Recognition, Retrieval, and Harmonisation for Multicentre Clinical Data Analysis

Yang Nan

Dissertation submitted in fulfilment of the requirements for the degree of:

Doctor of Philosophy

Department of Bioengineering, Imperial College London

August 2024

Declaration of originality

The studies presented in this thesis are my own work at the Department of Bioengineering and National Heart and Lung Institute, Imperial College London, except where stated otherwise through citations, references, or acknowledgements. All sources used have been cited and referenced in a manner consistent with the university's citation guidelines.

Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisors, Dr. Guang Yang and Dr. Simon LF Walsh, for their guidance and advice throughout my research journey. I would also like to extend my special thanks to the members of my thesis committee, Dr. Anil Anthony Bharath and Dr. Huiyu Zhou. Their critical thinking and in-depth insights helped me view my work from various perspectives, leading to a more robust dissertation. Additionally, my heartfelt thanks go to all the members of the Smart Imaging Lab. It has been an honour to work alongside such passionate and dedicated individuals.

I am also grateful to Imperial College London for providing the resources and environment conducive to my research. The facilities and support from the BioEngineering and National Heart and Lung Institute departments have been instrumental in my work.

My journey would not have been possible without the love and support of my family. I owe a deep sense of gratitude to my parents, Baoxin Nan and Qiuhui Ren, as well as my partner Cong Zhao, for their unconditional love, patience, and encouragement. They have been my pillar of strength and a constant source of motivation.

Abstract

Artificial Intelligence (AI) has emerged as a transformative force in medical imaging, significantly enhancing how professionals interpret complex data. Persistent issues mainly include (1) the inherent complexity and clinical variability of medical images, demanding highly adaptive AI for precise recognition and interpretation. (2) Developing effective AI solutions requires large, well-annotated datasets, a process that is time-consuming and resource-intensive for clinicians. (3) Research has largely concentrated on the preliminary stages such as diagnosis and feature extraction, neglecting the AI's potential for prognosis or long-term disease outcomes. (4) Variations in imaging devices, patient status, and protocol specifics limit the reproducibility and generalizability of AI models. This PhD program addresses these obstacles by concentrating on computational methods for clinical tasks such as recognition, retrieval, and data harmonization across multicenter datasets. Initial projects employed supervised learning to assess model performance in complex scenarios, including fine-grained recognition in kidney pathology and airway segmentation in lung CT scans. Innovations introduced include a fine-grained recognition model with uncertainty evaluation for differentiating different glomeruli and a fuzzy attention neural network enhanced by adversarial learning for more accurate airway segmentation, outperforming standard 3D segmentation approaches. To address the scarcity of annotated data, the research explored unsupervised learning with a novel deep mixture model that incorporates constraints to prevent segmentation errors. The AIIB23 challenge, held in conjunction with the MICCAI 2023 conference, established benchmarks for AI in airway segmentation and mortality prediction, helping to identify current research deficiencies. Furthermore, a clinical study involving patients with lung fibrosis compared the efficacy of image-derived biomarkers with conventional clinical metrics in mortality prediction, highlighting AI's potential in clinical applications. The program also explored computational data harmonization, introducing an image retrieval system to categorize and align images based on textural similarities, promising advances for future large-scale clinical studies.

Contents

Declaration of originality **i**

Copyright Declaration **ii**

Acknowledgements **iii**

Abstract **iv**

List of publications **1**

1 Introduction **5**

1.1 AI in digital healthcare 5

1.2 Challenges in digital healthcare 6

1.3 Solutions for the above challenges. 8

2 Fine-grained recognition in kidney pathology **12**

2.1 Abstract 12

2.2 Introduction 13

2.3 Method 16

2.3.1 Enhanced Glomeruli Segmentation with FISS 16

2.3.2 Fine-Grained Lesion Classification via UAAN 20

2.3.3 Uncertainty Aided Data Reconstitution 23

2.3.4 Solutions for the Imbalanced Dataset in Classification 25

2.4 Experiments Settings 26

2.4.1 Datasets 26

2.4.2 Preprocessing and Training Strategies 28

2.4.3 Evaluation Scheme and Metrics 28

2.5 Results 30

2.5.1 FISS Evaluation 30

2.5.2 Uncertainty-aided apportionment network evaluation 31

2.5.3 Slide-wise evaluation 35

2.6 Discussion 37

2.7 Conclusion 40

3 Airway modelling in High-resolution CT 41

3.1 Fuzzy attention neural network for airway segmentation 41

3.1.1 Introduction 41

3.1.2 Related Works 43

3.1.3 Methodology 47

3.1.4 Experiments and Results 54

3.1.5 Discussion 62

3.1.6 Conclusion 67

3.2 Adversarial Transformer for Airway Purification 68

3.2.1 Introduction 68

3.2.2 Related works 70

3.2.3 Method 71

3.2.4 Experiments and Results 74

3.2.5 Implementation Details 74

3.2.6 Ablation Study 80

3.2.7 Discussion and Conclusion 80

4 Unsupervised Learning in Medical AI 82

4.1 Introduction 83

4.2 Related Works 86

4.2.1 Conventional Unsupervised Segmentation 86

4.2.2 Deep Clustering and Mutual Information 86

4.2.3 Deep Generative Models and Log Likelihood 87

4.3 Methodology 89

4.3.1 Overview 89

4.3.2 Deep Constrained Gaussian Network 89

4.3.3 Preprocessing 94

4.4 Experiments 94

4.4.1 Datasets and Training Strategies 94

4.4.2 Experimental Details 95

4.4.3 Experimental Results 97

4.5 Discussion 102

4.6 Conclusion 109

5 Algorithms Benchmarking in Medical Imaging 110

5.1 introduction 110

5.2 Related work 113

5.2.1 Airway segmentation 113

5.2.2 AI-driven prognostication of lung fibrosis 115

5.2.3 Publicly available datasets and open challenges. 116

5.3 Method 117

5.3.1 The AIIB23 Dataset 117

5.3.2 Challenge setup 119

5.3.3 Participants 120

5.4 Results 121

5.4.1 Performance of airway modelling 121

5.4.2 Performance of mortality prediction 123

5.5 Discussion 125

5.6 Conclusion 134

6 A Clinical Translation Study: Prognostication of patients with fibrotic lung disease 136

6.1 Abstract . . . 136

6.2 Introduction . . . 137

6.3 Methods . . . 138

6.3.1 Australian IPF Registry Data. . . . 138

6.3.2 ATM22 Data . . . 139

6.3.3 SABRE- Smart Airway Biomarker Recognition Engine . . . 140

6.4 Results . . . 144

6.4.1 SABRE analysis on the AIPFR. . . . 144

6.4.2 Univariate analysis of airway indicators. . . . 145

6.4.3 Multivariate analysis against disease severity. . . . 146

6.5 Discussion and Conclusion. . . . 147

7 Data Harmonisation, Retrieval and Future Perspectives 149

7.1 Data harmonisation in digital healthcare . . . 149

7.1.1 Introduction . . . 149

7.1.2 Computational data harmonisation: definition, origin, what for and how? 151

7.1.3 Methods . . . 157

7.1.4 Data harmonisation strategies for information fusion . . . 158

7.1.5 Evaluation approaches of the data harmonisation strategies . . . 164

7.1.6 Meta-analysis . . . 166

7.1.7 Research directions . . . 169
7.1.8 Guidance of data harmonisation strategies and metrics . . . 171
7.1.9 Conclusion . . . 172
7.2 Deep Vectorised Hashing for Image Retrieval . . . 173
7.2.1 Introduction . . . 173
7.2.2 Related works . . . 176
7.2.3 Methods . . . 178
7.2.4 Experiments and Results . . . 184
7.2.5 Discussion . . . 189
7.2.6 Conclusion . . . 192

8 Summary and future work **194**

8.1 Summary of Thesis Achievements . . . 194
8.2 Future Work . . . 197

Bibliography **199**

List of Tables

2.1 Composition of The Warwick-QU Dataset 26

2.2 Composition of Renal Pathological Dataset 27

2.3 Glomeruli Distribution of RP-WSIs-G 27

2.4 Comparison Experiments for Gland Segmentation 30

2.5 Ablation study of UAAN 31

2.6 Performance of models on fine-grained classification in IgAN 32

2.7 Ablation Experiments of Uncertainty Factor 34

2.8 Performance of Glomeruli Detection 35

3.1 Details of datasets in this study 55

3.2 Ablation studies of the proposed method on BAS dataset 57

3.3 Comparison experiments on the open dataset BAS dataset 59

3.4 Comparison experiments on our in-house dataset 60

3.5 Comparison experiments of data sampling strategy 64

3.6 Detection rate of different-sized branches 65

3.7 Computational complexity of different methods 66

3.8 Preliminary segmentation generated for training 75

3.9 Comparison experiments on different datasets 75

3.10 Refined segmentation results 78

3.11 Ablation study of the proposed model. 80

4.1 Composition of the MoNuSeg dataset. 95

4.2 Performance of the cell segmentation (MoNuSeg dataset). 98

4.3 Performance of the glomerulus segmentation (RBI dataset). 99

4.4 Degeneration assessment (MoNuSeG and RBI dataset) 101

4.5 Performance of supervised methods (MoNuSeG Dataset) 103

4.6 Ablation studies of constrained weights 106

5.1 Demographics of data used in AIIB23 challenge. 117

5.2 Details of teams that participated in the validation phase and test phase (bolded).121

5.3 Summary of solutions proposed by top-10 teams in Task I. 122

5.4 Top-10 winners in AIIB23 Task I (n=140). 123

5.5 Summary of solutions proposed by top-5 teams in Task II. 124

5.6 Top-5 winners in AIIB23 Task II (n=90). 124

5.7 Detection rate of different sized branches 129

5.8 Model performance on unharmonised data. Colour Indicators for Significant Changes (Wilcoxon signed-rank test, $p<0.05$) in Performance Compared to Original Test Cases. Green represents positive trends, while red signifies negative trends. 130

5.9 Model performance on low-quality data. Color Indicators for Significant Changes (Wilcoxon signed-rank test, p<0.05) in Performance Compared to Original Test Cases. Green represents positive trends, while red signifies negative trends. . . . 131

5.10 Survival analysis for image-derived biomarkers 133

6.1 Baseline Demographics and Centralised MDM Review Characteristics. * indicates multidisciplinary diagnoses. 139

6.2 Patient characteristics for those with an HRCT amenable to SABRE versus those whose HRCT was not. 143

6.3 Segmentation performance of SABRE, † indicates significant differences (p<0.05) evaluated by Wilcoxon signed-rank test. 144

6.4 Univariable analysis Cox proportionate hazards models for airway segments . . . 145

6.5 Multivariate survival analysis of SABRE-based variables controlled for age and gender. 145

6.6 Multivariate analysis including all airway segment volumes. 145

6.7 Relationship between disease severity (using total fibrosis extent scores, DeepDTA score and the CPI) and SABRE-based airway segment volumes. 146

6.8 Multivariate survival analysis of each of SABRE-based airway segment volumes with disease severity 147

7.1 Existing data harmonisation reviews. 150

7.2 Terminology of computational data harmonisation. 152

7.3 Summary of the reproducibility/repeatability studies. 154

7.4 Data scale (image size) in previous studies. The small, middle, large image sizes are defined as $\epsilon \in [0, 256)$, $\epsilon \in [256, 512)$, $\epsilon > 512$, respectively, N/A indicates there is no report of image size. 168

7.5 Details of the BPS and RadIN-CTL Datasets. 185

7.6 Ablation studies 186

7.7 Comparison studies on CIFAR-10, BPS, and RadIN-CTL 187

List of Figures

2.1 Variations in glomeruli within IgAN PAS-stained WSI. Lesions within the sclerosis and crescent are highlighted by yellow and cyan boxes, respectively. 13

2.2 Histograms showcasing the colour intensity values of natural objects and glomeruli, depicted in red, green, and blue colours for the R, G, and B channels. The RoIs are illustrated with yellow curves. 14

2.3 Summary of the proposed method. The fine-grained detection is set up as two sub-tasks, comprising glomeruli location and fine-grained categorization. 16

2.4 Solution of conventional SSIM and the proposed ISS. 18

2.5 Understanding Different Loss Functions. (a) shows simulated glomerulus segmentation, with red circles for predicted glomeruli and blue for actual glomeruli. Red dots mark potential predicted glomerulus locations. Panels (b-d) demonstrate how various loss functions perform with different segmentation qualities and object sizes. 'Ours' is our unique loss function combining dice loss, issim loss, and focal loss. 19

2.6 Overview of the UAAN. 21

2.7 Architecture of the UAAN, with parent branch (a), child branch (c), and the feature apportion layer (b). 22

2.8 Visualization results of loss functions trained with the same training protocol. . 31

2.9 Heatmap visualization of comparison models and different stages in UAAN. The heatmaps of the final prediction of UAAN are marked by black boxes and heatmaps of f_{c-CSN} are marked by the green rectangle. 33

2.10 Data distribution without (a-c) and with (d-f) data reconstitution. (a) and (d) are the estimated probability vectors from the last dense layer, (b) and (e) refer to the raw probability vector from the GMP layer, (c) and (f) include both estimated and raw probability vectors. 34

2.11 Visualization results of the slide evaluation, the red, green, yellow and blue boxes represent annotations/predictions of C, NoA, GS and SS, respectively. The mispredictions were highlighted by the black arrows. 36

3.1 Challenges in airway segmentation and our solutions. (a) shows the varied intensity (details in yellow boxes) of airway trees and data imbalance issues. Green, red, and blue colours refer to TP, FN, and FP predictions. (b) describes our solutions regarding the aforementioned challenges, including fuzzy attention layer, JCAM loss for optimization, and a new metric to evaluate airway prediction continuity and completeness. 42

3.2 Architecture of the proposed fuzzy attention Neural Network. The fuzzy attention blocks are arranged with skip connections between encoders and decoders. Deep supervision was implemented for decoder blocks. 48

3.3 Architecture of (a) fuzzy attention layer, (b) fuzzy attention gate with Gaussian membership functions, and (3) accumulation maps of different panels and skeletonization of airway trees. 50

3.4 Visualization results of the competitive module, including nnUNet, Attention UNet, WingsNet, Rose and the proposed FANN on fibrosis data (first row), COVID-19 data (second row), and BAS data (third row), with the green, red, blue representing the correct airways (TP), missing airways (FN), and airway leakages (FP), respectively. 61

3.5 Challenges and solutions in airway modelling. (left) demonstrates issues like unclear boundaries, disease-related changes, and low image quality (marked in yellow boxes) causing segmentation breaks and overlooked branches (shown in yellow). (right) introduces our solution: an adversarial-based refinement network aimed at thorough and reliable segmentation. We use patch-scale discriminators to address local structural inaccuracies. 68

3.6 The refined airway segmentation model employs a U-Net generator with multi-scale supervision and a patch-scale discriminator. The discriminator is either (a) a PatchGAN with five convolution layers for feature extraction or (b) a Vision Transformer (ViT) that divides the image into patches and transforms them into fixed-size vectors with positional embeddings. 71

3.7 Visualization of the preliminary segmentation, refined segmentation and the ground truth on the BAS dataset (CASE 02) as well as our in-house dataset (CASE 162_01 in fibrosis and CASE RM451 in COVID-19). While the overall structure appears similar, differences are mainly observed at a fine scale. Therefore, zoomed-in views are provided in figures 3.8 to 3.11. 76

3.8 Visualization of a zoomed-in region in the preliminary segmentation, refined segmentation, and ground truth on the BAS dataset (CASE 02) and our in-house datasets (CASE 162_01 in fibrosis and CASE RM451 in COVID-19). . . . 77

3.9 Close-up of bronchioles in CASE02 from the EXACT'09 test set. Top to bottom rows show results from WingsNet, NaviAirway, and Wang et al. respectively. . . 78

3.10 Zoomed-in images of bronchioles in CASE 162_01 from the fibrosis test set. The images are organized with results from WingsNet, NaviAirway, and Wang et al., in top-to-bottom order. 79

3.11 A zoomed-in view of bronchioles in CASE RM451 from the COVID test set, with results from WingsNet, NaviAirway, and Wang et al. presented from top to bottom, respectively. 79

4.1 Challenges in Unsupervised Tissue Segmentation (a) and Our Solutions (b). (a) Displays common issues: empty classes (first row), redundant classes (second row), collapse (third row), and instability (fourth row). Red boxes mark three sections of the raw image, ground truth, and predictions using current methods. P1 and P2 are first and second predictions from experiments (last row). (b) Introduces our centralized constraint deep mixture network for unsupervised segmentation. Our model's effectiveness is shown in green boxes (last row), demonstrating improvements in handling empty classes, redundancies, collapse, and instability. Dice scores for all examples range between [0, 1]. 84

4.2 Variation in Estimated Parameters: (a) Shows a normal distribution for single-class samples. (b) Depicts a mixture model for multi-class samples with two classes (k=2). Note that μ_{est} represents the estimated mean of the mixture model, μ_{obs} is the observed mean of the minibatch data X, and μ_{real} is the ideal mean value of the mixture model. 92

4.3 Dice Score Box Plot from Repeated Experiments. The asterisk (*) marks models with a redundant class (pre-defined classes k=3 for cell segmentation). The dagger (‡) denotes results with highly significant differences (Wilcoxon signed-rank test, P¡0.001) compared to DCGN. Black dots represent outliers, white triangles show mean values and small orange dots are individual samples. 99

4.4 Unsupervised cell segmentation comparisons. * marks models with a redundant class (k=3). Green, yellow, and red colours show true positive, false positive, and false negative predictions, respectively. Red and cyan boxes emphasize regions of interest before and after zooming in. 100

4.5 Comparison of unsupervised glomeruli segmentation. Red bounding boxes indicate empty class problems. In the ground truth, red, blue, and almond colors represent (1) mesangial matrix and basement membrane, (2) intra-glomerular and macula densa cells, and (3) other areas like glomerular capillaries, respectively. 101

4.6 Unsupervised glomeruli segmentation comparison. Red bounding boxes highlight areas with empty class issues. In the ground truth images, red, blue, and almond colors signify (1) mesangial matrix and basement membrane, (2) intra-glomerular and macula densa cells, and (3) other regions such as glomerular capillaries. . . 104

4.7 Failed predictions of glomerular structures and cells. 106

4.8 Illustration of How Our Centralized Constraint Solves Issues: (a) shows handling of missing class and single class domination (collapse), (b) demonstrates addressing redundant class, and (c) deals with instability. Predictions by methods lacking this constraint are marked with dotted circles (left column). Class centroids are represented by yellow diamonds (from methods without the constraint) and yellow stars (from our proposed method). 108

5.1 High-Resolution CT (HRCT) Images: (a) shows patients with nodules from the LIDC dataset, (b) features patients with traction bronchiectasis (indicated by yellow arrows), and (c) highlights patients with honeycombing (marked by red arrows). Honeycombing complicates the extraction of airway trees, resulting in suboptimal airway modelling. 111

5.2 High-Resolution CT (HRCT) Images: (a) shows patients with nodules from the LIDC dataset, (b) features patients with traction bronchiectasis (indicated by yellow arrows), and (c) highlights patients with honeycombing (marked by red arrows). Honeycombing complicates the extraction of airway trees, resulting in suboptimal airway modelling. 112

5.3 Differences in Slice Count in AIIB Datasets. A) Violin plot showing the distribution of slice counts per case in training, validation, and two test sets. B) Comparative visualization of two CT scans with varying slice counts and their respective annotations. Top row: CT scan comprising 1145 slices. Bottom row: CT scan with 140 slices (B1-B6). 118

5.4 ROC of the top-5 teams in Task II. 127

5.5 Heatmap Showing the Difference Between Ground Truth and Predicted Outcomes on the Accumulated Axis Plane. Warm colors highlight false negative predictions (missed segments), while cold colors point out false positive predictions (extraneous segments). 128

5.6 Boxplot of the detected branch rate for terminal and small branches. 129

5.7 Reliant radiomic features in survival status prediction. 134

6.1 Data distribution of cohorts in this study. 137

6.2 Approach used for obtaining airway segments. 141

7.1 Visual Representation of Differences in (a) Radiomics and (b) Pathology Images. (a) Illustrates a lung tumour imaged using the same CT scanner but under six different acquisition protocols (Source: [1]). (b) Displays H&E stained tissue images collected from various sites (Source: [2]). 153

7.2 Scheme of computational data harmonisation. 156

7.3 Procedure of literature selection. 157

7.4 Taxonomy of computational data harmonisation strategies. 159

7.5 Description of adversarial learning methods. 163

7.6 Trend in Publications Over Years Based on Data Properties and Modalities. 'Public data' refers to open-source data accessible online, while 'in-house data' denotes datasets not available on the internet. The percentage shown in the top left subfigure represents the proportion of studies conducted using public datasets. 166

7.7 Harmonization Strategies Across Different Modalities. 'IFL' denotes Invariant Feature Learning approaches, and 'Img Pro' stands for Image Processing approaches. Each pie chart is annotated with the percentage of various submethods, identified by their abbreviations. 167

7.8 Evaluation metrics regarding different modalities. 169

7.9 Flowchart of how to choose harmonisation approaches (left) and evaluation metrics (right). 171

7.10 Challenges in Retrieval Tasks for Medical Images: This includes issues like over-centralization and poor performance on Out-of-Distribution (OOD) data. Over-centralization refers to assigning identical hash codes to samples within the same category without accounting for variations within the class. Weak OOD performance implies the model's inability to identify OOD samples, forcing it to incorrectly associate them with known classes. 174

7.11 Limitations of Measurement Techniques: (a) Relaxed Hamming Distance shows nearly constant patterns for small Hamming distances and sharp patterns for large distances. (b) Cosine Similarity evaluates only the directional aspect of hash vectors, overlooking other characteristics. 179

7.12 DVH integrates three key components: (1) A feature extraction stem with multi-head self-attention (MSA) and hierarchical feature fusion; (2) Vectorised correlation loss (VCL) applied to vectorised hash codes, enhanced by the IntraInterVariationMaximizer (IIVM) mechanism; (3) An Out-of-distribution (OOD) data detection module. Notably, during training, DVH's output h ranges from [-1, 1], with OOD samples incorporated only in the testing phase. 181

7.13 Confusion Matrix for HashFormer, CSQ, DPN, and DVH on the RadioIN-CTL Test Set (n=1200) with an Additional 200 OOD Samples from Various Datasets. The classification was determined by calculating the Hamming distance to the hash center of each class from the training set, considering samples with $h_D < \frac{K}{4} + 1$ as true positives. Labels: 'PD' for Parenchyma Destruction, 'NOR' for Normal, 'NODE' for Nodule, 'ILD' for Interstitial Lung Disease, 'BRON' for Bronchiectasis, and 'AO' for Airspace Opacification. 188

7.14 mAP on CIFAR-10, BPS, and RadIN-CTL datasets. 189

7.15 OOD samples that misrecognized by first (a-e) and second (f, e) steps. 191

7.16 Ranked Retrieval Outcomes by Our method, CSQ, and HashFormer. Displays the top-4 similar (in green background) and dissimilar (in orange background) samples for query images. Small green boxes signify that DVH arranges retrieved samples according to patterns observed in the query images. 192

List of publications

First/co-first author publications

Nan, Y., Li, F., Tang, P., Zhang, G., Zeng, C., Xie, G., ... & Yang, G. (2022). Automatic fine-grained glomerular lesion recognition in kidney pathology. Pattern Recognition, 127, 108648.

