The British Journal of Radiology*. 20201107. 10.1259/bjr.20201107.

B.3 Invited talks

Deep learning for real world medical imaging problems: novel lung applications. Research IT forum, The University of Sheffield. 16 December 2020. [Online].

The role of deep learning in structural and functional lung image analysis. Departmental research in progress, The University of Sheffield. 5 March 2021. [Online].

A hybrid model- and deep learning-based framework for functional lung image synthesis from non-contrast multi-inflation CT. American association of physicists in medicine (AAPM) CT ventilation working group. 18 August 2021. [Online].

B.4 Conference proceedings

Astley J.R., Biancardi A.M., Marshall H., Smith L.J., Hughes P.J.C., Collier G.J., Hatton M.Q., Wild J.M. and Tahir B.A. (2022). Deep learning–based synthesis of hyperpolarized gas MRI ventilation from 3D multi-inflation proton MRI. Medical imaging and deep learning (MIDL) 2022. *Zurich, Switzerland.*

Astley J.R., Biancardi A.M., Marshall H., Hughes P.J.C., Collier G.J., Smith L.J., Eaden J.A., Wild J.M. and Tahir B.A. (2022). A multi-channel deep learning approach for lung cavity estimation using hyperpolarized gas and proton MRI. Medical imaging and deep learning (MIDL) 2022. *Zurich, Switzerland.*

Smith L.J., Marshall H., Biancardi A., Collier G.J., Chan H.F., Hughes P.J.C., Brook M.L., **Astley J.R.,** Munro R., Rajaram S., Swift A.J., Capener D., Bray J., Ball J., Rodgers O., Jakymelen D., Smith I., Tahir B.A., Rao M., Norquay G., Weatherley N.D., Armstrong L., Hardaker L., Fihn-Wikander T., Hughes R. and Wild J.M. (2022). Physiological Assessment of Patients With Airways Dysanapsis Using ^{129}Xe MRI. European Respiratory Society (ERS) 2022. *Barcelona, Spain.*

Smith L.J., Marshall H., Jakymelen D., Biancardi A., Collier G.J., Chan H.F., Hughes P.J.C., Brook M.L., **Astley J.R.,** Munro R., Rajaram S., Swift A.J., Capener D., Bray J., Ball J., Rodgers O., Smith I., Tahir B.A., Rao M., Norquay G., Weatherley N.D., Armstrong L., Hardaker L., Fihn-Wikander T., Hughes R. and Wild J.M. (2022). ^{129}Xe MRI and lung function to phenotype ventilation heterogeneity in asthma and/or COPD. European Respiratory Society (ERS) 2022. *Barcelona, Spain.*

Marshall H., Smith L.J., Biancardi A., Collier G.J., Chan H.F., Hughes P.J.C., Brook M.L., **Astley J.R.,** Munro R., Rajaram S., Swift A.J., Capener D., Bray J., Ball J., Rodgers O., Jakymelen D., Smith I., Tahir B.A., Rao M., Norquay G., Weatherley N.D., Armstrong L., Hardaker L., Fihn-Wikander T., Hughes R. and Wild J.M. (2022). Physiological Phenotypes of Patients with Asthma and/or COPD Using ^{129}Xe MRI. American Thoracic Society (ATS) 2022. *San Francisco, CA, USA.*

Astley J.R., Biancardi A.M., Marshall H., Tofan M.M, Smith L.J., Hughes P.J.C., Collier G.J., Hatton M.Q., Blè F.X, Hughes R., Wild J.M. and Tahir B.A. (2022). Deep learning-based synthesis of hyperpolarized gas MRI ventilation from 3D multi-inflation proton MRI. The international society for magnetic resonance in medicine (ISMRM) 2022. *London, UK.*

Astley J.R., Biancardi A.M., Marshall H., Hughes P.J.C., Collier G.J., Smith L.J., Eaden J.A., Blè F.X, Hughes R., Wild J.M. and Tahir B.A. (2022). A multi-channel deep learning approach for lung cavity estimation from hyperpolarized gas and proton MRI. The international society for magnetic resonance in medicine (ISMRM) 2022. *London, UK.*

Tahir B.A., Hughes P.J.C., Atkinson J., Jadav I., **Astley J.R.,** Robinson S.D., Biancardi A., Marshall H., Hart K.A., Swinscoe J.A., Ireland R.H., Hatton M.Q. and Wild J.M. (2022). Multi-modal assessment of dose-related changes in regional lung function in non-small cell lung cancer patients receiving radiotherapy. The international society for magnetic resonance in medicine (ISMRM) 2022. *London, UK.*

Marshall H., Smith L.J, Biancardi A., Collier G.J., Chan H.F., Hughes P.J.C., Brook M.L., **Astley J.R**., Munro R., Rajaram S., Swift A.J., Capener D., Bray J., Hussain A.K., Ball J., Rodgers O., Jakymelen D., Smith I., Tahir B.A., Rao M., Norquay G., Weatherley N.D., Armstrong L., Hardaker L., Fihn-Wikander T., Blè F.X, Hughes R. and Wild J.M. (2022). ^{129}Xe MRI patterns of lung function in patients with asthma and/or COPD in the NOVELTY study. The international society for magnetic resonance in medicine (ISMRM) 2022. *London, UK.*

Chan H.., Baldwin T., Barker H., Stewart N., Eaden J., Hughes P.J.., Weatherley N., **Astley J.R.,** Tahir B.A., Johnson K.M., Karwoski R.A., Bartholmai B.J., Tibiletti M., Leonard C.T., Skeoch S., Chaudhuri N., Bruce I.N., Parker G.J.M., Bianchi S.M., and Wild J.M. (2022). Longitudinal comparison of quantitative UTE lung MRI and CT biomarkers in interstitial lung disease. The international society for magnetic resonance in medicine (ISMRM) 2022. *London, UK.*

Astley J.R., Biancardi A.M., Marshall H., Collier G.J., Hughes P.J.C., Wild J.M. and Tahir B.A. (2021). A hybrid model- and deep learning-based framework for functional lung image synthesis from non-contrast multi-inflation CT. Medical imaging and deep learning (MIDL) 2021. *Online.*

Astley J.R., Biancardi A.M., Hughes P.J.C., Marshall H., Eaden J.A., Weatherley N., Collier G.J., Wild J.M. and Tahir B.A. (2021). Comparison of 3D deep convolutional neural networks and training strategies for ventilated lung segmentation using multi-nuclear hyperpolarized gas MRI. American association of physicists in medicine (AAPM) 2021. *Online.*

Astley J.R., Biancardi A.M., Walker M., Hughes P.J.C., Marshall H., Collier G.J., Hatton M.Q., Wild J.M. and Tahir B.A. (2021). Hyperpolarized gas MRI ventilation synthesis from CT: comparison of conventional and deep learning methods. American association of physicists in medicine (AAPM) 2021. *Online.*

Astley J.R., Biancardi A.M., Hughes P.J.C., Smith L.J., Marshall H., Eaden J.A., Weatherley N.D., Collier G.J., Wild J.M. and Tahir B.A. (2021). Automatic Segmentation of Hyperpolarised Gas MRI via Deep Learning. The University of Sheffield Medical School Research Day (MSR) 2021. *Sheffield, UK.*

Astley J.R., Biancardi A.M., Hughes P.J.C., Smith L.J., Marshall H., Collier G.J., Eaden J.A., Weatherley N.D., Wild J.M. and Tahir B.A. (2021). Comparison of 3D deep convolutional neural networks and loss functions for ventilated lung segmentation using multi-nuclear hyperpolarized gas MRI. The international society for magnetic resonance in medicine (ISMRM) 2021. *Online.*

Astley J.R., Biancardi A.M., Marshall H., Smith L.J., Collier G.J., Hughes P.J.C., Walker M., Hatton M.Q., Wild J.M. and Tahir B.A. (2021). Generalizable deep learning for multi-resolution proton MRI lung segmentation in multiple diseases. The international society for magnetic resonance in medicine (ISMRM) 2021. *Online.*

Tahir B.A., Smith L.J., **Astley J.R.,** Walker M., Biancardi A.M., Collier G.J., Hughes P.J.C. and Wild J.M. (2021). Proton lung ventilation MRI in cystic fibrosis: comparison with hyperpolarized gas MRI, pulmonary function tests and multiple-breath washout. The international society for magnetic resonance in medicine (ISMRM) 2021. *Online.*

Astley J.R., Robinson S., Hatton M.Q., Wild J.M. and Tahir B.A. (2021). Deep learning-based survival prediction for non-small cell lung cancer patients undergoing radical radiotherapy. World lung cancer congress (WLCC) 2021*. Online.*

Mitchell T.A, **Astley J.R.,** Robinson S., Bryant H., Danson S., Tahir B.A. and Hatton M.Q. (2021). Artificial neural network-based tumor recurrence prediction in non-small cell lung cancer patients following radical radiotherapy. World lung cancer congress (WLCC) 2021*. Online.*

Astley J.R., Biancardi A.M., Hughes P.J.C., Smith L.J., Marshall H., Mussell G.T., Eaden J.A., Weatherley N., Collier G.J., Wild J.M. and Tahir B.A. (2020). Automatic Segmentation of Hyperpolarized Gas MRI via Deep Learning. Institute of pure and applied mathematics: deep learning and medical applications (IPAM-DLM) 2020. *Los Angeles, CA, USA.*

Astley J.R., Biancardi A.M., Hughes P.J.C., Smith L.J., Marshall H., Mussell G.T., Eaden J.A., Weatherley N., Collier G.J., Wild J.M. and Tahir B.A. (2020). Automatic Segmentation of Hyperpolarized Gas MRI via Deep Learning. The international society for magnetic resonance in medicine (ISMRM) 2020. *Online.*

Bibliography

Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado G S, Davis A, Dean J and Devin M 2016 Tensorflow: Large-scale machine learning on heterogeneous distributed systems *arXiv preprint arXiv:1603.04467*

Adamson A S and Smith A 2018 Machine Learning and Health Care Disparities in Dermatology *JAMA Dermatology* **154** 1247-8

Adnan M, Kalra S, Cresswell J C, Taylor G W and Tizhoosh H R 2022 Federated learning and differential privacy for medical image analysis *Scientific Reports* **12** 1953

Agarap A F 2018 Deep Learning using Rectified Linear Units (ReLU) *arXiv preprint arXiv:1803.08375*.