Nan, Y., Del Ser, J., Walsh, S., Schönlieb, C., Roberts, M., Selby, I., ... & Yang, G. (2022). Data harmonisation for information fusion in digital healthcare: A state-of-the-art systematic review, meta-analysis and future research directions. Information Fusion, 82, 99-122.

Nan, Y., Tang, P., Zhang, G., Zeng, C., Liu, Z., Gao, Z., ... & Yang, G. (2022). Unsupervised tissue segmentation via deep constrained gaussian network. IEEE Transactions on Medical Imaging, 41(12), 3799-3811.

Nan, Y., Del Ser, J., Tang, Z., Tang, P., Xing, X., Fang, Y., ... & Yang, G. (2023). Fuzzy attention neural network to tackle discontinuity in airway segmentation. IEEE Transactions on Neural Networks and Learning Systems. (Early ACCESS)

Nan, Y., Yang, G. (2024). Deep asymmetric mixture network for unsupervised cell segmentation. IEEE ISBI2024.

Tang, Z., **Nan, Y.**, Walsh, S., & Yang, G. (2023). Adversarial transformer for repairing human airway segmentation. IEEE Journal of Biomedical and Health Informatics. (Early ACCESS)

Last/co-last author publications

Li, H., Tang, Z.,**Nan, Y.**, & Yang, G. (2022). Human treelike tubular structure segmentation: A comprehensive review and future perspectives. Computers in Biology and Medicine, 106241.

Co-author publications

Li H, **Nan, Y.**, Del Ser J, et al. Region-based evidential deep learning to quantify uncertainty and improve robustness of brain tumor segmentation[J]. Neural Computing and Applications, 2023, 35(30): 22071-22085.

Zhang, M., Wu, Y., Zhang, H., Qin, Y., Zheng, H., Tang, W., **Nan, Y.**, ... & Gu, Y. (2023). Multi-site, multi-domain airway tree modeling. Medical Image Analysis, 90, 102957.

Xing, X., Felder, F., **Nan, Y.**, Papanastasiou, G., Simon, W., & Yang, G. (2023). You Don't Have to Be Perfect to Be Amazing: Unveil the Utility of Synthetic Images. arXiv preprint arXiv:2305.18337.

Yeung, M., Rundo, L., **Nan, Y.**, Sala, E., Schönlieb, C. B., & Yang, G. (2023). Calibrating the Dice loss to handle neural network overconfidence for biomedical image segmentation. Journal of Digital Imaging, 36(2), 739-752.

Li, H., **Nan, Y.**, Del Ser, J., & Yang, G. (2023). Large-kernel attention for 3d medical image segmentation. Cognitive Computation, 1-15.

Xing X, Huang J, **Nan, Y.**, et al. CS 2: A Controllable and Simultaneous Synthesizer of Images and Annotations with Minimal Human Intervention[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2022: 3-12.

Submitted papers (first author)

Nan, Y., Xing X, Papanastasiou G, et al. Deep Vectorised Hashing for Image Retrieval (submitted to the IEEE JBHI)

Nan, Y., Xing X, et al. Hunting imaging biomarkers in pulmonary fibrosis: Benchmarks of the AIIB23 challenge (submitted to the Medical Image Analysis)

Nan, Y., Felder F, et al. Prognostication of patients with fibrotic lung disease using quantitative airway analysis from HRCT: a retrospective study (submitted to European Respiratory Journal)

List of Abbreviations

2D	Two dimensional
3D	Three dimensional
AI	Artificial intelligence
CCC	Concordance Correlation Coefficient
CDF	Cumulative Distribution Functions
CECT	Consecutive Contrast-enhanced Computed Tomography
CNN	Convolutional Neural Network
CoV	Coefficient of Variation
CT	Computed tomography
DWI	Diffusion-Weighted Images
FCN	Fully Convolutional Network
FP	False Positive
FN	False Negative
FVC	Forced vital capacity
HRCT	High-Resolution Computed Tomography
HU	Hounsfield unit
ICC	Intraclass Correlation Coefficient
IPF	Idiopathic Pulmonary Fibrosis
IoU	Intersection over Union
MF	Membership Functions
MRI	Magnetic resonance imaging
NLP	Natural Language Processing
OOD	Out-of-distribution
PAS	Periodic Acid–Schiff
PDF	Probability Density Function
PET	Positron Emission Tomography
R^2	R-squared

RoI	Region of interest
SSIM	Structural similarity
SVM	Support Vector Machine
TP	True Positive
TN	True negative
WSI	Whole slide image

Chapter 1

Introduction

1.1 AI in digital healthcare

Artificial Intelligence (AI) refers to the simulation of human cognitive functions by machines, particularly those designed to mimic and replicate human thought processes and learning capabilities. Central to AI is its proficiency in processing substantial volumes of data swiftly and with high accuracy, thus enabling it to execute decisions or make predictions based on this data. AI systems are adept at undertaking tasks traditionally requiring human intellect, including visual perception, speech recognition, decision-making, and language translation [3]. The development of these intelligent systems entails training algorithms on extensive datasets. This training allows AI to identify patterns, execute informed decisions, and incrementally enhance its performance through continuous learning. The adaptability and efficiency of AI have established it as a pivotal, transformative technology in various industries, continually evolving and redefining the capabilities of machines.

In the sphere of biomedical and healthcare research, AI has proven to be an invaluable asset, revolutionizing the methodologies through which medical professionals diagnose diseases, administer treatments, and predict patient outcomes [4]. AI algorithms excel in analyzing intricate biomedical data, which include genetic markers, medical imaging, and comprehensive patient health records. This analytic capability facilitates more precise diagnoses, tailored treatment

regimens, and improved patient care outcomes. For instance, in the domain of medical imaging, AI algorithms have demonstrated a remarkable capacity to identify anomalies, such as tumours or fractures, with greater speed and accuracy than traditional methods employed by human practitioners [5]. Additionally, in the field of drug discovery and development, AI significantly accelerates the process by predicting potential interactions between drugs and various biological systems, thereby assisting in the identification of viable new therapeutic options [6]. The profound ability of AI to analyze and interpret complex datasets is particularly advantageous in medical image analysis, where it is employed across various imaging modalities including X-rays, MRI, and CT scans. Trained on extensive clinical data, AI algorithms have proven their efficacy in detecting subtle anomalies that may be overlooked by clinicians [5, 7]. This thesis mainly focuses on the challenges in multicentre clinical data analysis, including the recognition of pathological changes or basic structures in different manners (supervised/unsupervised), and the harmonisation of multicentre datasets.

1.2 Challenges in digital healthcare

The integration of AI in medical image analysis has offered great precision and efficiency in diagnostics. Despite these significant advancements, several limitations and challenges still remain.

Insufficient learning capability. Central to this achievement is supervised learning, where algorithms are trained on a vast array of labelled medical images [5]. Although efforts have been made to generalize computational tools, certain clinical tasks require the recognition of tiny pathological changes [8]. For instance, the algorithms are required to segment all terminal branches of the vessel and airways within the lung tissue from HRCT scans in the prognosis of patients with fibrotic lung diseases [9, 10]. Another scene is fine-grained recognition, which focuses on distinguishing objects within subordinate categories, such as different subtypes of lesions. It is characterized by relative inter-class variations and small intra-class variations, making it a challenging yet essential problem. For instance, the intra-class agreement for five

senior pathologists in recognizing the crescent and segmental glomerular sclerosis of patients with IgAN is relatively low, with correlation coefficients of 0.66 and 0.46, respectively [11]. Based on this 'unreliable' prior knowledge, the common models (developed on natural image data) cannot achieve the capacity to recognize subtypes of diseases that share similar visual patterns [12]. These subtle anomalies, often critical for early diagnosis and treatment, present a significant difficulty due to their small size and the complexity of the neighbouring biological structures. This complexity necessitates supervised learning techniques that are capable of discerning these fine-grained details with high precision.

Lack of annotation data. Besides the complexity patterns in medical images, another notable challenge is the scarcity of large, well-annotated datasets [4]. Unlike other domains where data might be abundant, medical images often come with constraints such as patient privacy, variability in disease presentation, and the sheer complexity of medical conditions. Moreover, the process of accurately labelling these images requires considerable time and expertise from medical professionals. This requirement not only limits the quantity of data available for training AI models but also introduces the issue of variability in annotations due to subjective interpretations by different experts. Additionally, the process of obtaining expert annotations is time-consuming and resource-intensive, further aggravating the challenges. As a result, there is an urgent need to develop robust self/semi/unsupervised learning techniques that can achieve maximum performance from limited data [13].

Lack of clinical studies. Most research only focused on downstream tasks such as disease diagnosis and pattern recognition, while barely investigating the AI-based clinical studies. This leads to a huge gap between clinicians and AI developers as both sides often lack a common understanding of other's needs and capabilities. Clinicians may not fully grasp the technological intricacies and limitations of AI models, while AI developers might not be deeply familiar with the clinical context and the practical challenges of healthcare settings.

Non-biological variances in multicentre data. Lastly, Multicentre datasets often introduce performance challenges for AI models in medical image analysis due to non-biological variances [14]. These variances include differences in imaging quality due to varied devices

across centres, divergent protocols and operating procedures, and environmental factors during data acquisition. For instance, imaging devices at different centres might have varying resolutions and contrast standards, leading to significant discrepancies in how the same medical condition appears in images from different sources. Additionally, patient positioning, imaging angles, or lighting conditions can vary between centres, further exacerbating the inconsistency in data [15]. These discrepancies pose a challenge as AI models struggle to apply patterns learned from data at one centre to data from another. Consequently, even highly accurate AI models can encounter reduced effectiveness when dealing with data affected by such non-biological variances, underscoring the need for robust models that can accommodate and adapt to these variations across multi-centre data sets.

1.3 Solutions for the above challenges.

This thesis aims to delve into addressing the aforementioned challenges that currently impede AI applications in clinical practice. It lies in developing innovative solutions and methodologies that tackle the challenging issues of AI model capabilities, data annotation scarcity, and non-biological variances in multi-centre data sets.

In order to develop highly efficient supervised learning algorithms under complex clinical scenarios, this thesis mainly involves approaches that employ sophisticated methods including well-designed architectures, enhanced feature extraction strategies, specific optimization functions, and specialized neural network layers. These techniques enable the algorithms to learn from complex patterns in medical images, ensuring that even the smallest pathological changes can be accurately identified. In particular, this thesis first presented a study that employed a hierarchical structure with an uncertainty-aided mechanism for pattern recognition in IgAN. This uncertainty factor could be used as a parameter for reorganizing and cleaning training sets, which improves learning efficiency by reconstituting high-quality data. Additionally, two carefully-designed techniques including the attention mechanism and fuzzy logic have been proposed for airway segmentation. The attention mechanism enhances the model's capability

to focus on relevant features and complex structures within CT images. It allows the model to prioritize certain areas over others, making it more efficient in identifying and delineating smaller airway structures. Fuzzy logic is employed to reduce uncertainty in feature representations. By incorporating fuzzy logic, the model can better handle the ambiguities and variances inherent in medical imaging data, leading to more reliable and consistent segmentation results. This study introduces an efficient airway segmentation method using a unique fuzzy attention neural network and optimization functions that improve the spatial continuity of predictions. Unlike existing attention mechanisms, this method employs a Gaussian membership function with learnable hyperparameters for heterogeneity in different feature channels. Additionally, the study presents a new evaluation metric specifically designed to assess both the continuity and completeness of airway predictions. The high precision and efficiency of fine-grained recognition in handling complex medical images represent a significant leap forward in diagnostic capabilities, contributing to improved patient outcomes and advancing healthcare.

Subsequently, this thesis enhances the ability of AI models to accurately understand complex clinical imaging data, even in the presence of few or no expert annotations. Among various clinical tasks, unsupervised segmentation has become the most challenging one as it requires the model to learn pixel-wise semantic representations without any manual annotations. In particular, it relies on models that can detect patterns, variations, and structures within the original images. For instance, methods like clustering approaches, including deep clustering, have been widely used [16, 17]. This thesis introduces a novel unsupervised learning approach, integrating an end-to-end deep mixture model with a constrained indicator for precise semantic tissue segmentation. This constraint is designed to centralize the components of deep mixture models during the optimization function calculation, thereby significantly reducing the common issues of redundant or empty classes in current unsupervised learning methods. Validated on both public and in-house datasets, our proposed deep-constrained Gaussian network demonstrates significantly better performance compared to current state-of-the-art models.

To better bridge the gap between clinicians and AI researchers, a benchmark study and a clinical translation study were proposed to help researchers better understand AI in lung disease. By introducing the 'Airway-Informed Quantitative CT Imaging Biomarker for Fibrotic Lung

Disease 2023' (AIIB23) challenge, we have not only provided the first publicly available dataset specifically tailored for fibrotic lung disease but also demonstrated the development of computational approaches for complex airway extraction. The results from this challenge underscore the potential of airway-derived biomarkers as strong predictors of survival, offering a hazard ratio greater than 1.5 with a p-value less than 0.0001. This finding represents a tremendous contribution to the field, as it suggests that automatic airway modelling can significantly enhance the prognostication of fibrotic lung diseases compared to existing clinical measurements and AI-based biomarkers. The robustness and generalization abilities of the proposed models set a new benchmark for future research and clinical applications, highlighting the critical role of AI in transforming pulmonary disease diagnosis and management.

In the clinical study, we highlight the significance of traction bronchiectasis's severity as a strong mortality predictor, underscoring its usefulness in monitoring disease progression and assisting in patient risk classification. We introduced a novel deep learning model, SABRE (Smart Airway Biomarker Recognition Engine), for quantifying the severity of traction bronchiectasis in patients with idiopathic pulmonary fibrosis (IPF). This work represents a significant advancement in the field as it addresses the critical need for reliable and objective biomarkers for predicting disease progression and treatment response in fibrotic lung diseases. By leveraging high-resolution computed tomography (HRCT) data, SABRE provides superior prognostic information compared to traditional measures of disease severity, offering robust and reproducible airway-derived biomarkers. These findings underscore the potential of SABRE to transform clinical practice by enhancing the accuracy of disease prognostication, thereby facilitating earlier and more targeted therapeutic interventions. These findings underscore the importance of SABRE as a prognostic tool, aiding clinicians in identifying patients at increased risk of rapid disease advancement and guiding treatment decisions, particularly concerning antifibrotic therapy.

Last but not least, to decrease the non-biological variances in multicentre data, a comprehensive review was performed to help the researchers better report their research findings and select appropriate harmonisation approaches. Then, a content-based image retrieval model emerges as a first step to data harmonization. By identifying and grouping similar images across multicentre

datasets, image retrieval facilitates the harmonization process by ensuring that only cases that include similar contents are harmonised. By employing the hash codes, similar samples can be effectively identified through the calculation of the Hamming distance.

Chapter 2

Fine-grained recognition in kidney pathology

2.1 Abstract

Identifying glomeruli lesions is crucial for the diagnosis and treatment planning in kidney pathology. However, the presence of coexisting glomerular structures, such as mesangial regions, complicates this task. This study outlines a strategy for identifying fine-grained glomeruli lesions in entire slide images. We first introduce a focal instance structural similarity loss, aimed at guiding the model to accurately identify different glomeruli. Following this, we develop an Uncertainty Aided Apportionment Network to facilitate detailed visual categorization, eliminating the need for bounding-box annotations. This dual-branch structure efficiently extracts subclass features from the parent class and generates an uncertainty factor, instrumental in restructuring the training dataset. The slide-wise evaluation results underscore the efficacy of our proposed approach, showing an 8–22% enhancement in mean Average Precision compared to notable existing detection methods.

2.2 Introduction

Immunoglobulin A nephropathy (IgAN), the most common cause of chronic kidney disease globally, is particularly prevalent in Asian regions. Notably, about 40% of IgAN patients progress to end-stage renal disease over several decades [18]. IgAN patients exhibit a diverse range of histological lesions, including crescentic glomerulonephritis, mesangial proliferation, and both global and segmental sclerosis.

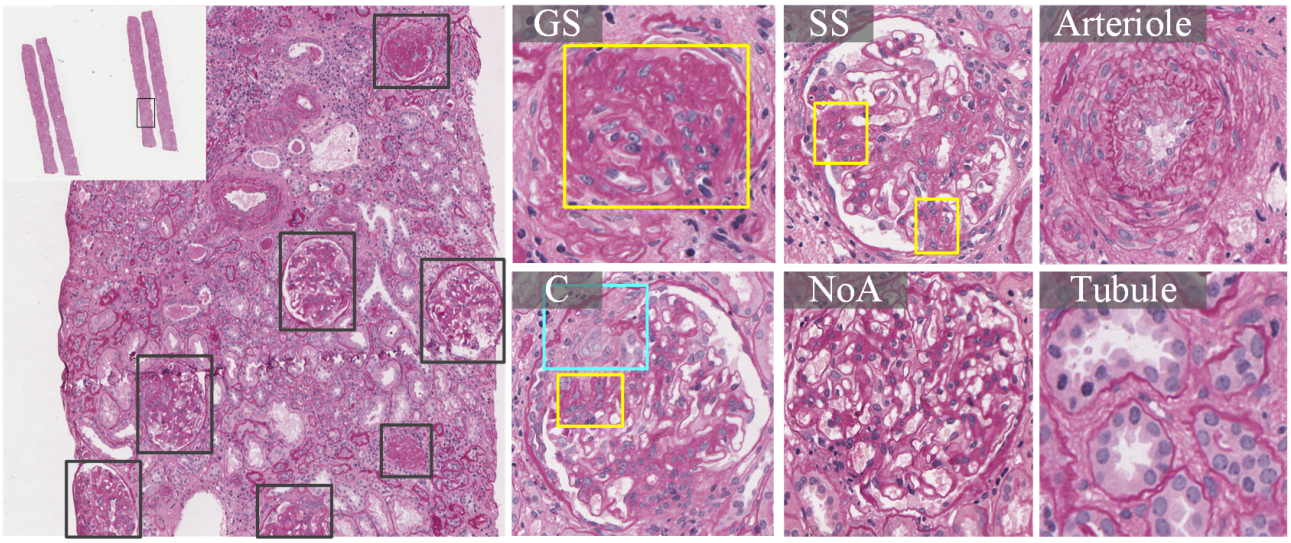


Figure 2.1: Variations in glomeruli within IgAN PAS-stained WSI. Lesions within the sclerosis and crescent are highlighted by yellow and cyan boxes, respectively.

The five primary structures in IgAN, as depicted in Fig. 2.1, are: Neg (Negative structures like tubules and arterioles), GS (Global Sclerosis, where sclerosis encompasses the entire glomerular tuft), C (Crescent, indicating the presence of at least two cell layers filling or encircling Bowman's space), SS (Segmental Sclerosis, where sclerosis affects part of the tuft but not its entirety), and NoA (None of the Above). Earlier research has shown that the extent of sclerosis and crescent formation are crucial for managing treatment decisions and predicting outcomes in IgAN patients [19]. The visual similarities among these pathological changes, due to the collapse and proliferation of capillary loops, pose a significant challenge. For instance, the mesangial matrix is prevalent in GS, C, SS, and NoA types, and the intrinsic glomerular cells closely resemble those in tubules and blood vessels. This resemblance sometimes leads to discrepancies even among pathologists. A previous study highlighted this issue, revealing low intra-class correlation coefficients for identifying segmental sclerosis (0.66) and crescent (0.46)

in IgAN, as assessed by three to five pathologists [11]. This underscores the pressing need for a more objective and quantifiable approach to the identification of pathological changes in IgAN.

The advancement of digital pathology has enabled the scanning of biopsy tissues into whole slide images (WSIs) using micro-scanners. Leveraging these achievements, researchers have been exploring computer-aided diagnosis systems for renal pathology, focusing on glomerulus localization and lesion classification. Current localization methods primarily rely on patch-wise predictions. This includes detecting [20] and segmenting [21–25] numerous glomeruli in large patches, as well as binary classification (glomerulus or non-glomerulus) from smaller patches [26,27]. Meanwhile, most existing studies predominantly classify between sclerosis and non-sclerosis [28–30], with limited focus on more complex lesions like C and SS, indicating a research gap concerning comprehensive analysis of IgAN.

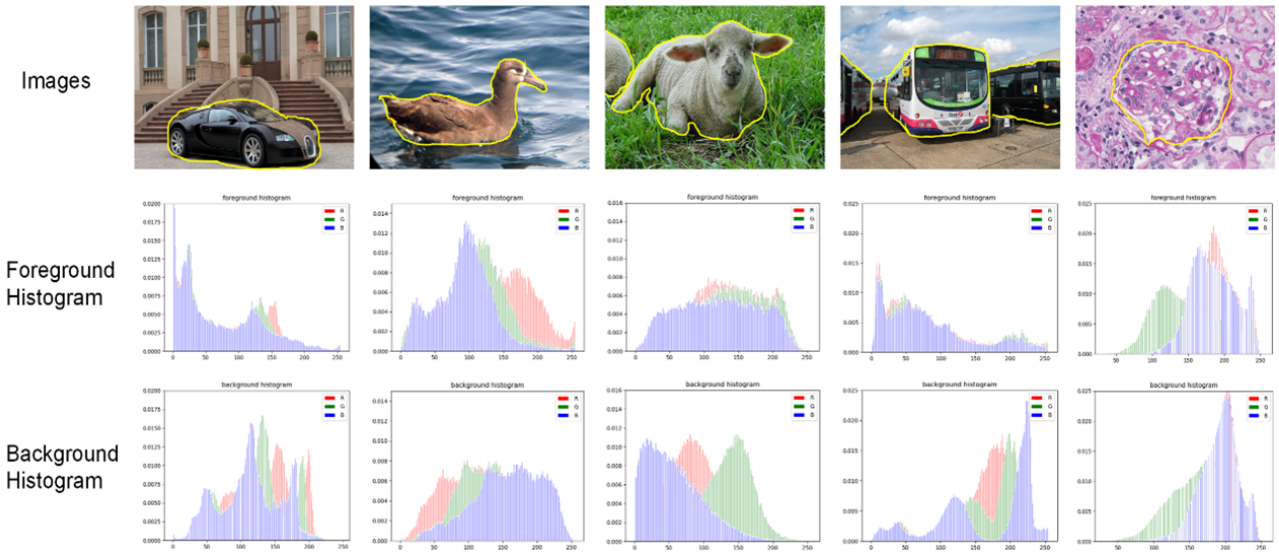


Figure 2.2: Histograms showcasing the colour intensity values of natural objects and glomeruli, depicted in red, green, and blue colours for the R, G, and B channels. The RoIs are illustrated with yellow curves.

Addressing the complex pathological changes in IgAN involves overcoming several technical challenges. Initially, localization of each glomerulus becomes an essential task, followed by accurately classifying glomerular lesions with low intra-class variances. Current fine-grained classification tasks are typically conducted on natural images with relatively uniform backgrounds (Fig. 2.2), compared with the various patterns in kidney pathology [31,32]. Moreover, methods relying on detailed annotations, like bounding boxes or masks for specific sub-regions, are impractical in medical imaging due to the high cost and effort of manual annotations [33].

For effective fine-grained recognition, a network should evaluate and express the 'confidence' level of its predictions. However, existing methods, such as Bayesian deep learning and Monte Carlo dropout, necessitate repeated inferences, making them inconvenient and time-consuming. Lastly, in the realm of medical image analysis, it's vital for the network to mitigate the effects of ambiguous annotations.

In this study, we introduce fine-grained visual recognition (FGVR) in large-scale WSIs, specifically targeting the location and identification of intricate pathological changes in IgAN. This approach diverges from previous methods by focusing on locating and differentiating subordinate targets (C, SS, and NoA in our case). To overcome the limitations [34, 35] of existing detection methods like Mask R-CNN [36] and FCOS [37] – such as weak capacity of fine-grained subcategories, subpar performance in data imbalanced issues, and absence of a confidence coefficient – we split the FGVR task into two phases: segmentation and fine-grained classification of glomeruli. Initially, we introduce a focal instance structural similarity (FISS) loss to achieve precise segmentation outcomes, merging with focal loss for accurate glomerulus boundary delineation. Subsequently, we categorize glomerular lesion types into two groups based on pathological representation and lesion severity: Neg, GS, CSN (a composite of C, SS, and NoA) as the parent class, and C, SS, NoA as the child class. Utilizing this classification, we propose an Uncertainty-Aided Apportionment Network (UAAN) designed to classify complex pathological changes in IgAN. This network generates two sets of predictions corresponding to the parent and child classes, along with their associated uncertainties. These uncertainties serve as confidence coefficients for the network's predictions, facilitating the reconstitution of training data, by identifying missing annotations, mislabeled and hard samples. To elucidate the functionality of our Uncertainty-Aided Apportionment Network (UAAN) and to enhance its interpretability, we employ Gradient Weighted Class Activation Mapping (Grad-CAM) to visualize heatmaps from various network layers [38]. Additionally, we use Uniform Manifold Approximation and Projection (UMAP) to showcase the data distribution both before and after the reconstitution process [39]. We validate our approach with experimental results from the Warwick-QU dataset [40] and our internal renal pathology dataset, conducting evaluations on both a patch-wise and a slide-wise basis.

2.3 Method

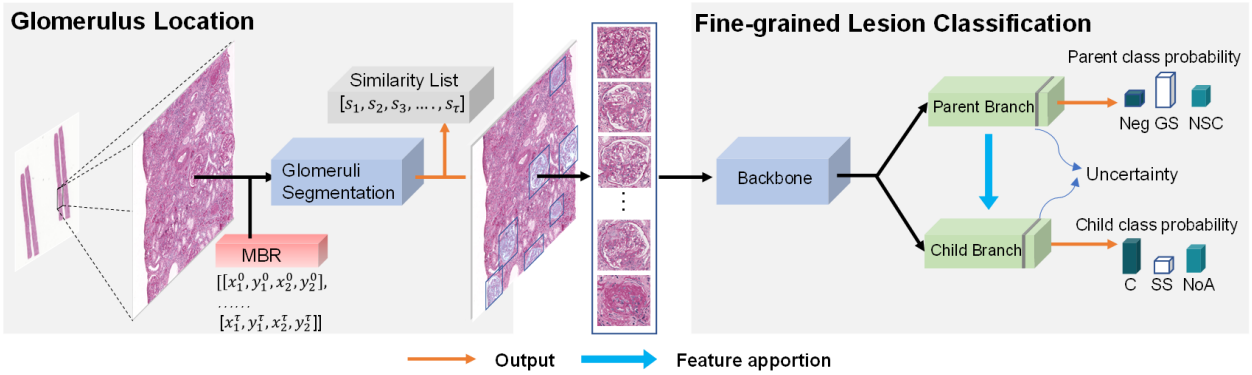


Figure 2.3: Summary of the proposed method. The fine-grained detection is set up as two sub-tasks, comprising glomeruli location and fine-grained categorization.

Given $\mathcal{H} \in \mathbb{R}^{W \times H \times 3}$, a whole slide image that is cropped into N_p large patches, each patch $P \in \mathbb{R}^{w \times h \times 3}$ and its corresponding binary mask $\mathcal{M} \in \mathbb{R}^{w \times h}$, along with the target bounding boxes $B \in \mathbb{R}^{n_i \times 4}$ (the coordinates (x_1, y_1, x_2, y_2) of the minimum bounding boxes of n_i targets) are first cropped for training the segmentation module $\mathcal{G}_\omega$ (Fig 2.3). Enhanced by the designed Focal Instance Structural Similarity loss ($\mathcal{F}_{FISS}$, shown in Eq. 2.2), $\mathcal{G}_\omega$ is tasked with locating all types of glomeruli for subsequent fine-grained classification. Following this, the classification module $\mathcal{D}_\theta$ is trained to identify the categories $\mathcal{J}$ of all segmented objects and generate the associated confidence coefficients C as

$$\mathcal{J}, C = \mathcal{D}(\mathcal{G}(P, B; \omega), \mathcal{F}, \ell_{CSN}; \theta) \tag{2.1}$$

where $\mathcal{F}$ represents the fixed samples, ℓ_{CSN} denotes index lists for CSN. During the inference, the glomeruli are extracted using the segmentation model $\mathcal{G}_\omega$ instead of ground truth masks.

2.3.1 Enhanced Glomeruli Segmentation with FISS

In our research, we utilize a u-shaped auto-encoder/decoder, labelled $\mathcal{G}_\omega$, for segmenting all types of glomeruli. To counteract the potential performance decrease associated with small batch sizes, we implement Group Normalization (GN) [41] rather than the traditional Batch Normalization (BN). Furthermore, we developed a generalized optimization function that as-

sessed the overlapped metric through the instance level. This function can be integrated into various network architectures.

Current segmentation losses heavily rely on overlap measurements such as the Dice coefficient or IoU score [42]. Training only in this way may lead to certain limitations. For instance, every pixel in the image is given equal weight during the loss computation. This results in unstable performance when the sizes of RoIs are various [43]. To address this issue, researchers have explored incorporating additional loss functions with overlapped losses. For instance, the focal loss [44], which has been shown to enhance performance with imbalanced data [45]. However, the focal loss might negatively impact the performance of the more prevalent or dominant class in the dataset. This highlights the need for a balanced approach in loss function design, to improve segmentation accuracy while maintaining stability across different types of samples.

In this study, we suggest an assisted loss FISS to more accurately evaluate the structural similarity between each object's ground truth and prediction in order to overcome these problems. Specifically, the Instance structural similarity (ISS) loss L_{ISS} and focal loss L_F are combined to create the FISS loss that both focus on the completeness and independence of the prediction

$$L_{FISS} = \alpha L_F(p, g) + \beta L_{ISS}(p, g, \mathcal{B}) \tag{2.2}$$

where α, β are weighted parameters, g and p indicate the ground truth and prediction, $\mathcal{B}$ refers to the cropped box of each object in g. α and β were set to 1 to balance the weight for L_F and L_{ISS}, which has been proved in an initial pilot study. The L_{ISS} [46] evaluates each instance's structural integrity and is calculated by

$$L_{ISS} = 1 - (ISS_P + ISS_N) \times 0.5, \tag{2.3}$$

where ISS_P and ISS_N are the instance similarity of the positive and negative samples, respectively. In the case of multi-object photos, ISS first assesses how well each object's prediction matches the actual image. Subsequently, a list containing the instances' similarity indices is obtained; e.g., when four examples are present in the input picture, list $[I_1, I_2, I_3, I_4]$ will be

obtained (shown in Fig. 2.4).

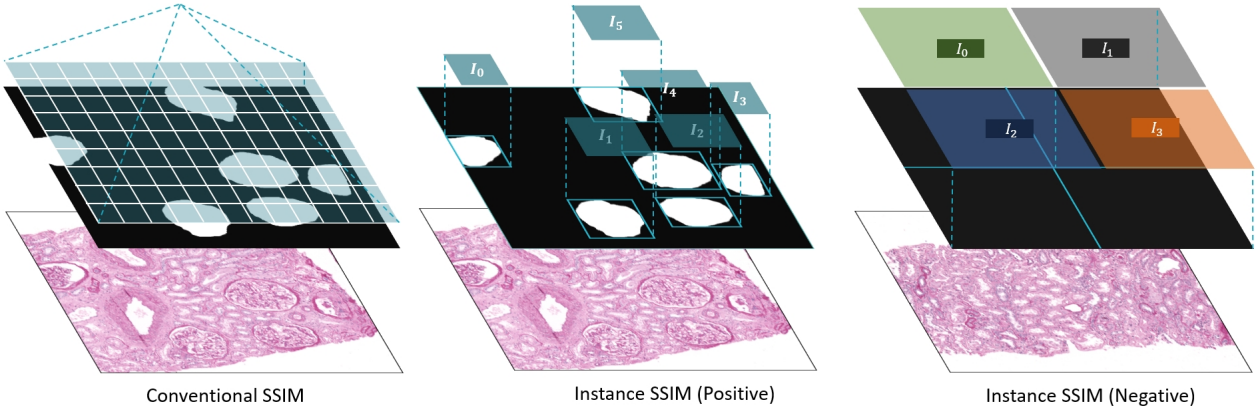


Figure 2.4: Solution of conventional SSIM and the proposed ISS.

Given an input image x with n_i objects, let σ_i and μ_i be the variance and mean of i-th object, the ISS_P is computed as

$$ISS_P = \frac{1}{n_i}\sum_{i=1}^{n_i}\frac{(2\mu_{p_i}\mu_{g_i}+c_{1_i})(2\sigma_{p_i}\sigma_{g_i}+c_{2_i})}{(\mu_{p_i}^2+\mu_{g_i}^2+c_{1_i})(\sigma_{p_i}^2+\sigma_{g_i}^2+c_{2_i})}, \tag{2.4}$$

where g_i and p_i indicate the cropped ground truth mask and the prediction of the i-th instance. In cases where instances were smaller than the sliding window during similarity calculation, zero padding was applied. This ensures that the kernel can slide across the entire object, accommodating instances of varying sizes. For images lacking a Region of Interest (ROI), we divide each image into λ patches ($\lambda = 4$). This division into patches is employed to compute ISS_N, an approach that facilitates the evaluation of similarity in a more segmented and detailed manner, rather than assessing similarity across the entire image directly.

$$ISS_N = \min[SSIM(G_1, P_1), \ldots, SSIM(G_\lambda, P_\lambda)]. \tag{2.5}$$

By integrating L_{Focal} and L_{ISS}, we aim to create a balanced approach that concentrates on imbalanced samples while preserving structural integrity. This combination is anticipated to yield smoother boundaries in the segmentation results and reduce the occurrence of false negative samples.

In order to thoroughly investigate the efficacy of segmenting circular objects with various loss

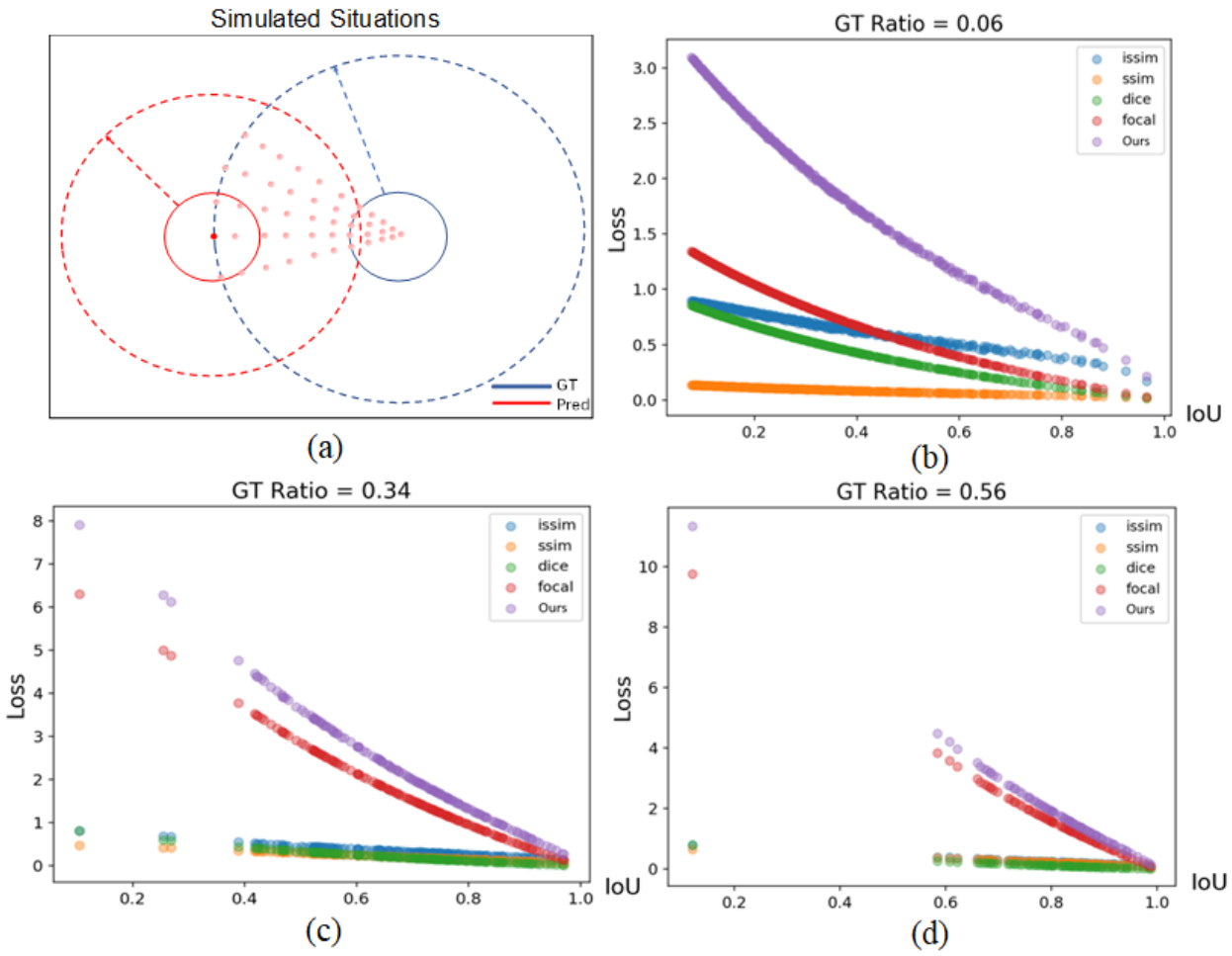


Figure 2.5: Understanding Different Loss Functions. (a) shows simulated glomerulus segmentation, with red circles for predicted glomeruli and blue for actual glomeruli. Red dots mark potential predicted glomerulus locations. Panels (b-d) demonstrate how various loss functions perform with different segmentation qualities and object sizes. 'Ours' is our unique loss function combining dice loss, issim loss, and focal loss.

functions, we simulate thousands of scenarios across diverse ground truth ratios, as depicted in Fig. 2.5. These simulations, reflecting varying sizes of target objects, are designed to assess the performance of different loss functions under a range of ground truth ratios. An image containing both small and large objects can be regarded as a series of mini-batch sub-images, each with a single object. This indicates that if a loss function demonstrates an effective response to a single object under diverse conditions, it should also be effective for images containing multiple targets. In this context, each red and blue circle in the simulations (Fig. 2.5) should be interpreted as the prediction and the ground truth, respectively, while the red points represent the centres of the predicted glomeruli. Firstly, we introduce the concept of the

'ground truth ratio,' which serves as a measure of the target sizes in terms of the image size.

$$GTR = \frac{a_{\text{obj}}}{a_{\text{img}}} \tag{2.6}$$

Subsequently, images with varying Ground Truth Ratios (GTR) are classified into three subcategories based on object size: small-sized objects (GTR not exceeding 20%), middle-sized objects (GTR above 20% but below 50%), and large-sized objects (GTR at least 50%).