Aggarwal R, Sounderajah V, Martin G, Ting D S W, Karthikesalingam A, King D, Ashrafian H and Darzi A 2021 Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis *npj Digital Medicine* **4** 65

Ajraoui S, Lee K J, Deppe M H, Parnell S R, Parra-Robles J and Wild J M 2010 Compressed sensing in hyperpolarized 3He lung MRI *Magn Reson Med* **63** 1059-69

Akila Agnes S, Anitha J and Dinesh Peter J 2018 Automatic lung segmentation in low-dose chest CT scans using convolutional deep and wide network (CDWN) *Neural Computing and Applications* **32**, 15845–15855

Alebiosu D O and Muhammad F P Medical Image Classification: A Comparison of Deep Pre-trained Neural Networks *IEEE Student Conference on Research and Development (SCOReD),15-17 Oct. 2019),* vol. Series*)* 306-10

Ali S and Rittscher J Conv2Warp: An Unsupervised Deformable Image Registration with Continuous Convolution and Warping. *Machine Learning in Medical Imaging, (Cham, 2019),* vol. Series*)* ed H-I Suk*, et al.*: Springer International Publishing) 489-97

Anthimopoulos M, Christodoulidis S, Ebner L, Geiser T, Christe A and Mougiakakou S 2019 Semantic Segmentation of Pathological Lung Tissue With Dilated Fully Convolutional Networks *IEEE J Biomed Health Inform* **23** 714-22

Antony J, Carlson D J, Keall P J and Xing L 2007 Clinical Impact of 4D-CT Imaging on Lung Cancer Radiotherapy Treatment Planning and Biological Response *International Journal of Radiation Oncology, Biology, Physics* **69** S526

Astley J R, Biancardi A, Marshall H, Smith L J, Collier G J, Hughes P J C, Walker M, Hatton M Q, Wild J M and Tahir B A Generalizable deep learning for multi-resolution proton MRI lung segmentation in multiple diseases. *Proceedings of the 29th Annual Meeting of ISMRM, (Online, 2021),* vol. Series (abstract 3224)*)*

Astley J R, Biancardi A M, Hughes P J C, Marshall H, Smith L J, Collier G J, Eaden J A, Weatherley N D, Hatton M Q, Wild J M and Tahir B A 2022 Large-scale investigation of deep learning approaches for ventilated lung segmentation using multi-nuclear hyperpolarized gas MRI *Scientific Reports* **12** 10566

Astley J R, Biancardi A M, Hughes P J C, Smith L J, Marshall H, Eaden J, Bray J, Weatherley N D, Collier G J, Wild J M and Tahir B A 3D Deep Convolutional Neural Network-Based Ventilated Lung Segmentation Using Multi-nuclear Hyperpolarized Gas MRI *Thoracic Image Analysis, (Cham, 2020// 2020a),* vol. Series*)* ed J Petersen*, et al.*: Springer International Publishing) 24-35

Astley J R, Wild J M and Tahir B A 2020b Deep learning in structural and functional lung image analysis *The British Journal of Radiology* **0** 20201107

Aurora P, Gustafsson P, Bush A, Lindblad A, Oliver C, Wallis C and Stocks J 2004 Multiple breath inert gas washout as a measure of ventilation distribution in children with cystic fibrosis *Thorax* **59** 1068-73

Avants B, Tustison N and Song G 2008 Advanced normalization tools (ANTS) *Insight J* **1–35**

Avants B B, Tustison N J, Stauffer M, Song G, Wu B and Gee J C 2014 The Insight ToolKit image registration framework *Front Neuroinform* **8** 44-4

Bae K, Jeon K N, Hwang M J, Lee J S, Ha J Y, Ryu K H and Kim H C 2019 Comparison of lung imaging using three-dimensional ultrashort echo time and zero echo time sequences: preliminary study *Eur Radiol* **29** 2253-62

Bakator M and Radosav D 2018 Deep Learning and Medical Diagnosis: A Review of Literature *Multimodal Technologies and Interaction* **2** 47

Balachandar N, Chang K, Kalpathy-Cramer J and Rubin D L 2020 Accounting for data variability in multi-institutional distributed deep learning for medical imaging *Journal of the American Medical Informatics Association* **27** 700-8

Bauman G, Puderbach M, Deimling M, Jellus V, Chefd'hotel C, Dinkel J, Hintze C, Kauczor H U and Schad L R 2009 Non-contrast-enhanced perfusion and ventilation assessment of the human lung by means of fourier decomposition in proton MRI *Magn Reson Med* **62** 656-64

Beauchemin M, Thomson K P B and Edwards G 1998 On the Hausdorff Distance Used for the Evaluation of Segmentation Results *Canadian Journal of Remote Sensing* **24** 3-8

Beaudry J, Esquinas P and Shieh C-C Learning from our neighbours: a novel approach on sinogram completion using bin-sharing and deep learning to reconstruct high quality 4DCBCT *Proc. SPIE Medical Imaging 2019, (San Diego, California, United States, March 1, 2019),* vol. Series*)*: SPIE- Intl Soc Optical Eng) 153-3

Berger L, Eoin H, Cardoso M J and Ourselin S *Annual Conference on Medical Image Understanding and Analysis, 2018),* vol. Series*)*: Springer) 277-86

Berners-Lee C M 1968 Cybernetics and Forecasting *Nature* **219** 202-3

Bezdek J C, Ehrlich R and Full W 1984 FCM: The fuzzy c-means clustering algorithm *Computers & Geosciences* **10** 191-203

Bi L, Kim J, Kumar A, Feng D and Fulham M Synthesis of positron emission tomography (PET) images via multi-channel generative adversarial networks (GANs) *RAMBO 2017 Lecture Notes in Computer Science, (Québec City, QC, Canada, Sept 14, 2017),* vol. Series 10555 LNCS*)*: Springer Cham) 43-51

Biancardi A, Acunzo L, Marshall H, Tahir B A, Hughes P J C, Smith L J, Weatherley N D, Collier G J and Wild J M A paired approach to the segmentation of proton and hyperpolarized gas MR images of the lungs *Proceedings of the 26th Annual Meeting of ISMRM, (Paris, 2018),* vol. Series (abstract 2442)*)*

Biancardi A M and Wild J M 2017 New Disagreement Metrics Incorporating Spatial Detail – Applications to Lung Imaging *Medical Image Understanding and Analysis, (Cham, 2017),* vol. Series*)* ed M Valdés Hernández and V González-Castro: Springer International Publishing) 804-14

Blendowski M and Heinrich M P 2019 Combining MRF-based deformable registration and deep binary 3D-CNN descriptors for large lung motion estimation in COPD patients *Int J Comput Assist Radiol Surg* **14** 43-52

Bluemke D A, Moy L, Bredella M A, Ertl-Wagner B B, Fowler K J, Goh V J, Halpern E F, Hess C P, Schiebler M L and Weiss C R 2019 Assessing Radiology Research on Artificial Intelligence: A Brief Guide for Authors, Reviewers, and Readers—From the Radiology Editorial Board *Radiology* **294** 487-9

Brennan D, Schubert L, Diot Q, Castillo R, Castillo E, Guerrero T, Martel M K, Linderman D, Gaspar L E, Miften M, Kavanagh B D and Vinogradskiy Y 2015 Clinical Validation of 4-Dimensional Computed Tomography Ventilation With Pulmonary Function Test Data *International Journal of Radiation Oncology, Biology, Physics* **92** 423-9

Capaldi D P I, Eddy R L, Svenningsen S, Guo F, Baxter J S H, McLeod A J, Nair P, McCormack D G and Parraga G 2018 Free-breathing Pulmonary MR Imaging to Quantify Regional Ventilation *Radiology* **287** 693-704

Capaldi D P I, Guo F, Xing L and Parraga G 2020 Pulmonary Ventilation Maps Generated with Free-breathing Proton MRI and a Deep Convolutional Neural Network *Radiology* **298** 427-38

Castillo E, Castillo R, Vinogradskiy Y, Dougherty M, Solis D, Myziuk N, Thompson A, Guerra R, Nair G and Guerrero T 2019 Robust CT ventilation from the integral formulation of the Jacobian *Med Phys* **46** 2115-25

Castillo R, Castillo E, Martinez J and Guerrero T 2010a Ventilation from four-dimensional computed tomography: density versus Jacobian methods *Phys Med Biol* **55** 4661-85

Castillo R, Castillo E, McCurdy M, Gomez D R, Block A M, Bergsma D, Joy S and Guerrero T 2012 Spatial correspondence of 4D CT ventilation and SPECT pulmonary perfusion defects in patients with malignant airway stenosis *Physics in Medicine and Biology* **57** 1855-71

Chassagnon G, Vakalopoulou M, Paragios N and Revel M P 2020 Artificial intelligence applications for thoracic imaging *Eur J Radiol* **123** 108774

Chen W, Wei H F, Peng S T, Sun J W, Qiao X and Liu B Q 2019 HSN: Hybrid Segmentation Network for Small Cell Lung Cancer Segmentation *Ieee Access* **7** 75591-603

Chlebus G, Meine H, Thoduka S, Abolmaali N, van Ginneken B, Hahn H K and Schenk A 2019 Reducing inter-observer variability and interaction time of MR liver volumetry by combining automatic CNN-based liver segmentation and manual corrections *PLoS One* **14** e0217228

Chuang K-S, Tzeng H-L, Chen S, Wu J and Chen T-J 2006 Fuzzy c-means clustering with spatial information for image segmentation *Computerized Medical Imaging and Graphics* **30** 9-15

Collier G J, Acunzo L, Smith L J, Hughes P J C, Norquay G, Chan H F, Biancardi A, Marshall H and Wild J M Linear binning maps for image analysis of pulmonary ventilation with hyperpolarized gas MRI: transferability and clinical applications *Proceedings of the 26th Annual Meeting of ISMRM, (Paris, 2018),* vol. Series (abstract 4482)*)*

Collier G J, Hughes P J C, Horn F C, Chan H-F, Tahir B, Norquay G, Stewart N J and Wild J M 2019 Single breath-held acquisition of coregistered 3D 129Xe lung ventilation and anatomical proton images of the human lung with compressed sensing *Magnetic Resonance in Medicine* **82** 342-7

Congleton J and Muers M 1995 The incidence of airflow obstruction in bronchial carcinoma, its relation to breathlessness, and response to bronchodilator therapy *Respiratory medicine* **89** 291-6

Corcoran H L, Renner W R and Milstein M J 1992 Review of high-resolution CT of the lung *Radiographics* **12** 917-39