In the case of small objects with a 6% area ratio in Fig. 2.5(b), L_{SSIM} demonstrates relative insensitivity compared to L_{Dice} and L_{ISS}. This characteristic renders it less effective for segmenting small objects. L_{ISS}, on the other hand, is more responsive to small objects, even when the prediction is near the ground truth (for instance, we consider an object to be well segmented if its IoU is greater than 0.8). For middle-sized objects, as depicted in Fig. 2.5(c), L_{Dice} and L_{SSIM} exhibit similar penalties, whereas L_{ISS} imposes a much stricter penalty. In the context of large-sized objects shown in Fig. 2.5(d), L_{SSIM} and L_{ISS} have comparable impacts and are adept at accurately reflecting structural integrity.

2.3.2 Fine-Grained Lesion Classification via UAAN

Hierarchical Structure Design Deep neural networks typically perform well when categories are independent and identically distributed. However, the intricate pathological changes in IgAN present a challenge to traditional classification methods. In response, our study introduces a hierarchical architecture specifically for fine-grained classification. This architecture comprises two branches: the parent branch B_p for classifying Neg, GS, and CSN (a collective term for C, SS, and NoA samples), and the child branch B_c for differentiating NoA, SS, and C, as depicted in Fig. 2.6.

The backbone of the Uncertainty-Aided Apportionment Network (UAAN) is an adaptation of the Inception-ResNetV2 [47], modified to extract sophisticated high-level features. This modification includes the integration of Group Normalization (GN) due to the small batch size, PReLU [48] (to enhance generalization capabilities), and a basic convolution block (consisting

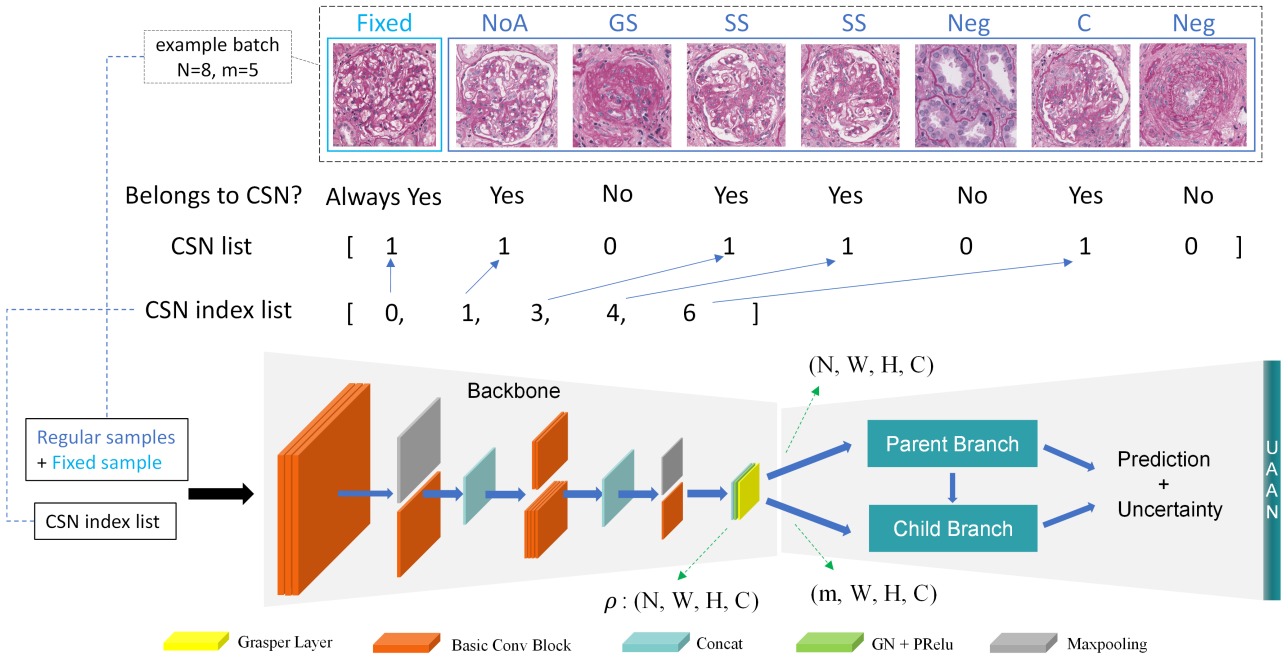


Figure 2.6: Overview of the UAAN.

of a convolution layer, GN, and PReLU). Then, these feature maps are transfered to the sub-branches B_p and B_c via the CSN index lists ℓ_{CSN}. The indices in ℓ_{CSN} denote the CSN samples within the batch $B_c \in \mathbb{R}^{N\times w'\times h'\times c'}$ containing N images with m CSN samples, where w', h', c' represent the width, height, and channel count in the last feature map of the backbone, respectively. Each batch must include at least one CSN sample under this arrangement.

Following the UAAN's backbone, the operations in the B_p and B_c branches are executed to yield predictions for the parent and child classes, respectively. Additionally, a feature apportionment mechanism is introduced between these two branches, designed to facilitate the transfer of valuable information about subcategories. Within both B_p and B_c, Global Max Pooling (GMP) layers are employed to extract the most intense response from each channel in the last feature map. Furthermore, dropout layers are incorporated, compelling the network to develop more robust features. These features prove advantageous when combined with various random subsets of neurons. For a more detailed depiction of the two branches, refer to Fig. 2.7.

Feature Apportionment Mechanism. Feature apportionment is a crucial part of UAAN and is performed via the Grasper layer, as illustrated in Fig. 2.7(b). The Grasper layer segregates CSN features within the latent feature space based on the ℓ_{CSN} index list. In UAAN,

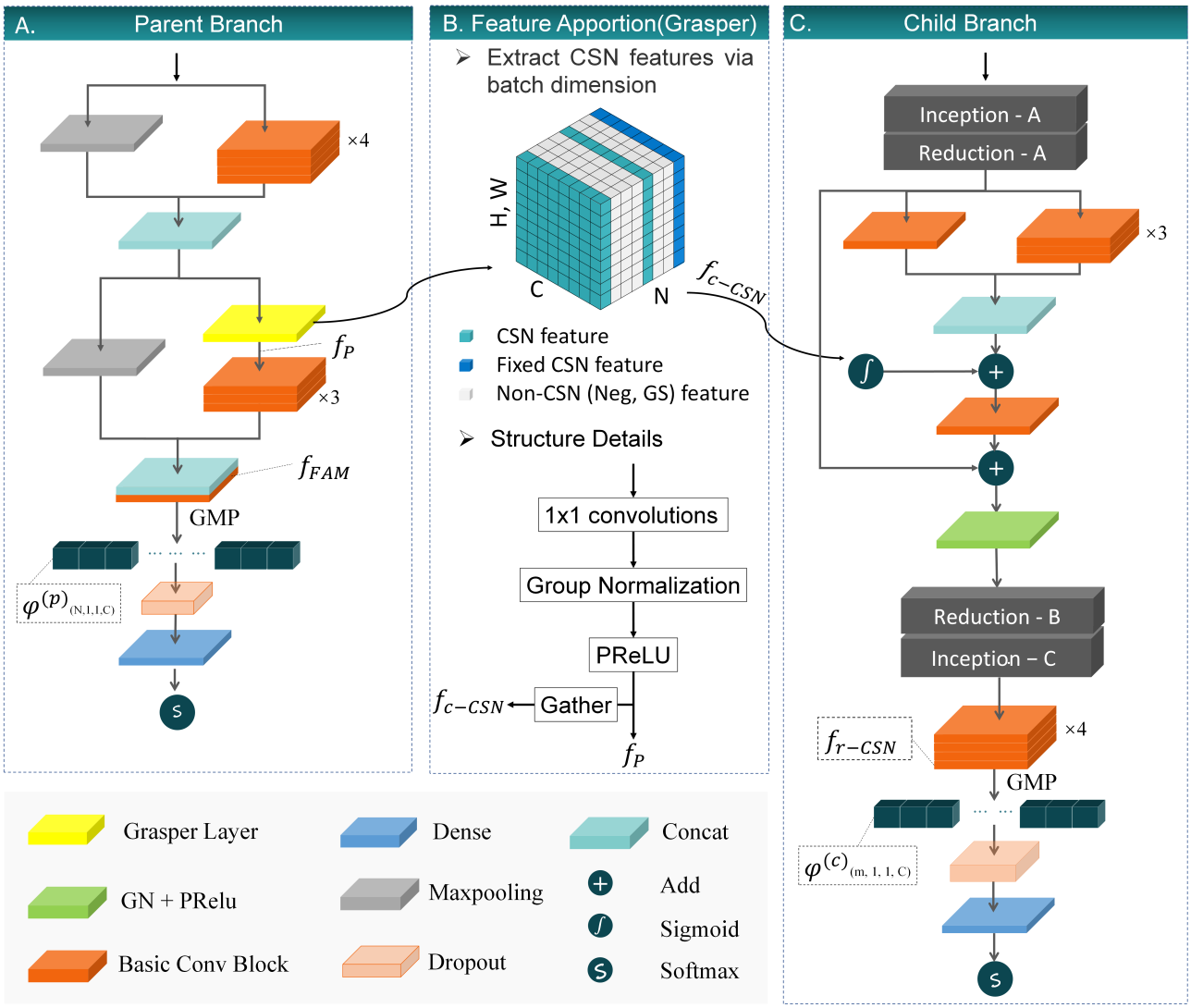


Figure 2.7: Architecture of the UAAN, with parent branch (a), child branch (c), and the feature apportion layer (b).

B_p is specifically designed to discriminate between Neg, GS, and CSN samples. As such, the layers in B_p are expected to be more adept at identifying and learning features unique to these categories. The CSN feature isolated by B_p, labelled as f_{c-CSN}, is presumed to represent common attributes of C, SS, and NoA that are relatively rare in GS and Neg samples. These attributes could include specific structures like capillary loops and mesangial regions. On the other hand, B_c is focused on the finer task of distinguishing between the child classes C, SS, and NoA. The layers in B_c are thus fine-tuned to extract nuanced features that are specific to each of these subcategories. By transferring f_{c-CSN} from B_p to B_c through a sigmoid activation layer, UAAN ensures that B_c not only harnesses these discriminative features but also incorporates the shared characteristics identified by B_p. This strategy allows the network to benefit from

a more comprehensive feature set, enhancing its ability to accurately classify and differentiate between the various subcategories within the CSN group.

2.3.3 Uncertainty Aided Data Reconstitution

'Hard samples' are defined as samples that share common features or morphological similarities among the C, SS, and NoA categories. For instance, a single glomerulus might simultaneously display mesangial proliferation and crescent changes, leading to subjective judgments by experts based on the severity of these changes. The inclusion of such hard samples in the training dataset can impede the network's ability to effectively discern and extract features of real subcategories, even with a substantial amount of data. In natural image analysis, 'unreliable' or hard samples are typically identified, relabeled, and augmented to bolster the performance of computational models. However, this approach is less feasible in biomedical image analysis due to the high cost and effort involved in expert annotation. Moreover, some samples might not contribute constructively to the training of a robust model, potentially causing detrimental fluctuations in classification boundaries.

To mitigate this challenge, we introduce an uncertainty assessment mechanism. This method outputs a confidence coefficient of the network's predictions, which helps in excluding 'unreliable' and hard samples from the initial training dataset. The confidence coefficient is calculated using the logits from the softmax layers $s^{(p)}(s^{(c)})$, and the vectors from the Global Max Pooling (GMP) layers $\phi^{(p)}(\phi^{(c)})$ in both B_c and B_p branches. The dense and GMP layers are tasked with dimension reduction, while the final prediction is typically made in the last dense layer, which integrates 'evidence' from previous layers [49]. The softmax layer's inputs were regarded as high-dimensional feature vector representations, which are close to the estimated probability density P_e. Therefore, the feature vectors $s^{(p)}$ and $s^{(c)}$ are interpreted as the estimated probability densities $p_e^{(p)}$ and $p_e^{(c)}$, respectively.

$$p_e^{(p)}, p_e^{(c)} = \phi(\mathcal{D}(I_g, \mathcal{F}, \ell_{CSN}; \theta)), \tag{2.7}$$

where I_g represents glomeruli segmented by $\mathcal{G}_\omega$, ϕ denotes the softmax activation function. The outputs $\phi^{(p)}$ and $\phi^{(c)}$ from the last convolution layer are regarded as the representations of I_g, whose distributions are close to the original distribution of the category that I_g belongs to. Therefore, we utilize $\phi^{(p)}$ and $\phi^{(c)}$ to calculate the raw PDFs to approximate the real distribution $P_r^{(p)}$ and $P_r^{(c)}$. Assume the glomeruli feature follows the Gaussian distribution, the feature vectors $\phi^{(p)}$ and $\phi^{(c)}$ are regarded as a sampling from latent feature space. This perspective allows for a more nuanced understanding and processing of the feature distributions within the UAAN, facilitating a more accurate and probabilistically informed approach to fine-grained classification. Given the mean μ and variance σ of feature embeddings $\phi^{(p)}$ ($\phi^{(c)}$), the raw probability density function P is calculated through

$$P_r(\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(k-\mu)^2}{2\sigma^2}}, \tag{2.8}$$

where K is the number of classes.

Given the estimated probability density function P_e and P_r, the uncertainty can be assessed through

$$U(P_r, P_e) = 1 - \frac{\sum_{k=0}^{K-1}(p_r^k - \bar{P}_r)(p_e^k - \bar{P}_e)}{\sqrt{\sum_{k=0}^{K-1}(p_r^k - \bar{P}_r)^2}\sqrt{\sum_{k=0}^{K-1}(p_e^k - \bar{P}_e)^2}}, \tag{2.9}$$

where P_r and P_e are computed by

$$\bar{P}_r = \frac{1}{K}\sum_{k=0}^{K-1} p_r^k, \quad \bar{P}_e = \frac{1}{K}\sum_{k=0}^{K-1} p_e^k. \tag{2.10}$$

By using this uncertainty factor, data reconstitution is employed to identify and remove 'unreliable' and hard samples from the dataset. After training the model, we inferenced on the raw training data D_r for data reconstitution. The label is considered 'credible' when there is a strong correlation between P_r and P_e. Using an uncertainty threshold of 0.5, we identified and removed 2.33% of NoA, 35.91% of SS, 15.15% of C, and 6.42% of GS samples from the training data to form a refined training set D_c, consisting only of 'reliable' samples. Notably, the evaluation datasets (valid and test) remained the same during this process.

2.3.4 Solutions for the Imbalanced Dataset in Classification

The structure of UAAN requires the inclusion of at least one CSN sample in each batch, which may not align well with the imbalanced datasets. To overcome this challenge, we introduce a strategy of a "fixed sample" mechanism to ensure each mini-batch contains a CSN sample. These "fixed" samples (one each from the C, SS, and NoA categories) are combined with regular training samples for network input. For example, in a setup where the number of input images per iteration (N) is 8, we would have 7 regular samples and 1 "fixed" sample. In a mini-batch with [fixed sample, NoA, GS, SS, SS, Neg, C, Neg], N would be 8, m (the number of CSN samples) would be 5, and the 'CSN index list' would be [0, 1, 3, 4, 6], as illustrated in Fig. 2.6. Importantly, to avoid overfitting, these "fixed" samples are not factored into the loss calculation and backpropagation process.

Additionally, we employ a weighted cross-entropy loss, in which the weight δ_k of kth class is given by

$$\delta_k = \frac{N_T/n_k}{\sum_{k=1}^{K}(N_T/n_k)}, \tag{2.11}$$

where K is the total number of classes, n_j is the number of samples in the kth class, and N_T denotes the number of all samples. Accordingly, the loss function for training UAAN is formulated as

$$\mathcal{L} = -\sum_{k=1}^{K} \delta_k t_i \log p_k, \tag{2.12}$$

where s_{p_k} denotes the softmax probability for the k^{th} class, and t_k is the true label. Additionally, we implement data augmentation strategies to address the severe class imbalance issue. These strategies include randomized hue adjustment (modifying the hue of input images within a range of $[-0.15, 0.15]$), randomized horizontal flips, randomized vertical flips, and randomized brightness adjustments (with an adjustment range of $[0, 0.15]$). These augmentation techniques help in creating a more robust and generalized dataset for effective training of the network.

2.4 Experiments Settings

2.4.1 Datasets

Datasets involved in this study are Warwick-QU and the in-house Renal Pathological WSIs (RP-WSIs). The Warwick-QU dataset, characterized by its significant variation in glandular morphology across different histological grades, serves as an excellent testbed for evaluating the effectiveness of the Focal Instance Structural Similarity (FISS) loss. The RP-WSIs is employed to assess the performance of the UAAN. This is due to the considerable variations in pathological changes observed in IgAN cases, which make it a suitable dataset for testing the effectiveness in handling inter-class variations. The specifics of these two datasets are comprehensively outlined in Table 2.1.

Table 2.1: Composition of The Warwick-QU Dataset

Dataset	Warwick-QU	RP-WSIs
Type	Colorectal cancer	IgA nephropathy
Resolution	20X (0.62005 µm/pixel)	40X (0.2468 µm/pixel)
Scanner	Zeiss MIRAX MIDI	Leica Aperio ScanScope AT Turbo
Number of samples	165 images	400 whole slide images
Staining	Hematoxylin and Eosin	Periodic acid–Schiff
Number of patients	n/a	400

Warwick-QU Dataset: The Warwick-QU dataset, employed in the GlaS challenge at MICCAI 2015, comprises 85 training images and 80 test images (detailed in Table 1), with annotations marking the outer margins of glands [40]. This challenge was divided into two parts: Part A, with 60 images for off-site testing, and Part B, with 20 images for on-site testing. However, this division introduces a certain bias in the performance evaluation of segmentation algorithms, as equal weight is given to performance across both parts. The images in this dataset are of a resolution of 775×522 pixels.

RP-WSIs: For this dataset, 400 PAS-stained WSIs from patients diagnosed with IgAN were selected. These samples, sourced from the National Clinical Research Center of Kidney Diseases at Jinling Hospital, were collected between 2016 and 2019 using various operators and scanners.

The size of the WSIs ranges from 10,000×30,000 to 30,000×50,000 pixels. Specifically, the dataset includes 300 WSIs, with 9700 patches (1024×1024) for training and validation, and 3700 patches for testing. All data were anonymized in accordance with the Declaration of Helsinki. The initial annotations of the outer margins and subcategories of glomeruli were performed by experienced junior pathologists using Aperio Image Scope software (version 1.2.3.3), and subsequently verified by three senior pathologists. The inter-observer agreement for glomeruli lesion evaluation is high (Kendall's W = 0.802, Fleiss' kappa = 0.761, p <0.0001). Based on these annotations, the RP-WSIs were divided into three sub-datasets, categorized by different scales, as shown in Table 2.2.

Table 2.2: Composition of Renal Pathological Dataset

Level	Training Samples	Test Samples
Slide (RP-WSIs-S)	300	100
Patches (RP-WSIs-P)	9700	3700
Glomeruli (RP-WSIs-G)	23113	8529

* **P** indicates the large patches used for the segmentation task, **G** represents the glomeruli patches used for the classification task, and **S** is the whole slide images used for the detection task.

The training and test sets for RP-WSIs-P and RP-WSIs-G were extracted from the corresponding slides, respectively. RP-WSIs-P was comprised of nearly 13,000 patches with 1024 1024 pixels, while RP-WSIs-G (details in Table 2.3) was set up by cropping the minimum bounding boxes of glomeruli.