Crockett C B, Samson P, Chuter R, Dubec M, Faivre-Finn C, Green O L, Hackett S L, McDonald F, Robinson C, Shiarli A-M, Straza M W, Verhoeff J J C, Werner-Wasik M, Vlacich G and Cobben D 2021 Initial Clinical Experience of MR-Guided Radiotherapy for Non-Small Cell Lung Cancer *Frontiers in Oncology* **11**

Dai W, Dong N, Wang Z, Liang X, Zhang H and Xing E P Scan: Structure correcting adversarial network for organ segmentation in chest x-rays *DLMIA 2018 Lecture Notes in Computer Science, (Granada, Spain, Sept 20, 2018),* vol. Series 11045 LNCS*)*: Springer Cham) 263-73

Dauphin Y, Vries H, Chung J and Bengio Y 2015 RMSProp and equilibrated adaptive learning rates for non-convex optimization *arXiv* **35** de Lange E E, Altes T A, Patrie J T, Gaare J D, Knake J J, Mugler III J P and Platts-Mills T A 2006 Evaluation of asthma with hyperpolarized helium-3 MRI: correlation with clinical severity and spirometry *Chest* **130** 1055-62

de Vos B D, Berendsen F F, Viergever M A, Sokooti H, Staring M and Isgum I 2019 A deep learning framework for unsupervised affine and deformable image registration *Med Image Anal* **52** 128-43

de Vos B D, Wolterink J M, de Jong P A, Leiner T, Viergever M A and Isgum I 2017 ConvNet-Based Localization of Anatomical Structures in 3-D Medical Images *IEEE Trans Med Imaging* **36** 1470-81

Dice L R 1945 Measures of the Amount of Ecologic Association Between Species *Ecology* **26** 297-302

Dietze M M A, Branderhorst W, Kunnen B, Viergever M A and de Jong H 2019 Accelerated SPECT image reconstruction with FBP and an image enhancement convolutional neural network *EJNMMI Phys* **6** 14

Ding K, Cao K, Fuld M, Du K, Christensen G, Hoffman E and Reinhardt J 2012 Comparison of image registration based measures of regional lung ventilation from dynamic spiral CT with Xe-CT *Medical physics* **39** **8** 5084-98

Dong X, Lei Y, Wang T, Thomas M, Tang L, Curran W J, Liu T and Yang X 2019 Automatic multiorgan segmentation in thorax CT images using U-net-GAN *Med Phys* **46** 2157-68

Donovan G M and Kritter T 2015 Spatial pattern formation in the lung *J Math Biol* **70** 1119-49

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G and Gelly S 2020 An image is worth 16x16 words: Transformers for image recognition at scale *arXiv preprint arXiv:2010.11929*

Duan C, Deng H, Xiao S, Xie J, Li H, Sun X, Ma L, Lou X, Ye C and Zhou X 2019 Fast and accurate reconstruction of human lung gas MRI with deep learning *Magn Reson Med* **82** 2273-85

Duchi J, Hazan E and Singer Y 2011 Adaptive Subgradient Methods for Online Learning and Stochastic Optimization *Journal of Machine Learning Research* **12** 2121-59

Ebner L, Kammerman J, Driehuys B, Schiebler M L, Cadman R V and Fain S B 2017 The role of hyperpolarized 129xenon in MR imaging of pulmonary function *European journal of radiology* **86** 343-52

Edmunds D, Sharp G and Winey B 2019 Automatic diaphragm segmentation for real-time lung tumor tracking on cone-beam CT projections: a convolutional neural network approach *Biomedical Physics & Engineering Express* **5**

Eichinger M, Heussel C-P, Kauczor H-U, Tiddens H and Puderbach M 2010 Computed tomography and magnetic resonance imaging in cystic fibrosis lung disease *Journal of Magnetic Resonance Imaging* **32** 1370-8

El Bitar Z, Lazaro D, Coello C, Breton V, Hill D and Buvat I 2006 Fully 3D Monte Carlo image reconstruction in SPECT using functional regions *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **569** 399-403

Eppenhof K A J and Pluim J P W 2019 Pulmonary CT Registration Through Supervised Learning With Convolutional Neural Networks *IEEE Trans Med Imaging* **38** 1097-105

Fain S B, Korosec F R, Holmes J H, O'Halloran R, Sorkness R L and Grist T M 2007 Functional lung imaging using hyperpolarized gas MRI. *J Magn Reson Imaging* **25** 910-23

Fechter T and Baltas D 2020 One-Shot Learning for Deformable Medical Image Registration and Periodic Motion Tracking *IEEE Trans Med Imaging* **39** 2506-17

Ferrante E, Oktay O, Glocker B and Milone D H On the adaptability of unsupervised CNN-based deformable image registration to unseen image domains *MLMI 2018. Lecture Notes in Computer Science, (Granada, Spain, Sept 16, 2018),* vol. Series 11046 LNCS*)*: Springer Cham) 294-302

Feyzmahdavian H, Aytekin A and Johansson M 2015 An Asynchronous Mini-Batch Algorithm for Regularized Stochastic Optimization *IEEE Transactions on Automatic Control* **61**

Fields B K K, Demirjian N L, Dadgar H and Gholamrezanezhad A 2021 Imaging of COVID-19: CT, MRI, and PET *Semin Nucl Med* **51** 312-20

Fu Y, Lei Y, Wang T, Higgins K, Bradley J D, Curran W J, Liu T and Yang X 2020 LungRegNet: An unsupervised deformable image registration method for 4D-CT lung *Med Phys* **47** 1763-74

Gaál G, Maga B and Lukács A 2020 Attention U-Net Based Adversarial Architectures for Chest X-ray Lung Segmentation *arXiv preprint arXiv:2003.10304*

Galib S M, Lee H K, Guy C L, Riblett M J and Hugo G D 2020 A fast and scalable method for quality assurance of deformable image registration on lung CT scans using convolutional neural networks *Med Phys* **47** 99-109

Gamaleldin O, Bahgat A Y, Anwar O, Seif-Elnasr M, Eissa L A, Razek A A, Shehata G M and Khalifa M H 2020 Role of dynamic sleep MRI in obstructive sleep apnea syndrome *Oral Radiology* 1-9

Gao M, Xu Z, Lu L, Harrison A P, Summers R M and Mollura D J Multi-label deep regression and unordered pooling for holistic interstitial lung disease pattern detection *MLMI 2016. Lecture Notes in Computer Science, (Athens, Greece, Oct 17, 2016),* vol. Series 10019 LNCS*)*: Springer Cham) 147-55

Garcia-Uceda Juarez A, Selvan R, Saghir Z and Bruijne M D A joint 3D UNet-graph neural network-based method for airway segmentation from chest CTs *International workshop on machine learning in medical imaging, 2019),* vol. Series*)*: Springer) 583-91

GBD15, Wang H D, Naghavi M and Allen C 2016 Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980-2015: a systematic analysis for the Global Burden of Disease Study 2015 *Lancet* **388** 1459-544

Ge Y, Su T, Zhu J, Deng X, Zhang Q, Chen J, Hu Z, Zheng H and Liang D 2020 ADAPTIVE-NET: deep computed tomography reconstruction network with analytical domain transformation knowledge *Quant Imaging Med Surg* **10** 415-27

Gerard S E, Herrmann J, Kaczka D W, Musch G, Fernandez-Bustamante A and Reinhardt J M 2020 Multi-resolution convolutional neural networks for fully automated segmentation of acutely injured lungs in multiple species *Med Image Anal* **60** 101592

Gerard S E, Herrmann J, Kaczka D W and Reinhardt J M Transfer learning for segmentation of injured lungs using coarse-to-fine convolutional neural networks *RAMBO 2018, BIA 2018, TIA 2018. Lecture Notes in Computer Science, (Granada, Spain, Sept 16-20, 2018),* vol. Series 11040 LNCS*)*: Springer Cham) 191-201

Gerard S E, Herrmann J, Xin Y, Martin K T, Rezoagli E, Ippolito D, Bellani G, Cereda M, Guo J, Hoffman E A, Kaczka D W and Reinhardt J M 2021 CT image segmentation for inflamed and fibrotic lungs using a multi-resolution convolutional neural network *Scientific Reports* **11** 1455

Gerard S E, Patton T J, Christensen G E, Bayouth J E and Reinhardt J M 2019 FissureNet: A Deep Learning Approach For Pulmonary Fissure Detection in CT Images *IEEE Trans Med Imaging* **38** 156-66

Gibson E, Giganti F, Hu Y, Bonmati E, Bandula S, Gurusamy K, Davidson B, Pereira S P, Clarkson M J and Barratt D C 2018a Automatic Multi-Organ Segmentation on Abdominal CT With Dense V-Networks *IEEE Transactions on Medical Imaging* **37** 1822-34

Gibson E, Li W, Sudre C, Fidon L, Shakir D I, Wang G, Eaton-Rosen Z, Gray R, Doel T, Hu Y, Whyntie T, Nachev P, Modat M, Barratt D C, Ourselin S, Cardoso M J and Vercauteren T 2018b NiftyNet: a deep-learning platform for medical imaging *Comput Methods Programs Biomed* **158** 113-22

Goldstein E B, Coco G, Murray A B and Green M O 2014 Data-driven components in a model of inner-shelf sorted bedforms: a new hybrid model *Earth Surf. Dynam.* **2** 67-82

Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y Generative adversarial nets *Advances in neural information processing systems, (Montréal CANADA Dec 8-13, 2014),* vol. Series 3*)*: Neural information processing systems foundation) 2672-80

Gou S, Liu W, Jiao C, Liu H, Gu Y, Zhang X, Lee J and Jiao L 2019 Gradient regularized convolutional neural networks for low-dose CT image enhancement *Phys Med Biol* **64** 165017

Grover A, Kapoor A and Horvitz E 2015 A Deep Hybrid Model for Weather Forecasting. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* (Sydney, NSW, Australia: Association for Computing Machinery) 379–86

Grover J, Byrne H L, Sun Y, Kipritidis J and Keall P 2022 Investigating the use of machine learning to generate ventilation images from CT scans *Medical Physics* **49** 5258-67

Guerrero T, Sanders K, Noyola-Martinez J, Castillo E, Zhang Y, Tapia R, Guerra R, Borghero Y and Komaki R 2005 Quantification of regional ventilation from treatment planning CT *Int J Radiat Oncol Biol Phys* **62** 630-4

Guo Z, Li X, Huang H, Guo N and Li Q 2019 Deep learning-based image segmentation on multimodal medical imaging *IEEE Transactions on Radiation and Plasma Medical Sciences* **3** 162-9

Hamilton F, Lloyd A L and Flores K B 2017 Hybrid modeling and prediction of dynamical systems *PLOS Computational Biology* **13** e1005655

Hammernik K, Klatzer T, Kobler E, Recht M P, Sodickson D K, Pock T and Knoll F 2018 Learning a variational network for reconstruction of accelerated MRI data *Magn Reson Med* **79** 3055-71

Hanania A N, Mainwaring W, Ghebre Y T, Hanania N A and Ludwig M 2019 Radiation-Induced Lung Injury: Assessment and Management *Chest* **156** 150-62

Hatamizadeh A, Hoogi A, Sengupta D, Lu W, Wilcox B, Rubin D and Terzopoulos D 2019 Deep Active Lesion Segmentation *Machine Learning in Medical Imaging* 98-105

Hatamizadeh A, Tang Y, Nath V, Yang D, Myronenko A, Landman B, Roth H R and Xu D Unetr: Transformers for 3d medical image segmentation *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision,2022),* vol. Series*)* 574-84

He K, Zhang X, Ren S and Sun J 2015 Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification *IEEE International Conference on Computer Vision (ICCV 2015)* **1502**

He M, Driehuys B, Que L G and Huang Y T 2016 Using Hyperpolarized (129)Xe MRI to Quantify the Pulmonary Ventilation Distribution *Acad Radiol* **23** 1521-31

He M, Kaushik S S, Robertson S H, Freeman M S, Virgincar R S, McAdams H P and Driehuys B 2014 Extending semiautomatic ventilation defect analysis for hyperpolarized (129)Xe ventilation MRI *Acad Radiol* **21** 1530-41

Hesamian M H, Jia W, He X and Kennedy P 2019 Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges *J Digit Imaging* **32** 582-96

Hofmanninger J, Prayer F, Pan J, Rohrich S, Prosch H and Langs G 2020 Automatic lung segmentation in routine imaging is a data diversity problem, not a methodology problem *ArXiv preprint* abs/2001.11767

Hollings N and Shaw P 2002 Diagnostic imaging of lung cancer *European Respiratory Journal* **19** 722-42

Hooda R, Mittal A and Sofat S 2018 An Efficient Variant of Fully-Convolutional Network for Segmenting Lung Fields from Chest Radiographs *Wireless Personal Communications* **101** 1559-79

Horn F C, Marshall H, Collier G J, Kay R, Siddiqui S, Brightling C E, Parra-Robles J and Wild J M 2017 Regional Ventilation Changes in the Lung: Treatment Response Mapping by Using Hyperpolarized Gas MR Imaging as a Quantitative Biomarker *Radiology* **284** 854-61

Horn F C, Tahir B A, Stewart N J, Collier G J, Norquay G, Leung G, Ireland R H, Parra-Robles J, Marshall H and Wild J M 2014 Lung ventilation volumetry with same-breath acquisition of hyperpolarized gas and proton MRI *NMR in Biomedicine* **27** 1461-7

Hu Q, de F S L F, Holanda G B, Alves S S A, Dos S S F H, Han T and Reboucas Filho P P 2020 An effective approach for CT lung segmentation using mask region-based convolutional neural networks *Artif Intell Med* **103** 101792

Huang L, Yang D, Lang B and Deng J *Huang, L., Yang, D., Lang, B., & Deng, J. (2018). Decorrelated batch normalization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (Salt Lake City, UT, June 18-23, 2018),* vol. Series*)*: IEEE) 791-800

Hughes P J C, Horn F C, Collier G J, Biancardi A, Marshall H and Wild J M 2018 Spatial fuzzy c-means thresholding for semiautomated calculation of percentage lung ventilated volume from hyperpolarized gas and (1) H MRI *J Magn Reson Imaging* **47** 640-6

Hughes P J C, Smith L, Chan H-F, Tahir B A, Norquay G, Collier G J, Biancardi A, Marshall H and Wild J M 2019 Assessment of the influence of lung inflation state on the quantitative parameters derived from hyperpolarized gas lung ventilation MRI in healthy volunteers *Journal of applied physiology (Bethesda, Md. : 1985)* **126** 183-92

Hwang S and Park S 2017 Accurate Lung Segmentation via Network-Wise Training of Convolutional Networks *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* **10553** 92-9

Imran A-A-Z, Hatamizadeh A, Ananth S P, Ding X, Tajbakhsh N and Terzopoulos D 2020 Fast and automatic segmentation of pulmonary lobes from chest CT using a progressive dense V-network *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* **8** 509-18

Ioffe S and Szegedy C Batch normalization: Accelerating deep network training by reducing internal covariate shift *Proceedings of the 32nd International Conference on International Conference on Machine Learning, (Lille, France, July 2015),* vol. Series 37*)*: International Machine Learning Society (IMLS)) 448-56

Ireland R H, Tahir B A, Wild J M, Lee C E and Hatton M Q 2016 Functional Image-guided Radiotherapy Planning for Normal Lung Avoidance *Clin Oncol (R Coll Radiol)* **28** 695-707

Isensee F, Kickingereder P, Wick W, Bendszus M and Maier-Hein K H *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, (Cham, 2019),* vol. Series*)* ed A Crimi*, et al.*: Springer International Publishing) 234-44

Isensee F, Petersen J, Klein A, Zimmerer D, Jaeger P F, Kohl S, Wasserthal J, Koehler G, Norajitra T and Wirkert S 2018 nnu-net: Self-adapting framework for u-net-based medical image segmentation *arXiv preprint arXiv:1809.10486*

Ivanovska T, Hegenscheid K, Laqua R, Gläser S, Ewert R and Völzke H *Visualization in Medicine and Life Sciences III, (Cham, 2016// 2016),* vol. Series*)* ed L Linsen*, et al.*: Springer International Publishing) 3-24

Jaeger S, Karargyris A, Candemir S, Folio L, Siegelman J, Callaghan F, Zhiyun X, Palaniappan K, Singh R K, Antani S, Thoma G, Yi-Xiang W, Pu-Xuan L and McDonald C J 2014 Automatic tuberculosis screening using chest radiographs *IEEE Trans Med Imaging* **33** 233-45

Jahangir R, Kamali-Asl A and Arabi H 2022 Deep Learning-Based Attenuation and Scatter Correction of Brain 18F-FDG PET Images in the Image Domain*. arXiv preprint arXiv:2206.14673.*

Jang B S, Chang J H, Park A J and Wu H G 2019 Generation of virtual lung single-photon emission computed tomography/CT fusion images for functional avoidance radiotherapy planning using machine learning algorithms *J Med Imaging Radiat Oncol* **63** 229-35

Javaid U, Dasnoy D and Lee J A Multi-organ Segmentation of Chest CT Images in Radiation Oncology: Comparison of Standard and Dilated UNet *ACIVS Lecture notes on computer science, (Granada, Spain, Sept 16-20, 2018),* vol. Series 11182 LNCS*)*: Springer Cham) 188-99

Jiang J, Hu Y C, Tyagi N, Zhang P, Rimner A, Deasy J O and Veeraraghavan H 2019 Cross-modality (CT-MRI) prior augmented deep learning for robust lung tumor segmentation from small MR datasets *Med Phys* **46** 4392-404

Jiang J, Hu Y C, Tyagi N, Zhang P, Rimner A, Mageras G S, Deasy J O and Veeraraghavan H *MICCAI 2018 Lecture notes on computer science, (Granada, Spain, Sept 16-20, 2018),* vol. Series 11071 LNCS*)*: Springer Cham) 777-85

Jiang Z, Yin F F, Ge Y and Ren L 2020 A multi-scale framework with unsupervised joint training of convolutional neural networks for pulmonary deformable image registration *Phys Med Biol* **65** 015011

Jin W, Li X, Fatehi M and Hamarneh G 2022 Guidelines and evaluation for clinical explainable AI on medical image analysis *arXiv preprint arXiv:2202.10553*

Jögi J, Jonson B, Ekberg M and Bajc M 2010 Ventilation-perfusion SPECT with 99mTc-DTPA versus Technegas: a head-to-head study in obstructive and nonobstructive disease *J Nucl Med* **51** 735-41

Juarez A G U, Tiddens H A W M and de Bruijne M Automatic airway segmentation in chest CT using convolutional neural networks *RAMBO 2018 Lecture notes on computer science, (Granada, Spain, Sept 16-20, 2018),* vol. Series 11040 LNCS*)*: Springer Cham) 238-50

Kaireit T F, Sorrentino S A, Renne J, Schoenfeld C, Voskrebenzev A, Gutberlet M, Schulz A, Jakob P M, Hansen G, Wacker F, Welte T, Tümmler B and Vogel-Claussen J 2017 Functional lung MRI for regional monitoring of patients with cystic fibrosis *PLOS ONE* **12** e0187483

Kalinovsky A, Liauchuk V and Tarasau A 2017 Lesion Detection in Ct Images Using Deep Learning Semantic Segmentation Technique *International Workshop Photogrammetric and Computer Vision Techniques for Video Surveillance, Biometrics and Biomedicine* **42-2** 13-7

Kanungo T, Mount D M, Netanyahu N S, Piatko C D, Silverman R and Wu A Y 2002 An efficient k-means clustering algorithm: Analysis and implementation *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **24** 881-92

Kazeminia S, Baur C, Kuijper A, van Ginneken B, Navab N, Albarqouni S and Mukhopadhyay A 2018 GANs for Medical Image Analysis *arXiv preprint arXiv:1809.06222*

Kida S, Bal M, Kabus S, Negahdar M, Shan X, Loo B W, Keall P J and Yamamoto T 2016 CT ventilation functional image-based IMRT treatment plans are comparable to SPECT ventilation functional image-based plans *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology* **118** 521-7

Kingma D and Ba J *3rd International Conference for Learning Representations, (San Diego, May 7-15 2015),* vol. Series abs/1412.6980*)*

Kipritidis J, Siva S, Hofman M, Callahan J, Hicks R and Keall P 2014 Validating and improving CT ventilation imaging by correlating with ventilation 4D-PET/CT using Ga-68-labeled nanoparticles *Medical physics* **41** 011910

Kipritidis J, Tahir B A, Cazoulat G, Hofman M S, Siva S, Callahan J, Hardcastle N, Yamamoto T, Christensen G E, Reinhardt J M, Kadoya N, Patton T J, Gerard S E, Duarte I, Archibald-Heeren B, Byrne M, Sims R, Ramsay S, Booth J T, Eslick E, Hegi-Johnson F, Woodruff H C, Ireland R H, Wild J M, Cai J, Bayouth J E, Brock K and Keall P J 2019 The VAMPIRE challenge: A multi-institutional validation study of CT ventilation imaging *Medical physics* **46** 1198-217