Table 2.3: Glomeruli Distribution of RP-WSIs-G

	NoA	SS	C	GS	Neg	Total
Training D_r	10401	1066	818	2223	8605	23113
Training D_c	10161	683	694	2080	8409	22027
Test	3764	639	444	1017	2665	8529

2.4.2 Preprocessing and Training Strategies

For preprocessing each input image x, the standard min-max normalization was utilized

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}}. \tag{2.13}$$

All trainable parameters were initialized by Glorot uniform initializer. For comparative analysis, networks were trained using the Adam optimizer on an NVIDIA Tesla V100 GPU for 100 epochs, starting with an initial learning rate of 0.0001 and a decay rate of 0.96 per epoch. To ensure an unbiased evaluation, all models were trained from scratch without any transfer learning techniques.

2.4.3 Evaluation Scheme and Metrics

Patch-wise Evaluation are conducted to validate the effectiveness of the FISS loss and UAAN. The evaluation of FISS loss is focused on demonstrating its efficacy without the influence of other factors such as network architecture, data augmentation, or specific training strategies. To provide persuasive results, ablation studies are carried out without altering the network architecture. The loss functions included in experiments are the Cross-Entropy loss (CE), Dice coefficient loss (DL), Tversky loss, Focal loss (FL), Structural Similarity Index Measure (SSIM) loss, Lovasz loss, Instance Structural Similarity loss (ISS), and our Focal Instance Structural Similarity loss (FISS). For the assessment of these loss functions, we used the Dice coefficient as the evaluation metric, where P denotes the prediction output from the segmentation network and G represents the ground truth.

$$\text{Dice} = \frac{2|P \cap G|}{|P| + |G|} \tag{2.14}$$

This evaluation is applied to both the Warwick-QU and RP-WSIs datasets, providing a comprehensive understanding of the performance and impact of each loss function.

The assessment of UAAN includes five aspects. To begin with, ablation experiments were used

to assess the efficacy of several strategies based on UAAN, involving data reconstitution and feature apportionment. In particular, all comparison models were optimized by weighted cross-entropy loss and trained using the same strategy. Next, we explored the performance of all comparisons on both raw dataset D_r and the rebuilt dataset D_c. Let K be the total number of classes and N represents the number of test samples. The micro-accuracy and macro-accuracy are calculated by

$$\mathrm{Acc}_{\mathrm{micro}} = \frac{TP + TN}{N} \tag{2.15}$$

$$\mathrm{Acc}_{\mathrm{macro}} = \frac{1}{K} \sum_{i=0}^{K} \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \tag{2.16}$$

where TP_i, FP_i, TN_i, and FN_i are the numbers of true positive, false positive, true negative, and false negative samples regarding the i-th sample, respectively. Additionally, we used Grad-CAM [38] to generate heatmaps for comparisons and certain layers inside the UAAN. Grad-CAM utilises a back-propagated signal to allocate RoIs within images for category prediction. Fourthly, we examined the impact of uncertainty thresholds by removing 'unreliable' data from D_r using various thresholds and subsequently have these data reviewed by experts. At last, we use Uniform Manifold Approximation and Projection (UMAP) [39] to visualize the data distribution in the latent space, both before and after the uncertainty-aided data reconstitution. The vectors used for UMAP were derived from the GMP and Dense layers.

Slide-Wise Evaluation. The performance of our proposed protocol on WSI was evaluated through Average Precision (AP) and its mean value. The ground truth bounding boxes (used for mAP calculation) were generated based on annotations from clinicians. During the assessment, the segmentation was first performed to locate all glomeruli (stage-I), followed by the fine-grained classification (stage-II), where the inputs for classification were the cropped glomeruli obtained from the prediction in stage-I. All segmented entities even for incorrect candidates (e.g., tubules and arteries), except for very small objects (with an area less than 100 pixels), were fed into the UAAN. Then, UAAN took over to predict the categories of these cropped rectangles, and the probability obtained from stage-II was utilized as the confidence threshold for calculating the mAP. This approach aligns with the evaluation typically used in detection-based methods, where objects exceeding a specified Intersection over Union (IoU) threshold are

classified as true positive samples. In our experiments, we set this threshold at 0.5, a common standard in object detection tasks. We conduct a comparative analysis of the AP_{50} (mean average precision at an IoU threshold of 0.5) for each subcategory, contrasting our scheme with various other detection methods. These include Mask Scoring R-CNN [50], Mask R-CNN [36], Cascade R-CNN [51], and FCOS [37]. A critical aspect of this comparison is that all models, including ours and the others mentioned, are trained from scratch.

2.5 Results

2.5.1 FISS Evaluation

The experimental results on RP-WSIs-P and Warwick-QU are presented in Table 2.4.

Table 2.4: Comparison Experiments for Gland Segmentation

	Warwick-QU			RP-WSIs-P
Loss Function	Dice_A	Dice_B	$\text{Dice}_{A\&B}$	Dice
DL	0.8512	0.8819	0.8665	0.742
DL+CE	0.8888	0.8917	0.8902	0.879
DL+FL	0.8949	0.8938	0.8943	0.848
DL+Tversky	0.8891	0.8701	0.8796	0.882
DL+Lovasz	0.8965	0.8683	0.8824	0.884
DL+SSIM	0.8825	0.8838	0.8831	0.885
DL+ISS	0.9036	0.8835	0.8935	0.897
DL+FISS	0.9049	0.8940	0.8995	0.906

Table 2.4 reveals that the model utilizing a combination of Dice and FISS loss obtains the highest performance, with the Dice scores of 0.9022 and 0.906 on the Warwick-QU and RP-WSIs datasets, respectively. This performance is closely followed by the model employing Dice and ISS loss. A notable observation is the significant variance in the performance of the model trained with Dice loss and Focal loss, ranking third on the Warwick-QU dataset but only seventh on the RP-WSIs dataset. This discrepancy can be attributed to the necessity for manual adjustments of specific parameters in the focal model. Conversely, the network trained solely with Dice loss exhibited the least favourable results on both datasets (Fig. 2.8). This

underperformance is likely due to the presence of hard examples in the datasets, such as glands with various histologic grades and arterioles resembling global sclerosis in glomeruli.

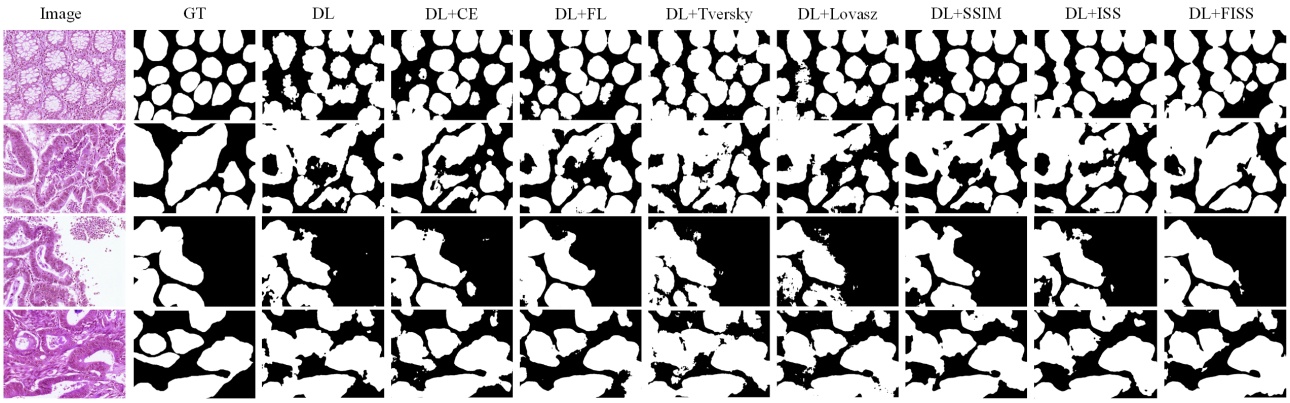


Figure 2.8: Visualization results of loss functions trained with the same training protocol.

2.5.2 Uncertainty-aided apportionment network evaluation

Ablation Studies. The ablation studies of the UAAN are detailed in Table 2.5, highlighting the impact of various techniques on the model's performance. Notably, the implementation of feature apportionment led to a significant improvement in micro-accuracy, achieving a relative gain of 3.34%. Similarly, the incorporation of uncertainty-aided data reconstitution effectively enhanced macro-accuracy, with a gain of 2.58%. The most substantial improvement was observed when implementing both feature apportionment and data reconstitution, with 5.57% micro-accuracy and 4.69% macro-accuracy gains compared to the baseline.

Table 2.5: Ablation study of UAAN

Feature Apportionment	Uncertainty	$\text{Acc}_{\text{micro}}$	$\text{Acc}_{\text{macro}}$
×	×	0.8960	0.8677
×	✓	0.9132	0.8935
✓	×	0.9294	0.8723
✓	✓	0.9517	0.9146

Comparison of Fine-Grained Classification. UAAN demonstrates exceptional proficiency in handling fine-grained classification tasks, surpassing all other competing architectures

(Table 2.6). The Uncertainty-Aided Apportionment Network (UAAN) presented a remarkable performance on both the raw dataset D_r and the reconstituted dataset D_c, achieving a micro-accuracy of 95.07% (a roughly gain of 3-11%) and macro-accuracy of 91.28% (a relative improvement of 12-26%). This performance notably exceeds that of other advanced architectures, surpassing EfficientNet-B6 by nearly 4%, SE-ResNeXt-101 by 5.2%, and both Inception-ResNet-V2 and DenseNet-121 by approximately 8%. Furthermore, the results demonstrate that reconstituting the training set using the uncertainty factor significantly enhances the performance of all comparison models. This improvement is reflected in a micro-accuracy gain of 1% to 4%, underscoring the effectiveness of incorporating uncertainty factors in the training process to boost overall model performance.

Table 2.6: Performance of models on fine-grained classification in IgAN

Model	$\text{Acc}_{\text{micro}-D_r}$	$\text{Acc}_{\text{micro}-D_c}$	$\text{Acc}_{\text{macro}-D_c}$
ResNet-50 [52]	0.8495	0.8717	0.7773
ResNeXt-50 [53]	0.8294	0.8474	0.6568
DenseNet-121 [54]	0.8635	0.8817	0.7757
EfficientNet-B6 [55]	0.9092	0.9235	0.7899
Inception-ResNet-v2 [47]	0.8507	0.8830	0.7396
Mean Teacher (70% annotation) [56]	/	0.7605	0.6067
DCL [57]	0.8176	0.8373	0.6770
SE-ResNeXt-101 [58]	0.8943	0.9106	0.7638
DenseNet+LSTM-GCNet [59]	0.9120	0.9295	0.8263
UAAN (ours)	**0.9294**	**0.9517**	**0.9146**

Explainable ability. Fig. 2.9 shows the heatmaps generated by Gradient-weighted Class Activation Mapping (Grad-CAM) for both the UAAN and other comparisons. The final predictions made by UAAN are delineated by black bounding boxes. Pathological changes categorized as GS, C, and SS are highlighted with yellow curves (in the first column in Fig. 2.9). It is crucial to note that the label "None" in Fig. 2.9) signifies the absence of heat maps for these two categories in the child branch. A critical observation from the heatmaps is that the areas marked by green rectangles, which represent f_{c-CSN}, show a strong correlation to the common features of CSN, such as capillary loops, intrinsic glomerular cells, and the mesangial matrix. This illustrates that during the training process, these specific features are effectively captured by the grasper layer and then transmitted to the child branch. This mechanism serves

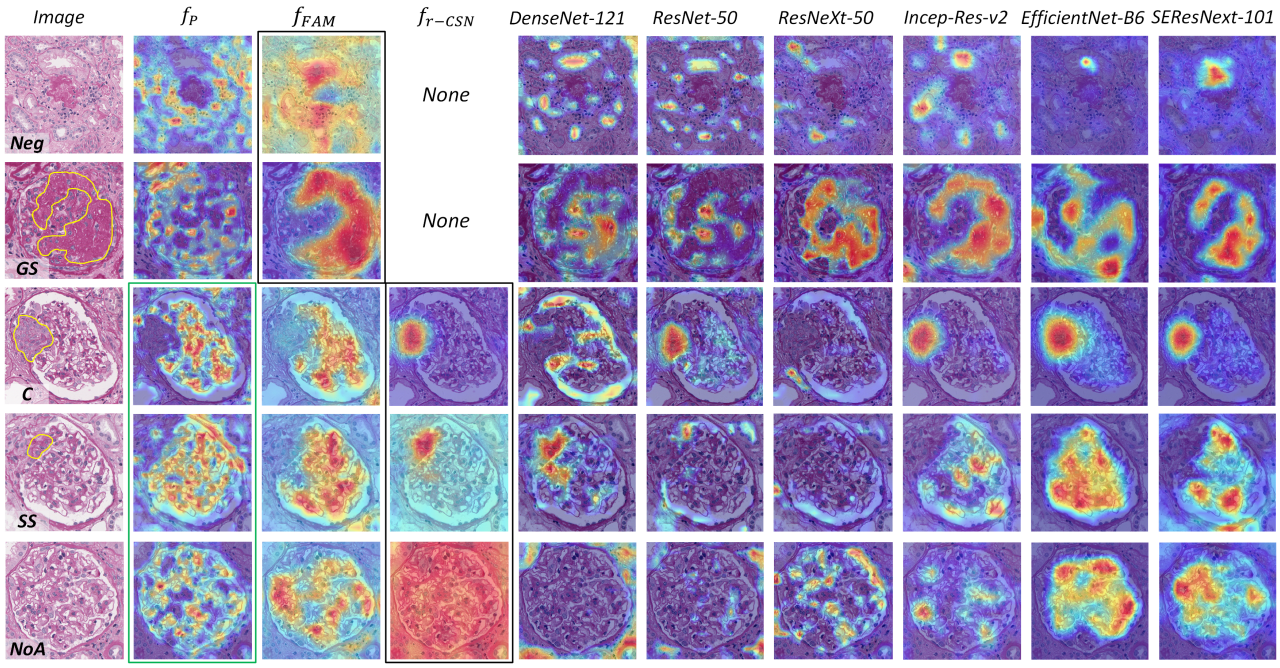


Figure 2.9: Heatmap visualization of comparison models and different stages in UAAN. The heatmaps of the final prediction of UAAN are marked by black boxes and heatmaps of f_{c-CSN} are marked by the green rectangle.

to amplify the shared information across these categories, thereby enhancing the model's ability to accurately recognize and classify these complex pathological changes.

The highlighted regions in f_{FAM} provide critical insights into the areas where the network focuses its attention, significantly influencing the predictions among CSN, GS, and Neg. For instance, the highlighted areas in $f_{\text{FAM}-SS}$, $f_{\text{FAM}-NoA}$, and $f_{\text{FAM}-C}$ are predominantly concentrated on the capillary loops. Conversely, the $f_{\text{FAM}-GS}$ primarily targets the mesangial region, while $f_{\text{FAM}-Neg}$ is focused on tubules and arterioles. These specific structures are key distinguishing features used by pathologists to differentiate Neg, GS, and CSN. As the output of B_c (the child branch) only includes NoA, SS, and C categories, heat maps for GS and Neg are not present in f_{r-CSN} and are denoted as 'None' in Fig. 2.6. Observing f_{r-CSN}, we see a high variance in the highlighted areas: $f_{r-CSN-C}$ focuses on proliferating cells in Bowman's space, $f_{r-CSN-SS}$ on the increased mesangial matrix and collapsed capillary loops, and $f_{r-CSN-NoA}$ on global information. This global focus is crucial in NoA samples, where the network evaluates each corner of the sample to determine the presence of any pathological changes.

Ablation Studies of Uncertainty Thresholds. The raw mislabelled rate in D_r is 6.7%,

evaluated by three senior clinicians, who randomly examined 100 samples from each class. Leveraging the uncertainty indicator U, we identified that 1-15% of the training set, corresponding to uncertainty values in the range [0.3,0.9], could be classified as "unreliable" data. We removed these "unreliable" samples from D_r and conducted a second expert review to check whether these images were real hard samples or mislabelled samples.

Table 2.7: Ablation Experiments of Uncertainty Factor

U_{thresh}	N/A	0.3	0.4	0.5	0.6	0.7	0.8	0.9
N_{select}	400	3217	1729	1086	718	469	262	140
N_m	27	453	318	215	145	99	51	26
$R_m(\%)$	6.7	14.1	18.4	19.8	20.2	21.1	19.5	19.2

* N_{select} refers to the number of selected data based on different uncertainty threshold, R_m and N_m are the percentage and number of the selected data points that were identified as mislabelled, respectively. 'N/A' indicates no threshold applied.

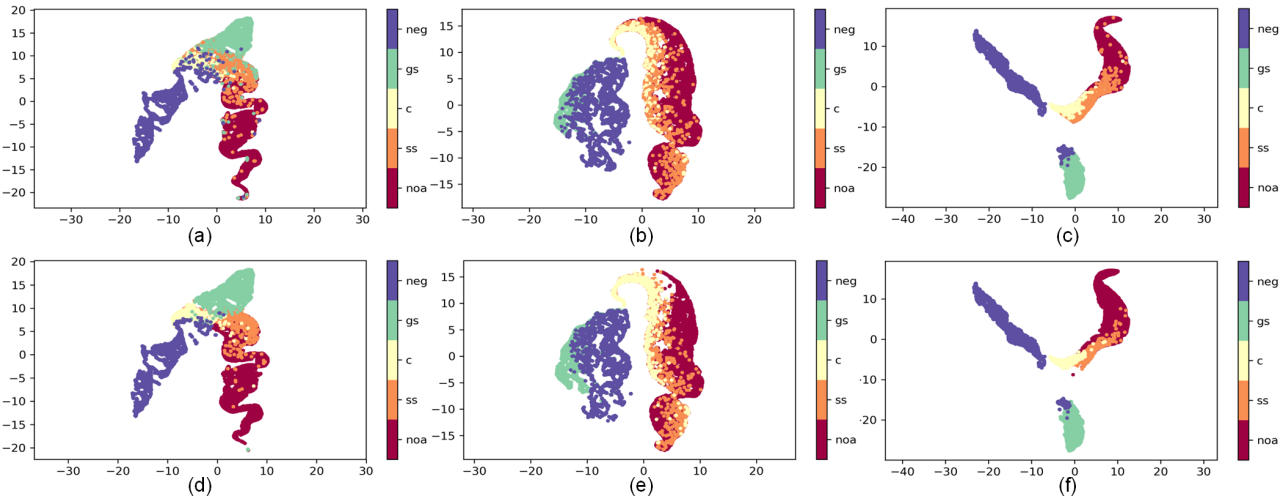


Figure 2.10: Data distribution without (a-c) and with (d-f) data reconstitution. (a) and (d) are the estimated probability vectors from the last dense layer, (b) and (e) refer to the raw probability vector from the GMP layer, (c) and (f) include both estimated and raw probability vectors.

Effectiveness of Data Reconstitution. The effectiveness of data reconstitution is illustrated in Fig. 2.10(a-c), which shows the data distribution before the reconstitution process. Initially, the distribution appears disordered, with categories such as SS, NoA, and C being densely intermingled. However, as observed in Fig. 2.10(e-f), the implementation of data reconstitution using the uncertainty factor significantly alters this scenario. It segregates the mislabelled or hard samples within the feature space, particularly for the CSN samples. On

the other hand, categories with lower variance, such as Neg and GS, are less affected. The distribution in Fig. 2.10 (f) demonstrates that the combination of probability vectors from the dense layer and GMP leads to a clear boundary.

2.5.3 Slide-wise evaluation

Table 2.8 demonstrates the model performance on 100 WSIs, and the visualization results are shown in Fig. 2.11.

Table 2.8: Performance of Glomeruli Detection

Model	AP_{NoA}	AP_{SS}	AP_{GS}	AP_{C}	mAP
Mask R-CNN [29]	0.705	0.103	0.707	0.240	0.439
FCOS [30]	0.789	0.280	0.751	0.336	0.539
MS R-CNN [44]	0.674	0.110	0.555	0.230	0.392
Cascade R-CNN [45]	0.785	0.165	0.780	0.194	0.481
Our Method (full 2-stage)	0.780	0.394	0.718	0.575	0.613

Experimental results indicate that the common detectors such as FCOS, MS R-CNN, Mask R-CNN, and Cascade R-CNN demonstrate a weaker performance compared to our proposed method, which achieved a mAP of 0.613. Specifically, these established detectors exhibited an 8-22% reduction in mAP relative to UAAN. To delve deeper into the results, we employed the Wilcoxon-Mann-Whitney test, a non-parametric statistical test, for analysis. The statistical outcomes revealed no significant differences ($p > 0.05$) between our method and the aforementioned detectors in identifying GS and NoA samples. In particular, the p-values were 0.0727 between FCOS and our method; 0.0619 between Cascade R-CNN and our method for GS; 0.0895 between FCOS and our method; and 0.0978 between Cascade R-CNN and our method for NoA. Although FCOS and Cascade R-CNN demonstrated comparable performance to our method in identifying GS and NoA samples, they exhibited limited capabilities in distinguishing C and SS samples ($p<0.05$). In contrast, our proposed method showed balanced performance across all categories, effectively detecting all types of glomerular lesions.