Kirby M, Heydarian M, Svenningsen S, Wheatley A, McCormack D G, Etemad-Rezai R and Parraga G 2012a Hyperpolarized 3He Magnetic Resonance Functional Imaging Semiautomated Segmentation *Academic Radiology* **19** 141-52

Kirby M, Mathew L, Heydarian M, Etemad-Rezai R, McCormack D G and Parraga G 2011 Chronic obstructive pulmonary disease: quantification of bronchodilator effects by using hyperpolarized He MR imaging *Radiology* **261** 283-92

Kirby M, Mathew L, Wheatley A, Santyr G E, McCormack D G and Parraga G 2010 Chronic obstructive pulmonary disease: longitudinal hyperpolarized 3He MR imaging *Radiology* **256** 280-9

Kirby M, Svenningsen S, Kanhere N, Owrangi A, Wheatley A, Coxson H O, Santyr G E, Paterson N A, McCormack D G and Parraga G 2013 Pulmonary ventilation visualized using hyperpolarized helium-3 and xenon-129 magnetic resonance imaging: differences in COPD and relationship to emphysema *Journal of applied physiology* **114** 707-15

Kirby M, Svenningsen S, Owrangi A, Wheatley A, Farag A, Ouriadov A, Santyr G E, Etemad-Rezai R, Coxson H O, McCormack D G and Parraga G 2012b Hyperpolarized 3He and 129Xe MR Imaging in Healthy Volunteers and Patients with Chronic Obstructive Pulmonary Disease *Radiology* **265** 600-10

Kjorstad A, Regier M, Fiehler J and Sedlacik J 2017 A decade of lung expansion: A review of ventilation-weighted (1)H lung MRI *Z Med Phys* **27** 172-9

Kläser K, Borges P, Shaw R, Ranzini M, Modat M, Atkinson D, Thielemans K, Hutton B, Goh V, Cook G, Cardoso M J and Ourselin S 2021 A multi-channel uncertainty-aware multi-resolution network for MR to CT synthesis *Appl Sci (Basel)* **11** 1667

Krizhevsky A, Sutskever I and Hinton G E ImageNet classification with deep convolutional neural networks 26th Annual Conference on Neural Information Processing Systems, (Lake Tahoe, NV United States, Dec 3-6, 2012), vol. Series 2) 1097-105

Kulasekara M, Dinh V Q, Fernandez-del-Valle M and Klingensmith J D 2022 Comparison of two-dimensional and three-dimensional U-Net architectures for segmentation of adipose tissue in cardiac magnetic resonance images *Medical & Biological Engineering & Computing* **60** 2291-306

Lafarge M W, Moeskops P, Veta M, Pluim J P W and Eppenhof K A J *SPIE Medical Imaging, 2018, (Houston, Texas, United States, Feb 10-15, 2018),* vol. Series*)*: SPIE-Intl Soc Optical Eng) 27-7

Le Roux P Y, Hicks R J, Siva S and Hofman M S 2019 PET/CT Lung Ventilation and Perfusion Scanning using Galligas and Gallium-68-MAA *Semin Nucl Med* **49** 71-81

Lee H, Grosse R, Ranganath R and Ng A Y 2011 Unsupervised Learning of Hierarchical Representations with Convolutional Deep Belief Networks *Communications of the Acm* **54** 95-103

Lee S M, Lee J G, Lee G, Choe J, Do K H, Kim N and Seo J B 2019 CT Image Conversion among Different Reconstruction Kernels without a Sinogram by Using a Convolutional Neural Network *Korean J Radiol* **20** 295-303

Levin D L, Schiebler M L and Hopkins S R 2017 Physiology for the pulmonary functional imager *Eur J Radiol* **86** 308-12

Li B N, Chui C K, Chang S and Ong S H 2011 Integrating spatial fuzzy clustering with level set methods for automated medical image segmentation *Comput. Biol. Med.* **41** 1-10

Li W, Wang G, Fidon L, Ourselin S, Cardoso M J and Vercauteren T On the compactness, efficiency, and representation of 3D convolutional networks: brain parcellation as a pretext task *International conference on information processing in medical imaging, 2017),* vol. Series*)*: Springer) 348-60

Liu H, Zheng L, Shi G, Xu Q, Wang Q, Zhu H, Feng H, Wang L, Zhang N, Xue M and Dai Y 2021a Pulmonary Functional Imaging for Lung Adenocarcinoma: Combined MRI Assessment Based on IVIM-DWI and OE-UTE-MRI *Frontiers in Oncology* **11**

Liu X, Xie L, Wang Y, Zou J, Xiong J, Ying Z and Vasilakos A V 2021b Privacy and Security Issues in Deep Learning: A Survey *IEEE Access* **9** 4566-93

Liu Y, Fu W, Selvakumaran V, Phelan M, Segars W P, Samei E, Mazurowski M, Lo J Y-C, Rubin G and Henao R Deep learning of 3D CT images for organ segmentation using 2D multi-channel SegNet model *In Medical Imaging 2019: Imaging Informatics for Healthcare, Research, and Applications, (San Diego, California, United States, Feb 16-21, 2019),* vol. Series Vol. 10954, p. 109541D*)*: SPIE-Intl Soc Optical Eng) 49-9

Liu Z, Miao J, Huang P, Wang W, Wang X, Zhai Y, Wang J, Zhou Z, Bi N, Tian Y and Dai J 2020 A deep learning method for producing ventilation images from 4DCT: First comparison with technegas SPECT ventilation *Medical Physics* **47** 1249-57

Lobo P and Guruprasad S Classification and Segmentation Techniques for Detection of Lung Cancer from CT Images *International Conference on Inventive Research in Computing Applications (ICIRCA), (Coimbatore, 2018),* vol. Series*)*: Institute of Electrical and Electronics Engineers Inc.) 1014-9

Loeve M, Lequin M H, de Bruijne M, Hartmann I J C, Gerbrands K, van Straten M, Hop W C J and Tiddens H A W M 2009 Cystic Fibrosis: Are Volumetric Ultra-Low-Dose Expiratory CT Scans Sufficient for Monitoring Related Lung Disease? *Radiology* **253** 223-9

Long Y, She X and Mukhopadhyay S HybridNet: integrating model-based and data-driven learning to predict evolution of dynamical systems *Conference on Robot Learning,2018),* vol. Series*)*: PMLR) 551-60

Lundervold A S and Lundervold A 2019 An overview of deep learning in medical imaging focusing on MRI *Z Med Phys* **29** 102-27

Lustberg T, van Soest J, Gooding M, Peressutti D, Aljabar P, van der Stoep J, van Elmpt W and Dekker A 2018 Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer *Radiother Oncol* **126** 312-7

Maas A L, *2013* Rectifier Nonlinearities Improve Neural Network Acoustic Models. ai.standford.edu

Mackay J and Crofton J 1996 Tobacco and the developing world *British Medical Bulletin* **52** 206-21

Magnant J, Vecellio L, de Monte M, Grimbert D, Valat C, Boissinot E, Guilloteau D, Lemarié E and Diot P 2006 Comparative Analysis of Different Scintigraphic Approaches to Assess Pulmonary Ventilation *Journal of Aerosol Medicine* **19** 148-59

Mahapatra D, Ge Z, Sedai S and Chakravorty R Joint registration and segmentation of xray images using generative adversarial networks *MLMI 2018 Lecture notes on computer science, (Granada, Spain, Sept 16-20, 2018),* vol. Series 11046 LNCS*)*: Springer Cham) 73-80

Manjon J V, Coupe P, Marti-Bonmati L, Collins D L and Robles M 2010 Adaptive non-local means denoising of MR images with spatially varying noise levels *J Magn Reson Imaging* **31** 192-203

Mansoor A, Bagci U, Foster B, Xu Z, Papadakis G Z, Folio L R, Udupa J K and Mollura D J 2015 Segmentation and Image Analysis of Abnormal Lungs at CT: Current Approaches, Challenges, and Future Trends *Radiographics* **35** 1056-76

Marshall H, Deppe M H, Parra-Robles J, Hillis S, Billings C G, Rajaram S, Swift A, Miller S R, Watson J H, Wolber J, Lipson D A, Lawson R and Wild J M 2012 Direct visualisation of collateral ventilation in COPD with hyperpolarised gas MRI *Thorax* **67** 613

Marshall H, Horsley A, Taylor C J, Smith L, Hughes D, Horn F C, Swift A J, Parra-Robles J, Hughes P J, Norquay G, Stewart N J, Collier G J, Teare D, Cunningham S, Aldag I and Wild J M 2017 Detection of early subclinical lung disease in children with cystic fibrosis by lung ventilation imaging with hyperpolarised gas MRI *Thorax* **72** 760

Martini K and Frauenfelder T 2018 Emphysema and lung volume reduction: the role of radiology *J Thorac Dis* **10** S2719-s31

Milletari F, Navab N and Ahmadi S A V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation *Proceedings of 2016 Fourth International Conference on 3d Vision (3dv)* 565-71

Mirza M and Osindero S 2014 Conditional Generative Adversarial Nets. *arXiv preprint arXiv:1411.1784*.