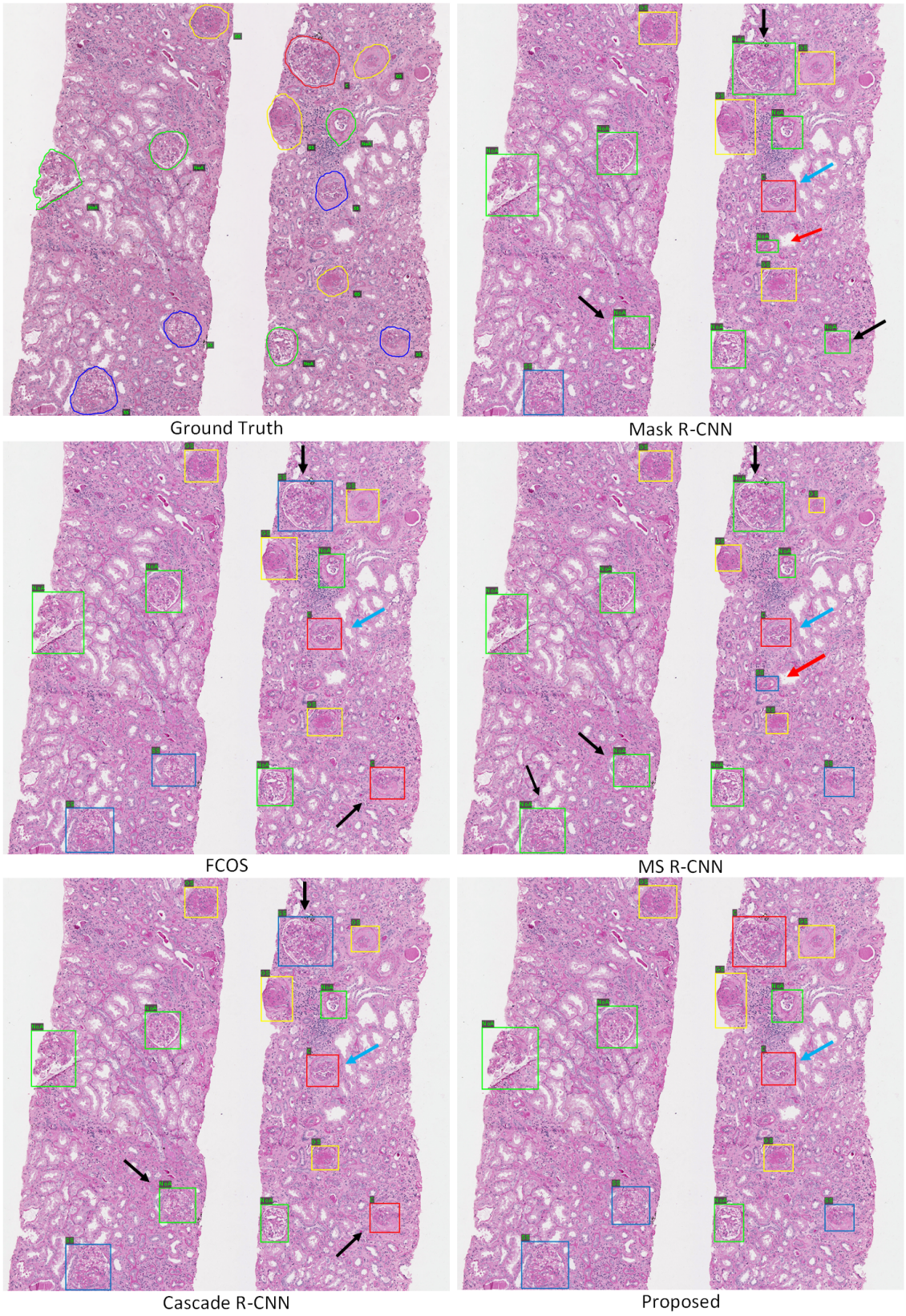


Figure 2.11: Visualization results of the slide evaluation, the red, green, yellow and blue boxes represent annotations/predictions of C, NoA, GS and SS, respectively. The mispredictions were highlighted by the black arrows.

2.6 Discussion

This section discusses the quantitative results of segmentation and fine-grained classification, as well as the uncertainty assessment and the limitations for detection methods when encountering FGVR tasks.

Segmentation Loss Functions. Various optimization functions were evaluated against the FISS loss by measuring their Dice scores in pilot studies and on Warwick-QU and renal pathology datasets. The results, illustrated in Fig.2.8, revealed that the proposed compound loss excels in sensitivity to varying target sizes. As detailed in Table 2.4, models utilizing FISS outperformed those using standard losses including Dice loss, focal loss, cross-entropy loss, Lovasz loss, Tversky loss, and conventional SSIM loss. Although combining Dice loss with other auxiliary losses effectively reduces false positives, it increases the rate of false negatives. In contrast, the DL+FISS loss strikes a balance, yielding more precise boundaries and smoother predictions (shown in Fig 2.8).

Fine-Grained Classification in Pathological Images. The ablation study presented in Table 2.5 emphasizes the importance of leveraging common features across different categories, rather than only concentrating on discriminative regions. It demonstrates that feature apportionment assists in distinguishing NoA from SS and C, while the accurate classification of SS and C necessitates a more refined data set. By combining uncertainty-aided data reconstitution with a feature apportionment strategy, there's a notable enhancement in both micro and macro accuracy. According to Table 2.4, the proposed method outperforms the comparisons significantly, achieving 95.07% micro-accuracy (with 3–11% gain) and 91.28% macro-accuracy (a relative improvement of 12–26%). This substantial disparity underscores the limited classification capabilities of current competitive classification models when handling data imbalance issues. The proposed UAAN distinguishes itself from comparison approaches in several ways: 1) It does not require numerous branches or generative models. 2) It considers both discriminative and common features among subcategories. 3) Its structure might be further enhanced by introducing complex operations such as attention modules [60]. 4) It could reduce the need for part annotations or additional subnetworks as our method takes raw images as inputs.

Interpretability of UAAN. Heat maps for comparison models were generated, focusing on different types of glomerular lesions. Among these categories, SS proved to be the most challenging class, with only the UAAN and DenseNet-121 models correctly focusing on the relevant areas. Figure 2.9 highlights that even some models with high accuracies, like SE-ResNeXt-101 and EfficientNet-B6, do not concentrate their attention regions entirely on the pathological changes. In contrast, UAAN offers a higher degree of interpretability in its predictions. This distinct focus of UAAN on relevant pathological areas, as opposed to more scattered attention in other models, underscores its effectiveness in accurately identifying and interpreting complex medical images.

Uncertainty in Deep Convolutional Neural Network. This study introduced a method for assessing uncertainty by estimating the correlation between two probability densities. While this approach does not offer rigorous uncertainty calculation as Bayesian inference, it still provides a useful estimate of the confidence coefficient for the predictions. Additionally, this uncertainty assessment was employed for selecting unreliable annotation and challenging samples, without necessitating multiple forward propagations. Ablation experiments, as shown in Table 2.7, reveal that increasing the uncertainty threshold results in the selection of fewer samples, but with a higher proportion of mislabeled ones. An in-depth review by three senior pathologists found that among the mislabeled images, 36% and 55% were correctly reclassified as C and SS, respectively, when the threshold was set at 0.5. This indicates that the output of the GMP correlates more closely with the real distribution, effectively simulating the raw probability density function. When there is a lack of correlation between P_e and P_r, it suggests a misrepresentation of features in the last dense layer. In such cases, the softmax layer's output represents an erroneous probability distribution due to inaccurate feature representation. This misrepresentation may explain why CNN+SVM models have shown superiority over pure CNN models in previous studies [61]. The implementation of this uncertainty assessment has led to significant improvements (1–4%) in micro-accuracy for current state-of-the-art models.

Comparisons with junior pathologists. The performance of UAAN was benchmarked against that of junior pathologists (1-2 years' experience) on a pilot dataset, consisting of 200 images for each class. The labels for this dataset were assigned by two senior pathologists,

achieving a high inter-observer agreement with a Cohen's Kappa score of 0.791 ($p < 0.0001$). The comparison revealed that junior pathologists obtained an accuracy of 89.4%, which is slightly lower than the 90.7% accuracy achieved by UAAN. This difference in accuracy demonstrates the capability of UAAN to perform at a level comparable to human experts in the field. Furthermore, in terms of efficiency, it took only 7.57ms for UAAN based on the Nvidia Tesla V100 16GB GPU with a batch size of 32. In contrast, junior pathologists require approximately 2 seconds per glomerulus (average speed), which makes the UAAN nearly 260 times faster than the junior pathologist.

Why two stages? The effectiveness of current state-of-the-art detection modules in locating various types of glomeruli, though competitive, falls short in classifying fine-grained lesions. These modules function optimally at a 4X resolution (2.468 μm/pixel), taking inputs with a size of 2048x2048 pixels. At this resolution, however, the model faces challenges in extracting the fine-grained features to differentiate between the subcategories of lesions. Another significant issue is the complexity and size of the classification heads added to these models, which leads to the exhaustion of GPU memory.

Limitations. This study acknowledges certain limitations, despite its superior performance compared to existing methods. One of the primary challenges is the complexity involved in the feature extraction, and the effectiveness of the second stage is heavily dependent on the outcomes of the segmentation stage. A specific issue arises when two adjacent glomeruli are not segmented individually; in such cases, the classification module struggles to perform accurately. Additionally, the stability of data reconstitution is influenced by various factors, such as the quality of annotations, the subjectivity of experts, and the characteristics of the dataset itself. These factors can introduce variability and affect the reliability of the results. Another limitation mentioned is the lack of multicentre validation.

2.7 Conclusion

In this study, a comprehensive approach for recognizing glomeruli lesions is introduced, incorporating focal instance structural similarity loss and an uncertainty-guided apportionment network. This research marks the first attempt to address the challenge of fine-grained visual recognition in pathological image analysis, with a particular focus on IgA Nephropathy in whole slide images. The study's innovative use of focal instance structural similarity loss and uncertainty-aided apportionment networks has led to significant advancements, demonstrating an 8–22% improvement in mAP over existing methods. The ability of the proposed method represents a substantial step forward in the field of digital pathology in IgAN.

Chapter 3

Airway modelling in High-resolution CT

3.1 Fuzzy attention neural network for airway segmentation

3.1.1 Introduction

The severity of traction bronchiectasis is a key indicator of outcomes in pulmonary diseases [9, 10, 62]. However, the reliability of manual grading on HRCT is limited due to considerable interobserver variability and weak reproducibility. Additionally, such visual assessments often lack sensitivity to seize minor changes over short follow-up terms [63]. These limitations underscore the need for the development of objective, computer-based methods for the quantification of the disease [64–66]. The capabilities of traditional approaches (e.g. region growing [67–69], fuzzy connectedness [70], intensity thresholding [71]) for airway segmentation were summarized in a benchmark paper [72]. However, these solutions have shown limitations in robustness and tend to produce coarse predictions due to the complex nature of bronchus patterns. Beyond traditional methods, deep neural networks, such as VNet [73] and nnUNet [74], have gained popularity in medical segmentation owing to their modelling capabilities. Despite significant

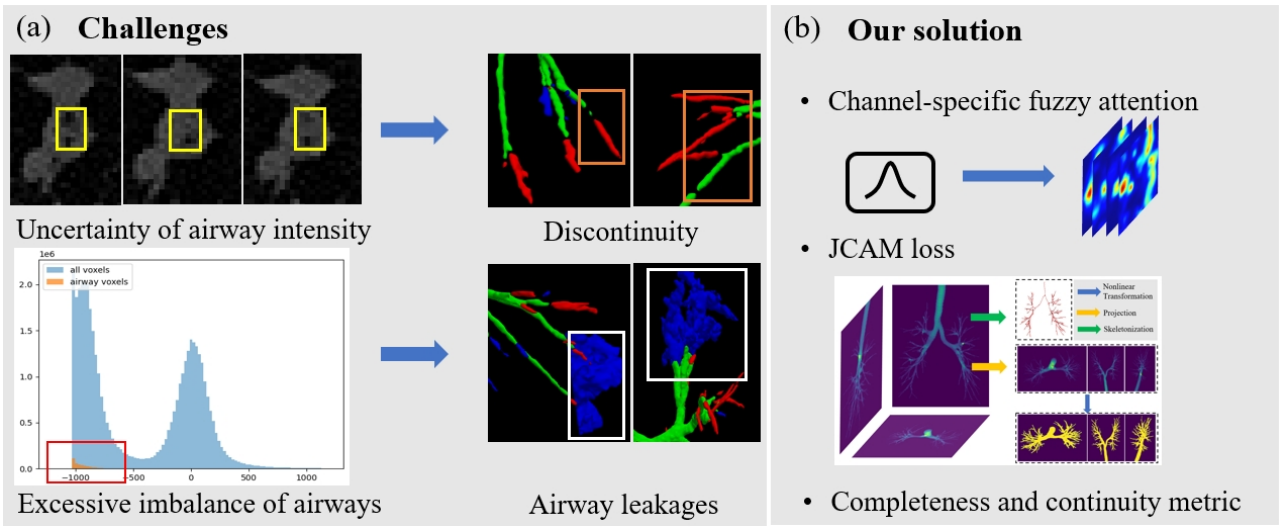


Figure 3.1: Challenges in airway segmentation and our solutions. (a) shows the varied intensity (details in yellow boxes) of airway trees and data imbalance issues. Green, red, and blue colours refer to TP, FN, and FP predictions. (b) describes our solutions regarding the aforementioned challenges, including fuzzy attention layer, JCAM loss for optimization, and a new metric to evaluate airway prediction continuity and completeness.

achievements in airway segmentation [75–79], existing methods still face challenges (Fig. 3.1) in achieving high continuity and completeness in their predictions of airway structures, as evidenced in our studies.

In airway segmentation, there are five key challenges: (1) Imbalanced distribution of foreground and background samples and the imbalanced voxel volume quantity between the trachea and terminal branches. This complicates the training of data-driven models and hampers the detection of terminal branches. Additionally, while most models on 3D medical data were trained on cropped patches, the methodology for patch extraction, often defaulting to sliding windows, lacks efficiency. (2) The similarity in voxel values between airway trees and normal lung regions, which refers to the inter-class homogeneity, necessitates the model learning beyond just voxel intensities to include more spatial and structural information. (3) Intra-class heterogeneity within airway trees, arising from incorrect labelling and imaging issues, requires robustness in the model to handle annotation errors and variations in voxel values. (4) The reliability of evaluation metrics, like intersection over union (IoU) or Dice coefficient, for model selection is limited [80]. The disparity in voxel volume between the main trachea and smaller branches renders these metrics less effective. (5) insufficient robustness and generalization on multicentre data, with many models requiring additional fine-tuning for different cohorts [80].

The attention mechanism has proven effective in segmenting complex structures by assigning varying degrees of 'importance' to feature representations [81]. However, challenges arise due to intra-class heterogeneity, often resulting from subjective annotation, and inter-class homogeneity introduced by similarities of HU intensities. These issues introduce a significant amount of 'uncertainty' into the features being analyzed. Implementing fuzzy logic has been shown to effectively reduce this uncertainty in feature representations, leading to notable improvements [82–84]. Therefore, it's hypothesized that combining the deep attention mechanism with fuzzy theory could effectively address the heterogeneity and homogeneity in data. Additionally, to improve the continuity of segmentation predictions, integrating the centerlines of airway branches with overlapped metrics in the loss calculation has been proposed. However, these overlapped centerline metrics tend to focus more on completeness, potentially overlooking continuity assessment. Thus, it is suggested that the loss function should be modified to better evaluate continuity and completeness by calculating their accumulation maps.

This section introduces an effective method for precise airway segmentation, comprising four key components: 1) a novel patch sampling strategy, 2) a unique channel-specific attention layer incorporated fuzzy theory, 3) a combined completeness and continuity optimization function, and 4) a new evaluation metric, CCF score. Extensive experimental studies were performed using both an open-access dataset consisting of 90 cases and our own in-house test-only datasets with 50 cases. These datasets (collected from 140 patients) include normal scans as well as cases with mild anomalies, COVID-19, cancer, and fibrosis. The effectiveness of the segmentation was evaluated according to the completeness and continuity of the airway trees. Additionally, the method involves estimating the radius of each branch in the airway trees, which is helpful to better understand lung disease.

3.1.2 Related Works

Airway Segmentation in HRCT.

The EXACT'09 results showed that no existing algorithm could achieve more than an average of 74% of the detected length of airway trees in the reference standard. To enhance performance,

Charbonnier et al. [75] developed a 2D CNN aimed at detecting airway leakages for post-processing. Yun et al. [76] introduced a 2.5D CNN, which utilizes three adjacent slices from each orthogonal direction (axial, sagittal, coronal) to augment spatial information. Beyond these, 3D CNNs have gained popularity for their superior consistency and integrity in predictions, as evidenced by several studies [77–79].

Several researchers treated airway segmentation as a class imbalance problem and have proposed methods to minimize interference from background regions. For example, Juarez et al. [77] applied a weighted cross-entropy loss to create a balance between foreground and background samples in their analysis. Garcia-Uceda et al. [85] and Qin et al. [86] both employed a technique where they cropped the full-size volume images to a bounding box that is centred around the pre-segmented lung region, effectively removing irrelevant background areas. Zheng et al. [87] adopted a two-stage training approach. Initially, they trained the network on sampled patches and identified hard-to-segment regions based on false-negative predictions. Subsequently, they fine-tuned the network in the second stage to enhance its segmentation performance. These varied approaches illustrate the ongoing efforts to tackle the complex challenge of accurate airway tree segmentation.

Others advanced models by targeting discontinuity issues and leakages. For instance, Qin et al., introduced a 3D UNet with an awareness of voxel connectivity specifically for airway segmentation [78]. Reynisson et al., [88] and Meng et al. [89] worked on improving the continuity of predictions by focusing on the airway centerlines. Zheng et al. delved into the causes of inaccurate segmentation, particularly highlighting the impact of gradient erosion and dilation on adjacent voxels [87]. Nadeem et al. [90], proposed a novel 3D UNet with a freeze-and-grow propagation approach to alleviate leakages and increase the accuracy of branch detection. Additionally, Wang et al., [91] proposed a radial distance loss technique to enhance the completeness of tubular structure predictions.

Despite these innovative efforts, challenges remain, particularly with discontinuous predictions in smaller airway branches. Furthermore, previous methods relied on multi-stage training or intricate post-processing procedures. This not only intensifies the computational cost but also

increases the time required for processing, posing practical challenges in real-world applications.

Fuzzy logic in Deep Semantic Segmentation

Beyond the established fuzzy clustering methods for segmentation, there have been several earlier attempts to blend deep neural networks with fuzzy logic, typically through a sequential learning protocol [92, 93]. An example of this is the use of neural networks to extract low-dimensional feature representations, which are then processed using fuzzy logic for classification [94]. However, a significant limitation of these multi-stage approaches is their inability to be trained in an end-to-end manner. Additionally, they do not integrate fuzzy theory directly into the training process of the deep neural networks, representing a gap between the potential of fuzzy logic and its actual application in deep learning models for segmentation tasks.

To tackle the integration of fuzzy logic with deep neural networks, researchers have developed end-to-end hierarchical schemes [82]. This has led to innovative approaches combining fuzzy logic with CNNs for segmentation tasks [83, 95, 96]. For instance, Huang et al. [96] developed a fuzzy FCN for segmenting breast ultrasound images, involving the transformation of data into the fuzzy domain for effective processing. Price et al. [95] introduced a flexible and efficient fuzzy layer that harnesses the robust aggregation capabilities of fuzzy integrals. Ding et al. [97] demonstrated the application of fuzzy-based learning in simultaneously segmenting brain tumours from T1 and T2 MRI scans. Guan et al. [83] created a fuzzy neural network aimed at lip segmentation, where the inclusion of a fuzzy logic module significantly improved the robustness of the segmentation results. Despite these advancements, a common limitation in existing approaches is their reliance on integrating a fuzzy integral within a residual block, primarily for information fusion. This approach falls short of addressing the specific challenges associated with airway segmentation. Additionally, there's a noticeable lack of diversity in the use of aggregation operators within these models, with most studies focusing on the fuzzy 'AND' operator and not exploring other potentially beneficial operators.