Mittal A, Hooda R and Sofat S 2018 LF-SegNet: A Fully Convolutional Encoder-Decoder Network for Segmenting Lung Fields from Chest Radiographs *Wireless Personal Communications* **101** 511-29

Moher D, Liberati A, Tetzlaff J, Altman D G, Altman D, Antes G, Atkins D, Barbour V, Barrowman N, Berlin J A, Clark J, Clarke M, Cook D, D'Amico R, Deeks J J, Devereaux P J, Dickersin K, Egger M, Ernst E, Gøtzsche P C, Grimshaw J, Guyatt G, Higgins J, Ioannidis J P A, Kleijnen J, Lang T, Magrini N, McNamee D, Moja L, Mulrow C, Napoli M, Oxman A, Pham B, Rennie D, Sampson M, Schulz K F, Shekelle P G, Tovey D and Tugwell P 2009 Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. American College of Physicians) 264-9

Morid M A, Borjali A and Del Fiol G 2021 A scoping review of transfer learning research on medical image analysis using ImageNet *Comput Biol Med* **128** 104115

Moriya T, Roth H R, Nakamura S, Oda H, Nagara K, Oda M and Mori K Unsupervised segmentation of 3D medical images based on clustering and deep representation learning *Proc. SPIE 10578, Medical Imaging 2018: Biomedical Applications in Molecular, Structural, and Functional Imaging, (Houston, Texas, USA, Feb 11-13, 2018),* vol. Series Volume 10578*)*: SPIE-Intl Soc Optical Eng) 71-1

Moro J, Ribes S, Caselles O and Parent L 2013 Evaluation of two registration techniques applied to lung adaptive radiotherapy *Physica Medica* **29**

Mukesh M, Benson R, Jena R, Hoole A, Roques T, Scrase C, Martin C, Whitfield G A, Gemmill J and Jefferies S 2012 Interobserver variation in clinical target volume and organs at risk segmentation in post-parotidectomy radiotherapy: can segmentation protocols help? *Br J Radiol* **85** e530-6

Negahdar M, Beymer D and Syeda-Mahmood T F Automated volumetric lung segmentation of thoracic CT images using fully convolutional neural network *SPIE Medical Imaging, 2018, (Houston, Texas, United States, Feb 10-15, 2018),* vol. Series*)*: SPIE-Intl Soc Optical Eng) 54-4

Nie D, Cao X, Gao Y, Wang L and Shen D Estimating CT image from MRI data using 3D fully convolutional networks *DLMIA 2016, LABELS 2016. Lecture Notes in Computer Science, (Athens, Greece, Oct 21, 2016),* vol. Series 10008 LNCS*)*: Springer Cham) 170-8

Norberg P, Bake B, Jacobsson L, Carlsson G A and Gustafsson A 2007 Evaluation of reconstruction techniques for lung single photon emission tomography: a Monte Carlo study *Nucl Med Commun* **28** 929-36

Norquay G, Collier G, Rao M, Stewart N and Wild J 2018a Xe 129-Rb spin-exchange optical pumping with high photon efficiency *Physical Review Letters* **121** 153201

Novikov A A, Lenis D, Major D, Hladuvka J, Wimmer M and Buhler K 2018 Fully Convolutional Architectures for Multiclass Segmentation in Chest Radiographs *IEEE Trans Med Imaging* **37** 1865-76

Oktay O, Schlemper J, Folgoc L L, Lee M J, Heinrich M P, Misawa K, Mori K, McDonagh S G, Hammerla N Y, Kainz B, Glocker B and Rueckert D 2018 Attention U-Net: Learning Where to Look for the Pancreas *ArXiv* abs/1804.03999

Olberg S, Zhang H, Green O L, Mazur T R, Yang D, Hugo G D, Bradley J D, Mutic S and Park J 2018 Deep Learning-Based Pseudo CT Reconstruction for MR Only-Guided Radiation Therapy of Lung SBRT *International Journal of Radiation Oncology Biology Physics* **102** E309-E10

Park B, Park H, Lee S M, Seo J B and Kim N 2019 Lung Segmentation on HRCT and Volumetric CT for Diffuse Interstitial Lung Disease Using Deep Convolutional Neural Networks *J Digit Imaging* **32** 1019-26

Park J, Yun J, Kim N, Park B, Cho Y, Park H J, Song M, Lee M and Seo J B 2020 Fully Automated Lung Lobe Segmentation in Volumetric Chest CT with 3D U-Net: Validation with Intra- and Extra-Datasets *J Digit Imaging* **33** 221-30

Pehrson L M, Nielsen M B and Ammitzbøl Lauridsen C 2019 Automatic Pulmonary Nodule Detection Applying Deep Learning or Machine Learning Algorithms to the LIDC-IDRI Database: A Systematic Review *Diagnostics (Basel)* **9**

Pennati F, Borzani I, Moroni L, Russo M C, Faelli N, Aliverti A and Colombo C 2021 Longitudinal Assessment of Patients With Cystic Fibrosis Lung Disease With Multivolume Noncontrast MRI and Spirometry *J Magn Reson Imaging* **53** 1570-80

Pennisi M, Kavasidis I, Spampinato C, Schinina V, Palazzo S, Salanitri F P, Bellitto G, Rundo F, Aldinucci M, Cristofaro M, Campioni P, Pianura E, Di Stefano F, Petrone A, Albarello F, Ippolito G, Cuzzocrea S and Conoci S 2021 An explainable AI system for automated COVID-19 assessment and lesion categorization from CT-scans *Artificial Intelligence in Medicine* **118** 102114

Petersson J, Sánchez-Crespo A, Larsson S A and Mure M 2007 Physiological imaging of the lung: single-photon-emission computed tomography (SPECT) *Journal of Applied Physiology* **102** 468-76

Plaut D C, Nowlan S J, Hinton G E and Nowlan S 1986 Experiments on Learning by Back Propagation. *Harvard University*

Poirot M, Bergmans R, Thomson B, Jolink F, Moum S, Gonzalez R, Lev M, Tan C and Gupta R 2019 Physics-informed Deep Learning for Dual-Energy Computed Tomography Image Processing *Scientific Reports* **9**

Portney P R and Mullahy J 1990 Urban air quality and chronic respiratory disease *Regional Science and Urban Economics* **20** 407-18

Preiswerk F, Cheng C C, Luo J and Madore B Synthesizing dynamic MRI using long-term recurrent convolutional networks *MLMI 2018, LNCS 11046 Lecture notes on computer science, (Granada, Spain, Sept 16-20, 2018),* vol. Series 11046 LNCS*)*: Springer Cham) 89-97

Qin C, Shi B, Liao R, Mansi T, Rueckert D and Kamen A Unsupervised Deformable Registration for Multi-modal Images via Disentangled Representations *Information Processing in Medical Imaging, (Cham, 2019),* vol. Series*)* ed A C S Chung*, et al.*: Springer International Publishing) 249-61

Rajchl M, Lee M C, Oktay O, Kamnitsas K, Passerat-Palmbach J, Bai W, Damodaram M, Rutherford M A, Hajnal J V, Kainz B and Rueckert D 2017 DeepCut: Object Segmentation From Bounding Box Annotations Using Convolutional Neural Networks *IEEE Trans Med Imaging* **36** 674-83

Raschka S 2018 Model evaluation, model selection, and algorithm selection in machine learning *arXiv preprint arXiv:1811.12808*

Reinhardt J, Ding K, Cao K, Christensen G, Hoffman E and Bodas S V 2008 Registration-based estimates of local lung tissue expansion compared to xenon CT measures of specific ventilation *Medical image analysis* **12 6** 752-63

Reinke A, Eisenmann M, Tizabi M D, Sudre C H, Rädsch T, Antonelli M, Arbel T, Bakas S, Cardoso M J and Cheplygina V 2021 Common limitations of image processing metrics: A picture story *arXiv preprint arXiv:2104.05642*

Ren G, Ho W Y, Qin J and Cai J 2019 Deriving Lung Perfusion Directly from CT Image Using Deep Convolutional Neural Network: A Preliminary Study *Artificial Intelligence in Radiation Therapy* 102-9

Ren G, Li B, Lam S-k, Xiao H, Huang Y-H, Cheung A L-y, Lu Y, Mao R, Ge H, Kong F-M, Ho W-y and Cai J 2022 A Transfer Learning Framework for Deep Learning-Based CT-to-Perfusion Mapping on Lung Cancer Patients *Frontiers in Oncology* **12**

Robbins H and Monro S 1951 A Stochastic Approximation Method *The Annals of Mathematical Statistics* **22** 400-7, 8

Ronneberger O, Fischer P and Brox T U-Net: Convolutional Networks for Biomedical Image Segmentation *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, (Cham, 2015),* vol. Series*)* ed N Navab*, et al.*: Springer International Publishing) 234-41

Rosenblatt F 1958 The perceptron: a probabilistic model for information storage and organization in the brain *Psychol Rev* **65** 386-408

Salehi S S M, Erdogmus D and Gholipour A Tversky loss function for image segmentation using 3D fully convolutional deep networks *International workshop on machine learning in medical imaging, 2017),* vol. Series*)*: Springer) 379-87

Salehinejad H, Colak E, Dowdell T, Barfett J and Valaee S 2019 Synthesizing Chest X-Ray Pathology for Training Deep Convolutional Neural Networks *IEEE Trans Med Imaging* **38** 1197-206

San Jose Estepar R, Ross J C, Harmouche R, Onieva J, Diaz A A and Washko G R 2015*.* Chest Imaging Platform: An Open-Source Library and Workstation for Quantitative Chest Imaging *LUNG IMAGING II: NEW PROBES AND EMERGING TECHNOLOGIES*: American Thoracic Society) A4975-A

Sandkühler R, Jud C, Bauman G, Willers C, Pusterla O, Nyilas S, Peters A, Ebner L, Stranziger E, Bieri O, Latzin P and Cattin P C 2019 Weakly Supervised Learning Strategy for Lung Defect Segmentation *Machine Learning in Medical Imaging* 541-8

Sentker T, Madesta F and Werner R GDL-FIRE4D: Deep Learning-Based Fast 4D CT Image Registration *MICCAI 2018 Lecture notes on computer science, (Granada, Spain, Sept 16-20, 2018),* vol. Series*)*: Springer Cham) 765-73

Seyyed-Kalantari L, Zhang H, McDermott M B A, Chen I Y and Ghassemi M 2021 Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations *Nature Medicine* **27** 2176-82

Shamshad F, Khan S, Zamir S W, Khan M H, Hayat M, Khan F S and Fu H 2022 Transformers in medical imaging: A survey *arXiv preprint arXiv:2201.09873*

Shapiro M D and Blaschko M B 2004 On hausdorff distance measures *Computer Vision Laboratory University of Massachusetts Amherst, MA* **1003**

Sheikh K, Coxson H O and Parraga G 2016 This is what COPD looks like *Respirology* **21** 224-36

Shiraishi J, Katsuragawa S, Ikezoe J, Matsumoto T, Kobayashi T, Komatsu K, Matsui M, Fujita H, Kodera Y and Doi K 2000 Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules *AJR Am J Roentgenol* **174** 71-4

Simon B 2004 Non-Invasive Imaging of Regional Lung Function using X-Ray Computed Tomography *Journal of Clinical Monitoring and Computing* **16** 433-42

Simon B A, Kaczka D W, Bankier A A and Parraga G 2012a What can computed tomography and magnetic resonance imaging tell us about ventilation? *J Appl Physiol* **113** 647-57

Simonyan K and Zisserman A 2014 Very deep convolutional networks for large-scale image recognition *arXiv preprint arXiv:1409.1556*