Evaluation Metrics of Airway Predictions

Given the prediction X and the corresponding mask Y, the common evaluation metrics for airway segmentation are comprised of overlapped measurements, detected length rate, branch rate and airway leakage rate, which are given as follows:

1. **Overlapped metrics** assess the overlap rate between the ground truth masks and predictions, such as Jaccard index score and dice score, denoted as:

$$\text{Jaccard}(X, Y) = \frac{XY}{X + Y - XY}, \tag{3.1}$$

$$\text{Dice}(X, Y) = \frac{2XY}{X + Y}, \tag{3.2}$$

   where XY are the intersections of X and Y.

2. **Detected branch ratio (DBR)** evaluates the percentage of detected branches concerning the expert annotations

$$\text{DBR} = \frac{N_x}{N_y} \times 100\%, \tag{3.3}$$

   where N_X is the amount of correctly detected branches, N_Y is the amount of branches in the corresponding annotation. Specifically, a branch is considered to be correctly detected when its IoU score is greater than 0.8 with the corresponding branch in the ground truth.

3. **Detected length ratio (DLR)** assesses the ratio of detected branch length for that of the expert annotations

$$\text{DLR} = \frac{L_x}{L_y} \times 100\%, \tag{3.4}$$

   where L_X denotes the total length of correctly recognized airway branches, L_Y refers to the whole length of branches in the ground truth.

4. **Airway leakage ratio (ALR)** denotes the ratio of total false positive predictions to the ground truth annotations

$$\mathrm{ALR} = \frac{V_x}{V_y} \times 100\%, \tag{3.5}$$

where V_X denotes the number of FP predictions, and V_Y refers to the volume of ground truth annotation.

It is of note that none of these metrics can be independently adopted for airway evaluation.

3.1.3 Methodology

Overview

Given an input raw 3D volume $V \in \mathbb{R}^{W \times H \times Z}$, we first introduce a smart patch sampling strategy to extract patches $V_P \in \mathbb{R}^{w \times h \times c}$ for model development. Subsequently, a fuzzy attention neural network (FANN) is devised for segmenting airway trees, under the supervision of a comprehensive optimization function. This function encompasses dice loss, accumulation mapping loss, and airway continuity loss. During the training phase, the CCF score was employed to evaluate the model and to save the optimal values of the trainable network parameters. The proposed Fuzzy Attention Neural Network is an extension of the 3D U-Net, enhanced by incorporating deep supervision, fuzzy attention layers, and JCAM (Jaccard continuity and accumulation mapping) loss. The size of patches V_P is determined based on the same ratio as the median shape of the ground truth annotations in the training set. FANN comprises several basic 3D layers such as 3D (transpose) convolution layers, instance normalization (IN), and fuzzy attention layers, as depicted in Fig. 3.2. Notably, the network was designed with multiple outputs for deep supervision, where the final prediction was obtained from the main output. The predictions from low-level outputs were used to calculate losses as the main output but with different weights.

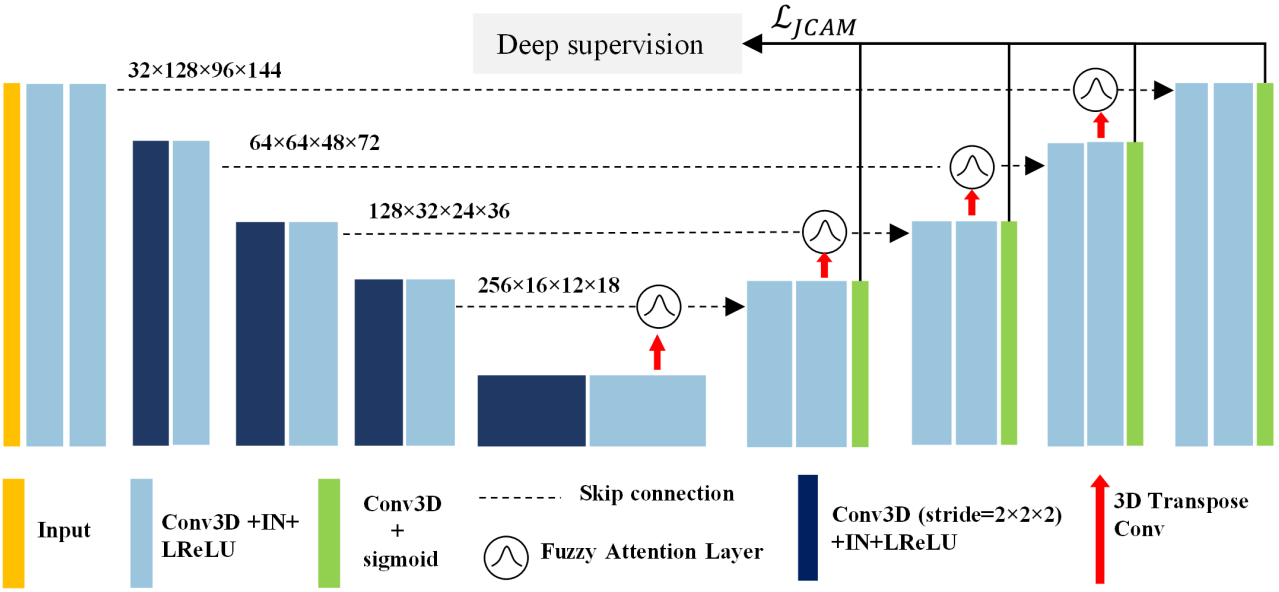


Figure 3.2: Architecture of the proposed fuzzy attention Neural Network. The fuzzy attention blocks are arranged with skip connections between encoders and decoders. Deep supervision was implemented for decoder blocks.

Smart Patch Sampling

Data-driven methods heavily rely on the data quantity, therefore, a balanced class distribution is essential for training a neural network. It has been mentioned that the significant imbalance between foreground and background samples can lead to a gradient erosion issue [87]. Due to the extensive volume size of HRCT, segmentation modules for HRCT often employ patch-based training strategies. Nevertheless, most studies applied sliding windows for patch extraction, which do not fully address the issue of class imbalance. Some researchers have attempted to mitigate class imbalance by capturing hard samples [87, 98], while these methods typically require an additional training phase to gather these hard samples. In our study, we propose a smart patch sampling (SPS) strategy for data sampling without fine-tuning or additional training. Considering a dataset containing 3D images and their corresponding annotations, the SPS workflow can be summarized as follows:

The smart patch sampling strategy starts with calculating the average size $S(z \times y \times x)$ of the MBR of the ground truth annotations, which informs the scale for the patch size selection. Subsequently, the centerline of the manual annotations is extracted using a skeletonization process [99]. To obtain the image patches, mask patches, and centerline patches, SPS employs

overlapping sliding windows. Finally, patches are selected based on two criteria: a centerline ratio, which is the ratio of centerline voxels to the total number of centerline voxels within a patch, that exceeds 15%; and a voxel volume ratio, which is the ratio of ground truth voxels within a patch to the total ground truth voxels, that exceeds 10%. Patches meeting these criteria are preserved for further development.

Fuzzy Attention Layer

One of the key challenges in achieving an effective model for airway segmentation is the uncertainty inherent in annotations and HU intensities of the airway regions. Efforts have been made to refine the focus of neural networks on RoIs. For example, the Attention U-Net [81] introduces an attention gate mechanism aimed at enhancing accuracy by lightening features in irrelevant regions. Nonetheless, we hypothesize that applying sigmoid activations in the attention layer may not represent the optimal approach.

Besides the "gradient vanishing", the sigmoid activation function is criticised by its sharp transition, where only a narrow input interval results in the output spanning the entire range from 0 to 1. Consequently, it is challenging to define a robust "boundary" with the sigmoid function that can effectively distinguish relevant features. Another limitation is its monotonicity; akin to the intensity distribution of airway structures, which has both negative and positive variations, the feature representation for airways should also exhibit bidirectional variations. However, with sigmoid activation, 1×1 convolution layers are often forced to shift the values of all features of interest to the positive side to obtain a response. Specifically, there must be a "threshold" in the feature representation reconstitution, usually achieved by 1×1 convolution layers within the attention gate, to decide the importance of a region. Moreover, the current attention maps are non-channel-specific, applying the same "attention" coefficient to all features along the channel dimension. Given a feature vector $F \in \mathbb{R}^{C \times W \times H \times D}$, the conventional attention map is expressed as $\alpha \in \mathbb{R}^{W \times H \times D}$, implying that all features along the channel C are equally "important". This might be problematic because features in different channels are extracted by distinct convolutional kernels, which suggests that a more effective approach would

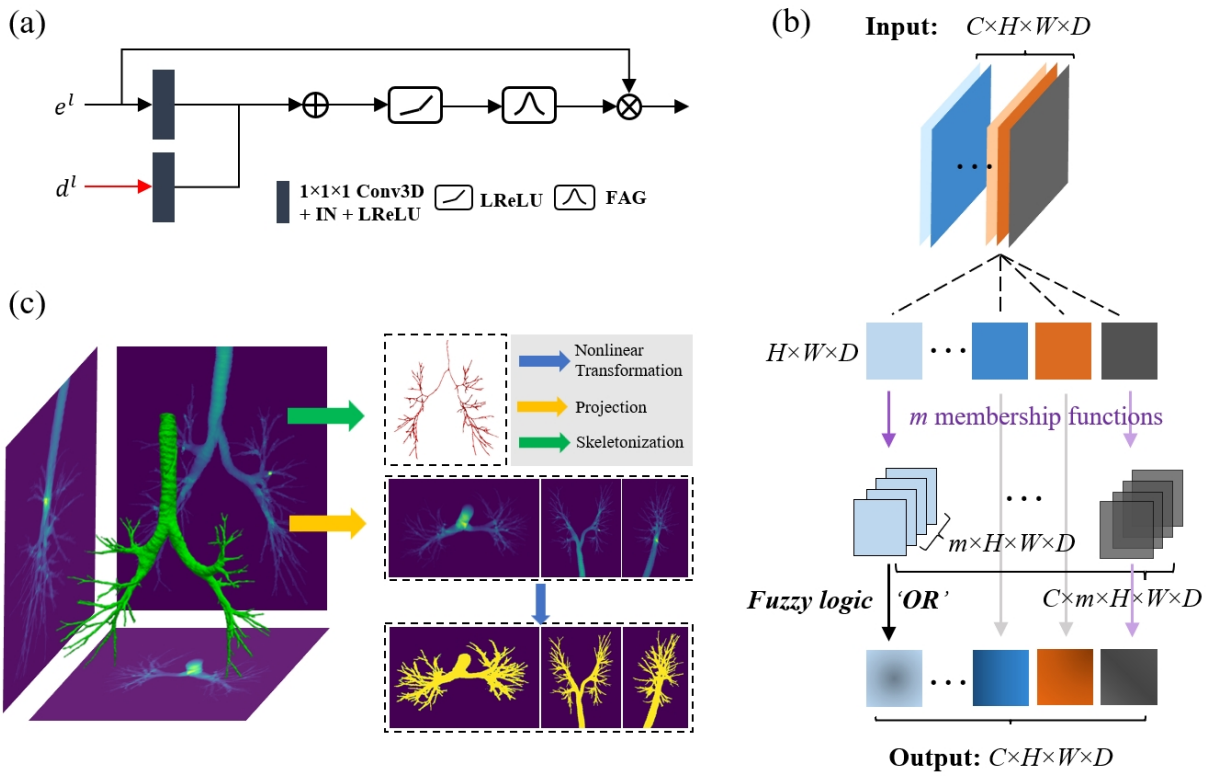


Figure 3.3: Architecture of (a) fuzzy attention layer, (b) fuzzy attention gate with Gaussian membership functions, and (3) accumulation maps of different panels and skeletonization of airway trees.

be for the attention map to be channel-specific.

Studies have indicated that both fuzzy logic and neural networks are potent for data representation [82]. Typically, a neural network is utilized to diminish noise and distil useful feature representations from raw data, while fuzzy logic aims to derive fuzzy representations that mitigate the uncertainty inherent in the data. In contrast to the sigmoid activation function, we hypothesize that the Gaussian function is more appropriate for formulating the attention gate, owing to its symmetric shape and modifiable parameters (the mean and variance of the Gaussian function can be adapted during learning). Additionally, its smoothness and non-zero nature across its entire domain well address the issues related to gradient vanishing and potential collapse. Consequently, we integrated fuzzy logic into the attention mechanism, employing trainable Gaussian membership functions to enhance the segmentation network's ability to concentrate on RoIs, thereby diminishing the variations and uncertainties through representation aspects.

The fuzzy attention layer we propose is integrated within the skip connections, utilizing fea-

tures from both the l-th encoder e^l and decoder d^l, as depicted in Fig. 3.3(a). These input feature vectors are first proceeded by a $1 \times 1 \times 1$ 3D convolution layer, followed by IN and a LeakyReLU for feature reconstitution. Subsequently, a voxel-wise addition operation is used to merge the information, which is then followed by another LeakyReLU activation. The resulting feature representations are then supplied to the Fuzzy Attention Gate to produce a voxel-wise attention map, as illustrated in Fig. 3.3(b). Let X, with dimensions $C \times H \times W \times D$, represent the input to the fuzzy attention gate, disregarding batch size for the moment. Owing to the smoothness and streamlined representation of Gaussian membership functions, we utilize learnable Gaussian MFs to define the deep fuzzy sets. Each feature map, characterized by height H, width W, and depth D, is fed into m Gaussian membership functions (with center $\mu_{i,j}$ and spread $\sigma_{i,j}$)

$$f_{i,j}(X, \mu, \sigma) = e^{\frac{-(X_j - \mu_{i,j})^2}{2\sigma_{i,j}^2}} \in \mathbb{R}^{C \times m \times H \times W \times D}, \tag{3.6}$$

where $i = 1, \ldots, m, j = 1, \ldots, C$. In contrast to prior studies that employed the 'AND' operator to derive fuzzy feature representations, our approach aims to utilize the membership function to ascertain the 'importance' of target features. We hold that the 'OR' aggregation operator is more effective in preserving information while suppressing irrelevant ones. Given two fuzzy sets $\tilde{A}$ and $\tilde{B}$, the 'OR' operator is mathematically characterized as follows:

$$f_{\tilde{A} \cup \tilde{B}}(y) = f_{\tilde{A}}(y) \vee f_{\tilde{B}}(y), \ \forall y \in U, \tag{3.7}$$

where y is the element of the universe, and U is the universe of information. For derivation available, we modified the function to

$$f_{\tilde{A} \cup \tilde{B}}(y) = \max(f_{\tilde{A}}(y), f_{\tilde{B}}(y)), \tag{3.8}$$

As a result, the fuzzy degree $f_j(X, \mu, \sigma) \in \Theta^{H \times W \times D}, \Theta \in [0, 1]$ of $j - th$ channel is calculated by

$$f_j(X, \mu, \sigma) = \bigvee_{i=1}^{m} e^{\frac{-(X_j - \mu_{i,j})^2}{2\sigma_{i,j}^2}} = \max(e^{\frac{-(X_j - \mu_{i,j})^2}{2\sigma_{i,j}^2}}) \tag{3.9}$$

where $\bigvee$ refers to the union process.

Ultimately, the output tensor from the FAG remains the same dimensions as the input tensors, yielding a voxel-wise attention map. The pseudocode for the fuzzy attention layer is demonstrated in Algorithm 1.

Algorithm 1 Pseudo code of fuzzy attention layer

Input: feature representations $e^l, d^l \in \mathbb{R}^{C\times W\times H\times D}$,from l-th encoder and decoder, , weights $w_e^{(l)}, w_d^{(l)}$ and bias $b_e^{(l)}, b_d^{(l)}$ connecting to l-th encoder and decoder

Output: fused feature representation $y \in \mathbb{R}^{C\times W\times H\times D}$

1: randomly initialise parameters $\mu \in \mathbb{R}^{m\times C}$ and $\sigma \in \mathbb{R}^{m\times C}$ of membership functions $f(X, \mu, \sigma)$.
2: **compute** the input $X \in \mathbb{R}^{C\times H\times W\times D}$ of FAG, $X = \mathrm{LReLU}(\mathrm{IN}(w_e^{(l)} e^l + b_e^{(l)})) + \mathrm{LReLU}(\mathrm{IN}(w_d^{(l)} d^l + b_d^{(l)}))$
3: **for** i in C **do**
4: **Specify** the fuzzy sets $f_{i,j}(X, \mu, \sigma)$ through m Gaussian membership functions as Eq. (6)
5: **compute** fuzzy degrees $f_j(X, \mu, \sigma)$ of j-th channel by adopting aggregation operator "OR" to fuzzy sets $f_{i,j}$ through Eq. (9).
6: **compute** output of j-th channel $y_j = e^{(l)} f_j(X, \mu, \sigma)$
7: **end for**

Jaccard Continuity and Accumulation Mapping Loss

The research community has confidently embraced strategies like cascade modules and cropping operations to enhance the segmentation of small organs or structures. These solutions effectively mitigate the challenge of inter-class imbalance. However, there remains a challenge of intra-class imbalance, attributable to the varying volumes of distinct branches within the airway tree, such as the trachea and secondary or tertiary bronchi. Predominantly, the trachea and secondary bronchi compose the bulk of the airway volume, which often results in discontinuous predictions for smaller bronchial structures.

To address the issue, we introduce a Jaccard continuity and accumulation mapping loss L_{JCAM}, which encourages the network to emphasize the continuity of airway predictions. The L_{JCAM} evaluates the discrepancies in different projections (through coronal, sagittal, and axial planes) and centrelines between the prediction and the ground truth. Given the prediction array X and the corresponding ground truth annotation Y, the centrelines Y_{CL} of the ground truth masks are delineated through skeletonization [99]. The continuity is effectively assessed by calculating

the ratio of correctly predicted centrelines to the ground truth centrelines Y_{CL}, irrespective of the airway branch size.

$$L_C = 1 - C = 1 - \frac{\sum X \cdot Y_{CL}}{\sum Y_{CL}} \tag{3.10}$$

Alongside continuity, we also introduce two variants of accumulation maps: the Linear Accumulation Maps (LAMs) and Nonlinear Accumulation Maps (nLAMs), as illustrated in Fig.3.3(c). LAMs are generated by aggregating the volume array along three orthogonal axes, while nLAMs result from applying a nonlinear transformation to the LAMs. Incorporating LAM into the loss function enhances the accuracy of the 3D prediction assessment since each mispredicted voxel contributes to discrepancies in the various projection maps. Moreover, to diminish intra-class imbalance, a nonlinear transformation using the tanh activation function is applied to LAMs to obtain nLAMs. Although nLAM, with its range confined between [0, 1], eliminates the spatial information inherent in LAM, it provides a more refined measure of continuity and veracity between the predictions and the ground truth masks.

Assume $A \oplus \tau$ as the operation that sums the array $A \in \mathbb{R}^{N \times C \times W \times H \times D}$ along τth channel. The LAM and nLAM are calculated by

$$L_{LAM} = \sum_{\tau \in [W,H,D]} l_1(X \oplus \tau, Y \oplus \tau), \tag{3.11}$$

$$L_{nLAM} = \sum_{\tau \in [W,H,D]} L_{\mathcal{J}}(\tanh(X \oplus \tau), \tanh(Y \oplus \tau)), \tag{3.12}$$

where $L_{\mathcal{J}}$ is the Jaccard loss which is defined as

$$L_{\mathcal{J}} = 1 - \mathcal{J} = 1 - \frac{XY + \varepsilon}{X + Y - XY + \varepsilon} \tag{3.13}$$

and ε is the smoothing factor ε>0.

Overall, the L_{JCAM} can be summarised as

$$L_{JCAM} = \alpha L_{\mathcal{J}}(X, Y) + \beta L_C(X, Y_{CL}) + \varphi L_{CE}(X, Y) + \gamma L_{LAM}(X, Y) + \delta L_{nLAM}(X, Y) \tag{3.14}$$

where $\alpha, \beta, \gamma, \varphi, \delta$ are the weights of Jaccard loss $L_{\mathcal{J}}$, continuity loss L_C, cross-entropy loss L_{CE}, linear and nonlinear accumulation mapping loss L_{LAM}, L_{nLAM}, respectively.

Continuity and Completeness F-score

In addition to exploring novel layers, the selection of modules is pivotal in machine learning. Specifically for airway segmentation, relying on the highest regional overlapping metrics to determine the best model is problematic due to (1) the intra-class imbalance of airway structures and (2) the limitations of manual annotations. Metrics such as the Dice coefficient or IoU score predominantly reflect the overall overlap of prediction voxels but fall short in evaluating the continuity and completeness of the structures. Consequently, there is a need to formulate a metric capable of accurately assessing the continuity and completeness of tree-like structures, such as airways, blood vessels, and neural pathways.