Smith L J, Collier G J, Marshall H, Hughes P J, Biancardi A M, Wildman M, Aldag I, West N, Horsley A and Wild J M 2018 Patterns of regional lung physiology in cystic fibrosis using ventilation magnetic resonance imaging and multiple-breath washout *European Respiratory Journal* **52**

Soans R E and Shackleford J A Organ localization and identification in thoracic CT volumes using 3D CNNs leveraging spatial anatomic relations *SPIE Medical Imaging, 2018, (Houston, Texas, United States, Feb 10-15, 2018),* vol. Series*)*: SPIE-Intl Soc Optical Eng) 68-8

Sokooti H, De Vos B, Berendsen F, Ghafoorian M, Yousefi S, Lelieveldt B, Isgum I and Staring M 2019 3D Convolutional Neural Networks Image Registration Based on Efficient Supervised Learning from Artificial Deformations. *arXiv preprint arXiv:1908.10235.*

Sokooti H, de Vos B, Berendsen F, Lelieveldt B P F, Išgum I and Staring M Nonrigid image registration using multi-scale 3D convolutional neural networks *MICCAI 2017. Lecture Notes in Computer Science, (Quebec City, QC, Canada, Sept 11-13, 2017),* vol. Series 10433 LNCS*)*: Springer Cham) 232- 9

Soliman A, Shaffie A, Ghazal M, Gimel'Farb G, Keynton R and El-Baz A A Novel CNN Segmentation Framework Based on Using New Shape and Appearance Features *25th IEEE International Conference on Image Processing (ICIP), (Athens, Greece, Oct 7-10, 2018),* vol. Series vol 10433*)*: IEEE Computer Society) 3488-92

Sousa P, Galdran A, Costa P and Campilho A Learning to Segment the Lung Volume from CT Scans Based on Semi-Automatic Ground-Truth *IEEE 16th International Symposium on Biomedical Imaging (Venice, Italy, April 8-11, 2019),* vol. Series*)*: Institute of Electrical and Electronics Engineers (IEEE)) 1202-6

Souza J C, Bandeira Diniz J O, Ferreira J L, Franca da Silva G L, Correa Silva A and de Paiva A C 2019 An automatic method for lung segmentation and reconstruction in chest X-ray using deep neural networks *Comput Methods Programs Biomed* **177** 285-96

Springenberg J T, Dosovitskiy A, Brox T and Riedmiller M 2014 Striving for simplicity: The all convolutional net *arXiv preprint arXiv:1412.6806*

Stergios C, Mihir S, Maria V, Guillaume C, Marie-Pierre R, Stavroula M and Nikos P Linear and Deformable Image Registration with 3D Convolutional Neural Networks *Image Analysis for Moving Organ, Breast, and Thoracic Images, (Granada, Spain, Sept 16-20, 2018),* vol. Series*)* ed D Stoyanov*, et al.*: Springer International Publishing) 13-22

Stewart N J, Chan H F, Hughes P J C, Horn F C, Norquay G, Rao M, Yates D P, Ireland R H, Hatton M Q, Tahir B A, Ford P, Swift A J, Lawson R, Marshall H, Collier G J and Wild J M 2018 Comparison of (3) He and (129) Xe MRI for evaluation of lung microstructure and ventilation at 1.5T *J Magn Reson Imaging* **48** 632-42

Stewart N J, Norquay G, Griffiths P D and Wild J M 2015 Feasibility of human lung ventilation imaging using highly polarized naturally abundant xenon and optimized three-dimensional steady-state free precession *Magn Reson Med* **74** 346-52

Stewart N J, Smith L J, Chan H F, Eaden J A, Rajaram S, Swift A J, Weatherley N D, Biancardi A, Collier G J, Hughes D, Klafkowski G, Johns C S, West N, Ugonna K, Bianchi S M, Lawson R, Sabroe I, Marshall H and Wild J M 2021 Lung MRI with hyperpolarised gases: current & future clinical perspectives *Br J Radiol* 20210207

Suzuki Y, Yamagata K, Yanagawa M, Kido S and Tomiyama N Weak supervision in convolutional neural network for semantic segmentation of diffuse lung diseases using partially annotated dataset (((*Medical Imaging 2020: Computer-Aided Diagnosis,2020),* vol. Series 11314*)*: SPIE) 528-33

Taha A A and Hanbury A 2015 Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool *BMC Med Imaging* **15** 29

Tahir B A, Bragg C M, Wild J M, Swinscoe J A, Lawless S E, Hart K A, Hatton M Q and Ireland R H 2017 Impact of field number and beam angle on functional image-guided lung cancer radiotherapy planning *Phys Med Biol* **62** 7114-30

Tahir B A, Hughes P J C, Robinson S D, Marshall H, Stewart N J, Norquay G, Biancardi A, Chan H F, Collier G J, Hart K A, Swinscoe J A, Hatton M Q, Wild J M and Ireland R H 2018 Spatial Comparison of CT-Based Surrogates of Lung Ventilation With Hyperpolarized Helium-3 and Xenon-129 Gas MRI in Patients Undergoing Radiation Therapy *Int J Radiat Oncol Biol Phys* **102** 1276-86

Tahir B A, Marshall H, Hughes P J C, Brightling C E, Collier G, Ireland R H and Wild J M 2019 Comparison of CT ventilation imaging and hyperpolarised gas MRI: effects of breathing manoeuvre *Phys Med Biol* **64** 055013

Tahir B A, Smith L J, Astley J R, Walker M, Biancardi A, Collier G J, Hughes P J C, Marshall H and Wild J M Proton lung ventilation MRI in cystic fibrosis: comparison with hyperpolarized gas MRI, pulmonary function tests and multiple-breath washout *Proceedings of the 29th Annual Meeting of ISMRM, (Online, 2021),* vol. Series (abstract 3220)*)*

Tahir B A, Swift A J, Marshall H, Parra-Robles J, Hatton M Q, Hartley R, Kay R, Brightling C E, Vos W, Wild J M and Ireland R H 2014 A method for quantitative analysis of regional lung ventilation using deformable image registration of CT and hybrid hyperpolarized gas/1H MRI *Phys Med Biol* **59** 7267-77

Tahir B A, Van Holsbeke C, Ireland R H, Swift A J, Horn F C, Marshall H, Kenworthy J C, Parra-Robles J, Hartley R, Kay R, Brightling C E, De Backer J, Vos W and Wild J M 2016 Comparison of CT-based Lobar Ventilation with 3He MR Imaging Ventilation Measurements *Radiology* **278** 585-92

Tahmasebi N, Boulanger P, Noga M and Punithakumar K 2018 A Fully Convolutional Deep Neural Network for Lung Tumor Boundary Tracking in MRI *Conf Proc IEEE Eng Med Biol Soc* **2018** 5906-9

Tajbakhsh N, Shin J Y, Gurudu S R, Hurst R T, Kendall C B, Gotway M B and Jianming L 2016 Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? *IEEE Trans Med Imaging* **35** 1299-312

Tanno R, Saeedi A, Sankaranarayanan S, Alexander D C and Silberman N Learning from noisy labels by regularized estimation of annotator confusion *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Salt lake City, UT, USA, 18-23 June 2019),* vol. Series*)* 11244-53

Thomen R P, Walkup L L, Roach D J, Higano N, Schapiro A, Brody A, Clancy J P, Cleveland Z I and Woods J C 2020 Regional structure-function in cystic fibrosis lung disease using hyperpolarized 129Xe and ultrashort echo magnetic resonance imaging *American journal of respiratory and critical care medicine* **202** 290-2

Togao O, Tsuji R, Ohno Y, Dimitrov I and Takahashi M 2010 Ultrashort echo time (UTE) MRI of the lung: assessment of tissue density in the lung parenchyma *Magn Reson Med* **64** 1491-8

Tomasi C and Manduchi R Bilateral filtering for gray and color images *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271),7-7 Jan. 1998 1998),* vol. Series*)* 839-46

Torre L A, Bray F, Siegel R L, Ferlay J, Lortet-Tieulent J and Jemal A 2015 Global cancer statistics, 2012 *CA Cancer J Clin* **65** 87-108

Torres-Duque C, Maldonado D, Pérez-Padilla R, Ezzati M and Viegi G 2008 Biomass Fuels and Respiratory Diseases *Proceedings of the American Thoracic Society* **5** 577-90

Tustison N J, Avants B B, Cook P A, Zheng Y, Egan A, Yushkevich P A and Gee J C 2010 N4ITK: improved N3 bias correction *IEEE Trans Med Imaging* **29** 1310-20

Tustison N J, Avants B B, Flors L, Altes T A, de Lange E E, Mugler Iii J P and Gee J C 2011 Ventilation-based segmentation of the lungs using hyperpolarized 3He MRI *Journal of Magnetic Resonance Imaging* **34** 831-41

Tustison N J, Avants B B, Lin Z, Feng X, Cullen N, Mata J F, Flors L, Gee J C, Altes T A, Mugler Iii J P and Qing K 2019 Convolutional Neural Networks with Template-Based Data Augmentation for Functional Lung Image Quantification *Acad Radiol* **26** 412-23

Tzeng Y S, Lutchen K and Albert M 2009 The difference in ventilation heterogeneity between asthmatic and healthy subjects quantified using hyperpolarized 3He MRI *J Appl Physiol (1985)* **106** 813-22

Umehara K, Ota J and Ishida T 2018 Application of Super-Resolution Convolutional Neural Network for Enhancing Image Resolution in Chest CT *J Digit Imaging* **31** 441-50

Vakalopoulou M, Chassagnon G, Bus N, Marini R, Zacharaki E I, Revel M P and Paragios N AtlasNet: Multi-atlas Non-linear Deep Networks for Medical Image Segmentation *MICCAI 2018, LNCS 11073 Lecture notes on computer science, (Granada, Spain, Sept 16-20, 2018),* vol. Series 11073 LNCS*)*: Springer Cham) 658-66

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I 2017 Attention is all you need *Advances in neural information processing systems* **30**

Vos T, Abajobir A A, Hassen Abate K, Abbafati C, Abbas K M, Abd-Allah F, Suliankatchi Abdulkader R and Abdulle A M 2017 Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016 *Lancet* **390** 1211-59

Voskrebenzev A, Gutberlet M, Klimes F, Kaireit T, Schönfeld C, Rotärmel A, Wacker F and Vogel-Claussen J 2017 Feasibility of quantitative regional ventilation and perfusion mapping with phase-resolved functional lung (PREFUL) MRI in healthy volunteers and COPD, CTEPH, and CF patients *Magnetic Resonance in Medicine* **79**

Voskrebenzev A and Vogel-Claussen J 2021 Proton MRI of the Lung: How to Tame Scarce Protons and Fast Signal Decay *J Magn Reson Imaging* **53** 1344-57

Wang G, Li W, Zuluaga M A, Pratt R, Patel P A, Aertsen M, Doel T, David A L, Deprest J, Ourselin S and Vercauteren T 2018a Interactive Medical Image Segmentation Using Deep Learning With Image-Specific Fine Tuning *IEEE Trans Med Imaging* **37** 1562-73

Wang C, Hayashi Y, Oda M, Itoh H, Kitasaka T, Frangi A F and Mori K Tubular Structure Segmentation Using Spatial Fully Connected Network with Radial Distance Loss for 3D Medical Images *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019,* 348-56

Wang G, Li W, Zuluaga M A, Pratt R, Patel P A, Aertsen M, Doel T, David A L, Deprest J, Ourselin S and Vercauteren T 2018a Interactive Medical Image Segmentation Using Deep Learning With Image-Specific Fine Tuning *IEEE Trans Med Imaging* **37** 1562-73

Wang X, Teng P, Lo P, Banola A, Kim G, Abtin F, Goldin J and Brown M High throughput lung and lobar segmentation by 2D and 3D CNN on chest CT with diffuse lung disease *RAMBO 2018/BIA Lecture* *notes on computer science, (Granada, Spain, Sept 16-20, 2018 2018b),* vol. Series 11040 LNCS*)*: Springer Cham) 202-14

Wang Z, Bovik A, Sheikh H and Simoncelli E 2004 Image Quality Assessment: From Error Visibility to Structural Similarity *Image Processing, IEEE Transactions on* **13** 600-12

Werbos P and John P 1974 Beyond regression : new tools for prediction and analysis in the behavioral sciences *Harvard University*

Westcott A, Capaldi D P I, McCormack D G, Ward A D, Fenster A and Parraga G 2019 Chronic Obstructive Pulmonary Disease: Thoracic CT Texture Analysis and Machine Learning to Predict Pulmonary Ventilation *Radiology* **293** 676-84

Wetzels M, Broers E, Peters P, Feijs L, Widdershoven J and Habibovic M 2018 Patient Perspectives on Health Data Privacy and Management: "Where Is My Data and Whose Is It?" *Int J Telemed Appl* **2018** 3838747

Whiting P, Singatullina N and Rosser J H 2015 Computed tomography of the chest: I. Basic principles *BJA Education* **15** 299-304

Wild J M, Ajraoui S, Deppe M H, Parnell S R, Marshall H, Parra-Robles J and Ireland R H 2011 Synchronous acquisition of hyperpolarised 3He and 1H MR images of the lungs - maximising mutual anatomical and functional information *NMR Biomed* **24** 130-4

Wild J M, Marshall H, Bock M, Schad L R, Jakob P M, Puderbach M, Molinari F, Van Beek E J R and Biederer J 2012 MRI of the lung (1/3): methods *Insights into Imaging* **3** 345-53

Willard J, Jia X, Xu S, Steinbach M and Kumar V 2020 Integrating physics-based modeling with machine learning: A survey *arXiv preprint arXiv:2003.04919*

Willemink M J and Noel P B 2019 The evolution of image reconstruction for CT-from filtered back projection to artificial intelligence *Eur Radiol* **29** 2185-95

Wilmet V, Verma S, Redl T, Sandaker H and Li Z 2021 A Comparison of Supervised and Unsupervised Deep Learning Methods for Anomaly Detection in Images *arXiv preprint arXiv:2107.09204*

Woodhouse N, Wild J M, Paley M N, Fichele S, Said Z, Swift A J and van Beek E J 2005 Combined helium-3/proton magnetic resonance imaging measurement of ventilated lung volumes in smokers compared to never-smokers *J Magn Reson Imaging* **21** 365-9

Xu J and Liu H 2017 Segmentation of Pulmonary CT Image by Using Convolutional Neural Network Based on Membership Function *IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), (Guangzhou, China, Jul 22-23, 2017),* vol. Series 1*)*: Institute of Electrical and Electronics Engineers Inc.) 198-203

Xu M, Qi S, Yue Y, Teng Y, Xu L, Yao Y and Qian W 2019 Segmentation of lung parenchyma in CT images using CNN trained with the clustering algorithm generated dataset *Biomed Eng Online* **18** 2

Xu Y Pulmonary Vessel Segmentation via Stage-Wise Convolutional Networks With Orientation-Based Region Growing Optimization *International Conference on Health Information Science, 2019),* vol. Series*)*: Springer) 193-200

Yablonskiy D A, Sukstanskii A L and Quirk J D 2017 Diffusion lung imaging with hyperpolarized gas MRI *NMR Biomed* **30**

Yamamoto T, Kabus S, Berg J V, Lorenz C, Goris M, Loo B and Keall P Evaluation of Four-dimensional (4D) Computed Tomography (CT) Pulmonary Ventilation Imaging by Comparison with Single Photon Emission Computed Tomography (SPECT) Scans for a Lung Cancer Patient *3rd International Workshop on Pulmonary Image Analysis 2010 117-128*

Yamamoto T, Kabus S, Lorenz C, Mittra E, Hong J C, Chung M, Eclov N, To J, Diehn M, Loo B W and Keall P J 2014 Pulmonary Ventilation Imaging Based on 4-Dimensional Computed Tomography: Comparison With Pulmonary Function Tests and SPECT Ventilation Images *International Journal of Radiation Oncology*Biology*Physics* **90** 414-22

Yang J, Veeraraghavan H, Armato S G, 3rd, Farahani K, Kirby J S, Kalpathy-Kramer J, van Elmpt W, Dekker A, Han X, Feng X, Aljabar P, Oliveira B, van der Heyden B, Zamdborg L, Lam D, Gooding M and Sharp G C 2018 Autosegmentation for thoracic radiation treatment planning: A grand challenge at AAPM 2017 *Med Phys* **45** 4568-81

Yang X, Tjio G, Yang F, Ding J, Kumar S, Leng S, Zhao X, Tan R-S, Zhong L and Su Y 2019 A Multi-channel Deep Learning Approach for Segmentation of the Left Ventricular Endocardium from Cardiac Images *41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2019),* vol. Series*)*: IEEE) 4016-9

Ye J C 2019 Compressed sensing MRI: a review from signal processing perspective *BMC Biomedical Engineering* **1**

Yuan S T, Frey K A, Gross M D, Hayman J A, Arenberg D, Curtis J L, Cai X-W, Ramnath N, Kalemkerian G P, Ten Haken R K, Eisbruch A and Kong F-M S 2011 Semiquantification and classification of local pulmonary function by V/Q single photon emission computed tomography in patients with non-small cell lung cancer: potential indication for radiotherapy planning *J Thorac Oncol* **6** 71-8

Yun J, Park J, Yu D, Yi J, Lee M, Park H J, Lee J G, Seo J B and Kim N 2019 Improvement of fully automated airway segmentation on volumetric computed tomographic images using a 2.5 dimensional convolutional neural net *Med Image Anal* **51** 13-20

Zapke M, Topf H-G, Zenker M, Kuth R, Deimling M, Kreisler P, Rauh M, Chefd'hotel C, Geiger B and Rupprecht T 2006 Magnetic resonance lung function–a breakthrough for lung imaging and functional assessment? A phantom study and clinical trial *Respir Res* **7** 1-9

Zeng J, Liu Z, Shen G, Zhang Y, Li L, Wu Z, Luo D, Gu Q, Mao H and Wang L 2019 MRI evaluation of pulmonary lesions and lung tissue changes induced by tuberculosis *International Journal of Infectious Diseases* **82** 138-46

Zha W, Fain S B, Schiebler M L, Evans M D, Nagle S K and Liu F 2019 Deep convolutional neural networks with multiplane consensus labeling for lung function quantification using UTE proton MRI *J Magn Reson Imaging* **50** 1169-81

Zha W, Kruger S J, Cadman R V, Mummy D G, Evans M D, Nagle S K, Denlinger L C, Jarjour N N, Sorkness R L and Fain S B 2018 Regional heterogeneity of lobar ventilation in asthma using hyperpolarized helium-3 MRI *Academic radiology* **25** 169-78

Zhang L, Tanno R, Bronik K, Jin C, Nachev P, Barkhof F, Ciccarelli O and Alexander D C Learning to Segment When Experts Disagree 2020a *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, (Cham, 2020),* vol. Series*)* ed A L Martel*, et al.*: Springer International Publishing) 179-90

Zhang L, Tanno R, Xu M-C, Jin C, Jacob J, Cicarrelli O, Barkhof F and Alexander D 2020b Disentangling human error from ground truth in segmentation of medical images *Advances in Neural Information Processing Systems* **33** 15750-62

Zhao X, Li L, Lu W and Tan S 2018 Tumor co-segmentation in PET/CT using multi-modality fully convolutional neural network *Phys Med Biol* **64** 015011

Zhong Y, Vinogradskiy Y, Chen L, Myziuk N, Castillo R, Castillo E, Guerrero T, Jiang S and Wang J 2019a Technical Note: Deriving ventilation imaging from 4DCT by deep convolutional neural network *Med Phys* **46** 2323-9

Zhong Z, Kim Y, Plichta K, Allen B G, Zhou L, Buatti J and Wu X 2019b Simultaneous cosegmentation of tumors in PET-CT images using deep fully convolutional networks *Med Phys* **46** 619-33

Zhou B, Crawford R, Dogdas B, Goldmacher G and Chen A A progressively-trained scale-invariant and boundary-aware deep neural network for the automatic 3D segmentation of lung lesions *IEEE Winter Conference on Applications of Computer Vision, (Hawaii, March 2019),* vol. Series*)*: Institute of Electrical and Electronics Engineers Inc.) 1-10

Zhou X, Takayama R, Wang S, Hara T and Fujita H 2017 Deep learning of the sectional appearances of 3D CT images for anatomical structure segmentation based on an FCN voting method *Med Phys* **44** 5221-33

Zhu J, Zhang J, Qiu B, Liu Y, Liu X and Chen L 2019 Comparison of the automatic segmentation of multiple organs at risk in CT images of lung cancer between deep convolutional neural network-based and atlas-based techniques *Acta Oncol* **58** 257-64

Zhu J-Y, Park T, Isola P and Efros A A *2017.* Unpaired image-to-image translation using cycle-consistent adversarial networks. *In Proceedings of the IEEE international conference on computer vision (Venice, Italy, 22-29 Oct 2017)* 2223-32*.*