Here we propose a continuity and completeness F-score (CCF score)

$$\mathrm{CCF}_s = (1+\omega^2) \times \frac{J \times C}{\omega^2 \times J + C} \tag{3.15}$$

where $\omega \in [0, 1]$ is the parameter that weights the preference between completeness and continuity, J and C are the Jaccard index and Continuity in Eq.(2.25) and Eq. (2.28), respectively. Based on this novel metric, the model can evaluate both continuity and completeness simultaneously.

3.1.4 Experiments and Results

Datasets and Training Strategies

In our investigation, the training of airway segmentation and comparison models was conducted utilizing 90 clinical thoracic CT scans sourced from the publicly available BAS dataset. For testing purposes, we exclusively used in-house datasets compiled from patients diagnosed with

fibrotic lung disease and COVID-19. Comprehensive details regarding all utilized datasets are tabulated in Table 3.1.

Table 3.1: Details of datasets in this study

Dataset	Composition	Volume size	Slice thickness
BAS	90 cases*	$s\times512\times512$, $s \in [157, 764]$	0.45mm-1.80mm
COVID-19	25 cases (test only)	$s\times512\times512$, $s \in [512, 1145]$	0.40mm-0.65mm
Fibrosis	25 cases (test only)	$s\times512\times512$, $s \in [516, 945]$	0.40mm-0.70mm

* 90 cases are split into 54,18 and 18 for training, validation and testing, respectively. BAS was collected from healthy volunteers or patients with pulmonary disease.

Binary airway segmentation (BAS) dataset: The BAS dataset comprises a total of 90 cases, with 20 cases sourced from EXACT'09 and 70 cases from the LIDC database. (1) The EXACT'09 dataset [72] consists of 20 training cases (with annotations) and 20 test cases (annotations not publicly available), with patient conditions ranging from normal to severe pulmonary diseases. (2) The LIDC dataset: The Lung Image Database Consortium image collection (LIDC-IDRI) contains 1018 cases of diagnostic and lung cancer screening thoracic CT scans [100]. Out of these, 70 cases with a slice thickness of $\leq$ 0.625mm were randomly chosen and annotated by experts [86, 87].

Our in-house dataset: (1) **COVID-19 dataset:** This subset comprises 25 HRCT scans from patients diagnosed with COVID-19, obtained from Wuhan Renmin Hospital. (2) **Fibrosis dataset:** This subset consists of 25 cases from patients with fibrotic lung disease, sourced from the OSIC dataset[1]. Annotations for these cases were provided by medical experts at the Royal Brompton Hospital.

Training Strategies: All parameters were initialized using He-normal initialization. Data augmentation during training includes randomized rotation (with degrees ranging from -10 to 10) and randomized flips (4 directions). Each model was trained over 200 epochs with an initial learning rate of 1×10^{-3} and a learning rate decay of 0.5 applied at the 20th, 50th, 80th, 110th, and 150th epochs. To ensure equal contributions from $L_{\mathcal{J}}, L_C, L_{CE}$, the hyper-parameters α, β, γ in L_{JCAM} were set to 1. Meanwhile, φ, δ were set to 0.3 to constrain the accumulation mapping loss, as it is computed across three dimensions. The ω in the CCF score was fixed at

[1]https://www.osicild.org/

0.9 to minimize excessive leakages.

Experimental Settings and Evaluation Metrics

The experiments were performed in the following environment:
(1) Ubuntu 20.04.2 LTS with i9-10980XE CPU and 128GB memory, equipped with NVIDIA RTX 3090 TI.
(2) CUDA 11.1.74 and cuDNN 8.0.5
(3) Python 3.8.5 with numpy=1.19.5, SimpleITK=2.1.1, scikit-image=0.18.2.

We randomly split the 90 cases from the BAS dataset, allocating 72 for training and validation, and reserving 18 for testing. All modules, including those used in ablation studies and comparative models, were trained from the ground up using an identical set of patches provided by the Smart Patch Sampling (SPS) method, employing the same optimizer and consistent training strategies.

To ensure a fair comparison, all models underwent identical post-processing, which involved extracting the largest connected component to eliminate noise and adhering to the same inference criteria using a non-overlapping sliding window technique. The significance of the outcomes between our proposed method and the comparative methods was determined using the Wilcoxon signed-rank test, with a p-value <0.05 denoting statistically significant improvements (or deteriorations).

Ablation studies. To investigate the efficacy of the proposed sub-modules, we conducted ablation studies for airway segmentation. These included: the basic 3D U-Net (BL); 3D U-Net augmented with a fuzzy attention layer (BL+Fuzzy); 3D U-Net with both a fuzzy attention layer and JCAM loss (BL+Fuzzy+JCAM); and 3D U-Net with a fuzzy attention layer, JCAM loss, and CCF score (BL+Fuzzy+JCAM+CCF_s), among other combinations (BL+JCAM, BL+Fuzzy+CCF_s). It is noteworthy that the baseline 3D UNet, serving as the structure for FAAN minus the fuzzy attention layer (depicted in Fig. 3.2), was selected to maintain a fair comparison.

Comparisons. The effectiveness of our proposed method was benchmarked against conventional methods, previous solutions for airway extraction, and high-performance models adept at segmenting tree-like structures. The standard approaches included nnUNet [74], VNet [73], and Attention UNet [81]. Previous solutions for airway segmentation encompassed methods by Juarez et al. [77], Wang et al. [91], WingsNet [87], NaviAirway [101], and BronchiNet [102]. Moreover, we examined high-performance modules for human tree-like structure segmentation, including ERNet [103], CENet [104], and ROSE [105]. Specifically, a modified 3D fuzzy network from [83] was employed to investigate the potential of extant fuzzy theory-driven studies. All modules were trained using the identical patch dataset derived from SPS criteria. Furthermore, all comparative analyses were executed on both the open-access BAS and our in-house datasets to evaluate robustness and performance, respectively.

Evaluation metrics. Five metrics were utilized to assess the model performance, including IoU, DLR, DBR, precision and the airway missing ratio

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{3.16}$$

$$\text{AMR} = \frac{FN}{V_Y} \times 100\%. \tag{3.17}$$

Experimental Results

All the experimental results were shown as mean ± standard deviation.

Ablation studies. Results of the ablation experiments are presented in Table 3.2, illustrating the mean and standard deviation of the different evaluation metrics.

Table 3.2: Ablation studies of the proposed method on BAS dataset

Model	IoU	Precision	DLR(%)	DBR(%)	AMR(%)	CCF_s
BL	**0.8749±0.0385**	0.9190±0.0302	87.81±8.74	81.74±11.15	5.11±4.10	0.8763±0.0514
BL+Fuzzy	0.8731±0.0493	**0.9205±0.0326**	89.20±10.73	83.92±12.48	5.45±5.41	0.8814±0.0650
BL+JCAM	0.8584±0.0448	0.9027±0.0388	88.71±7.83	82.93±10.90	5.33±4.09	0.8710±0.0554
BL+Fuzzy+JCAM	0.8625±0.0414	0.9039±0.0366	**93.95±7.96**	**90.64±10.42**	**4.92±4.22**	0.8913±0.0527
BL+Fuzzy+CCF_s	0.8701±0.0522	0.9112±0.0315	89.63±3.27	85.24±12.53	5.71±5.82	0.8816±0.0412
BL+Fuzzy+JCAM+CCF_s	0.8738±0.0445	0.9187±0.0320	92.71±7.93	89.01±10.3	5.22±4.50	**0.8969±0.0554**

% indicates the ratio metrics, Model abbreviations **BL**: baseline 3D UNet with deep supervision.

It is notable that the baseline 3D U-Net attained competitive performance in airway segmenta-

tion, achieving an IoU score of 0.8749, a DLR of 87.81%, and a DBR of 81.74%. The incorporation of the fuzzy attention layer into the 3D-UNet framework led to a marked enhancement in DBR, with an average increase of approximately 2.2%. The addition of the JCAM loss further advanced the DBR to 82.93%, and the inclusion of the CCF score also contributed to the model's improvement, with a 1.3% rise in DBR. The most substantial enhancement was observed when combining the fuzzy attention layer with JCAM loss, culminating in a DBR of 90.64%, a DLR of 92.95%, and a minimal AMR of 4.92%. By integrating all these innovative approaches (BL+Fuzzy+JCAM+CCF_s), the proposed FANN secured the highest CCF score of 0.8969, alongside competitive metrics for DBR (89.01%), DLR (92.17%), and AMR (5.22%), albeit with a slight decrease in IoU and the dice score.

Comparison experiments on BAS data. On the BAS dataset, the proposed FANN model demonstrates state-of-the-art performance in airway segmentation, as evidenced in Table 3.3. The model achieved a CCF score of 0.8969, a DLR of 92.71%, and a DBR of 89.01%. Both ERNet [103], Rose [105], LFSCNN [83], and V-Net [73] exhibit high voxel-wise precision (0.9587, 0.9499, 0.9837, and 0.9781, respectively). However, V-Net shows a severe discontinuity in airway predictions, indicated by a high AMR of 27.15%, along with low DBR and DLR (22.04% and 33.96%, respectively). Attention UNet [81] presents competitive results with 0.8762 IoU, 88.61% DLR, and 82.43% DBR, followed closely by nnUNet [74] (with the DLR of 86.84%, IoU of 0.8805, and DBR of 79.21%) and WingsNet [87] (with a DLR of 88.00%, DBR of 83.11%, and IoU of 0.8445). Rose [105] registers the highest IoU score at 0.8825, accompanied by comparable DLR and DBR scores of 84.84% and 77.31%, respectively. The approach by Juarez et al. [77] also garners strong IoU (0.8371) and precision (0.9344) scores, albeit with a lower DLR (73.46%) and DBR (64.25%). NaviAirway achieves the lowest airway missing ratio at 4.13% but suffers from comparatively low precision at 0.8672. Contrary to expectations, BronchiNet exhibits limited effectiveness with a high airway missing ratio of 43.21%.

Comparison experiments on COVID-19 data. On the COVID-19 dataset, the proposed FANN model outperformed all competitors, achieving the highest CCF score of 0.9270, accompanied by a DLR of 93.30%, a DBR of 90.18%, and an AMR of 2.37% (refer to Table 3.4). The nnUNet [74] recorded the second-best CCF score at 0.9148, with a DLR of 91.36% and

Table 3.3: Comparison experiments on the open dataset BAS dataset

Model	IoU	Precision	DLR(%)	DBR(%)	AMR(%)	CCF_s
ERNet [103]*	0.8639±0.0395‡	0.9587±0.0205‡	83.72±8.40‡	75.86±10.59‡	9.58±3.70‡	0.8517±0.0518‡
Rose [105]*	**0.8825±0.0457**‡	0.9499±0.0208‡	84.84±7.88‡	77.31±9.59‡	7.41±4.83‡	0.8669±0.0562‡
CENet [104]*	0.8601±0.0410‡	0.9498±0.0240‡	81.33±11.43‡	73.37±14.50‡	8.82±3.79‡	0.8385±0.0575‡
LFSCNN [83]*	0.8193±0.0676‡	0.9537±0.0224‡	69.09±15.08‡	58.26±16.92‡	12.58±5.03‡	0.7564±0.0898‡
BronchiNet [102]*	0.4984±0.1596‡	0.8588±0.0621‡	64.28±16.28‡	57.52±18.01‡	43.21±18.65‡	0.5541±0.1610‡
NaviAirway [101]*	0.8348±0.0335‡	0.8672±0.0406‡	87.34±7.15‡	80.99±9.50‡	**4.13±3.04**‡	0.8516±0.0439‡
Juarez et al. [77]†	0.8371±0.0752‡	0.9344±0.0314	73.46±14.19‡	64.25±15.55‡	11.17±7.62‡	0.7879±0.0952‡
Wang et al. [91]†	0.7330±0.0786‡	0.7636±0.0737‡	85.05±12.27‡	78.58±14.20‡	5.07±6.19	0.7813±0.0937‡
WingsNet [87]*	0.8445±0.0596‡	0.9373±0.0303‡	88.00±11.95‡‡	83.11±13.46‡	10.32±6.89‡	0.8600±0.0768‡
Attention UNet [81]†	0.8762±0.0414	0.9207±0.0309	88.61±8.32‡	82.43±10.65‡	5.16±4.18	0.8806±0.0534‡
nnUNet [74]*	0.8805±0.0313	0.9436±0.0234‡	86.84±7.00‡	79.21±9.43‡	6.96±4.02‡	0.8750±0.0416‡
V-Net [73]*	0.7140±0.1716‡	**0.9781±0.0150**‡	33.96±17.96‡	22.04±14.22‡	27.15±17.99‡	0.4779±0.1751‡
Proposed	0.8738±0.0445	0.9187±0.0320	**92.71±7.93**	**89.01±10.3**	5.22±4.50	**0.8969±0.0554**

* refers to results obtained by an open-source module (with trained weights) or modules trained from open-source codes. † indicates the reproduced results. ‡ represents statistical significance (with Wilcoxon signed-rank test p-value <0.05) compared with the proposed method.

a DBR of 87.33%. Attention UNet [81] also showcased commendable performance with the highest IoU score of 0.923, a DLR of 88.53%, and a DBR of 82.97%. WingsNet [87] reached a performance similar to the proposed FAAN in terms of DLR (92.50%) and DBR (90.67%), with a significantly lower IoU of 0.8282. Surprisingly, V-Net [73] continued to exhibit subpar results, achieving only a DLR of 34.96% and a DBR of 24.07%. Intriguingly, ERNet [103], Rose [105], and NaviAirway [101] attained similar CCF scores (0.8801, 0.8801, and 0.8882, respectively), with NaviAirway presenting superior branch metrics compared to ERNet, achieving a DLR of 87.29% and a DBR of 80.60%, as opposed to ERNet's 84.97% DLR and 78.55% DBR.

Comparison experiments on fibrosis data. The models' performance experienced a notable decline in datasets derived from patients with fibrotic lung disease, as detailed in Table 3.4. The proposed FANN demonstrated exceptional performance and generalization capability, evidenced by a CCF score of 0.8099, a DLR of 78.98%, a DBR of 73.44%, and a precision of 0.8904. Attention UNet [81] secured impressive results with a DLR of 72.83%, a DBR of 66.88%, and precision of 0.8844, while WingsNet [87] attained a DLR of 69.46%, a DBR of 63.01%, a precision of 0.9401, and an AMR of 21.98%. nnUNet [74] did not mirror its success from the BAS and COVID datasets, reflected by lower DLR (58.15%) and DBR (50.18%) scores. Similarly, ERNet [103], Rose [105], LFSCNN [83], and NaviAirway [101] reported competitive IoU scores (0.7840, 0.7898, 0.7998, and 0.8074, respectively), yet their branch continuity metrics were relatively modest. Despite achieving comparable DBR and DLR to WingsNet, the

Table 3.4: Comparison experiments on our in-house dataset

	Model	IoU	Precision	DLR(%)	DBR(%)	AMR(%)	CCF_s
COVID-19	ERNet [103]*	0.9064±0.0458‡	0.9655±0.0253‡	84.97±8.42‡	78.55±10.89‡	5.54±2.21‡	0.8801±0.0575‡
	Rose [105]*	0.9132±0.0338‡	0.9552±0.0157‡	85.91±8.12‡	78.89±9.83‡	4.09±3.57‡	0.8882±0.0457‡
	CENet [104]*	0.8883±0.0559‡	0.9611±0.0157‡	77.30±14.19‡	68.86±15.24‡	5.61±3.98‡	0.8328±0.0767‡
	LFSCNN [83]*	0.8883±0.0559‡	0.9611±0.0157‡	77.30±14.19‡	68.86±15.24‡	5.61±3.98‡	0.8328±0.0767‡
	BronchiNet [102]*	0.4200±0.1762‡	0.7253±0.0532‡	64.07±23.74‡	62.20±23.91‡	49.29±23.32‡	0.4965±0.1992‡
	NaviAirway [101]*	0.8861±0.0298‡	0.9100±0.0245‡	87.29±7.92‡	80.60±10.34‡	2.86±2.39‡	0.8801±0.0413‡
	Juarez et al. [77]†	0.8637±0.1016‡	0.9566±0.0134	71.73±15.07‡	63.02±15.07‡	10.08±10.79‡	0.7914±0.1189‡
	Wang et al. [91]†	0.7433±0.1010‡	0.7749±0.0736‡	84.87±13.20‡	79.93±14.09‡	5.41±10.35‡	0.7870±0.1129‡
	WingsNet [87]*	0.8282±0.0672‡	0.9537±0.0215‡	92.50±6.89	**90.67±8.95**	13.68±7.15‡	0.8689±0.0680‡
	Attention UNet [81]†	**0.9230±0.0279**	0.9464±0.0175	88.53±6.66‡	82.97±8.34‡	2.62±2.21	0.9057±0.0377‡
	nnUNet [74]*	0.9158±0.0381‡	0.9685±0.0407‡	91.36±6.72‡	87.33±8.89‡	2.61±1.73‡	0.9148±0.0473‡
	V-Net [73]*	0.7898±0.0871‡	**0.9893±0.0040**‡	34.96±9.66‡	24.07±8.80‡	20.33±8.86‡	0.5052±0.0911‡
	Proposed	0.9222±0.0261	0.9431±0.0166	**93.30±5.29**	90.18±7.59	**2.37±1.64**	**0.9270±0.0338**
Fibrosis	ERNet [103]*	0.7840±0.0948‡	0.9359±0.0404‡	55.87±14.92‡	49.09±15.35‡	12.46±3.38‡	0.6641±0.1133‡
	Rose [105]*	0.7898±0.0705‡	0.8979±0.0687	65.17±14.41‡	58.47±16.33‡	12.79±7.21‡	0.7213±0.0913‡
	CENet [104]*	0.7602±0.1020‡	0.9459±0.0252‡	50.42±16.40‡	43.68±16.85‡	14.67±4.89‡	0.6194±0.1228‡
	LFSCNN [83]*	0.7998±0.0527‡	0.9308±0.0169‡	57.95±13.05‡	50.43±14.12‡	11.85±3.65‡	0.6835±0.0719‡
	BronchiNet [102]*	0.5709±0.1146‡	0.8929±0.0432‡	76.38±15.51‡	69.49±15.09‡	34.22±7.71‡	0.6436±0.1298‡
	NaviAirway [101]*	0.8074±0.0533‡	0.9247±0.0165‡	59.94±14.40‡	51.48±14.90‡	13.45±6.45‡	0.6989±0.0742‡
	Juarez et al. [77]†	0.7911±0.0532‡	0.9296±0.0294‡	56.16±10.76‡	48.29±12.18‡	15.68±6.37‡	0.6688±0.0688‡
	Wang et al. [91]†	0.6979±0.0647‡	0.7468±0.0773‡	69.61±9.24‡	62.61±11.17‡	8.22±3.88	0.6971±0.0747‡
	WingsNet [87]*	0.7436±0.0805‡	0.9401±0.0156‡	69.46±9.71‡	63.01±11.39‡	21.98±8.48‡	0.7208±0.0872‡
	Attention UNet [81]†	0.8202±0.0553	0.8844±0.0640	72.83±9.41‡	66.88±10.84‡	7.98±2.97	0.7764±0.0678‡
	nnUNet [74]*	**0.8312±0.0495**‡	0.9381±0.0314‡	58.15±6.80‡	50.18±7.93‡	11.74±2.93‡	0.6972±0.0564‡
	V-Net [73]*	0.5002±0.0967‡	**0.9768±0.0006**‡	8.54±3.76‡	3.40±2.99‡	49.36±9.88‡	0.1576±0.0568‡
	Proposed	0.8269±0.0402	0.8904±0.0373	**78.98±8.00**	**73.44±9.54**	**7.95±2.37**	**0.8099±0.0517**

* refers to results obtained by an open-source module (with trained weights) or module trained from open-source codes. † indicates the reproduced results. ‡ represents statistical significance (with Wilcoxon signed-rank test p-value <0.05) compared with the proposed method.

method by Wang et al. [91] was more prone to airway leakages (evidenced by a lower precision of 0.7468) and missing tracheal structures (indicated by low IoU scores). Remarkably, BronchiNet achieved improved performance on the fibrosis dataset compared to its results on the BAS and COVID-19 datasets, with a DLR of 76.38%, a DBR of 69.49%, and a CCF score of 0.6436.

Visualization results. The visualization of correct predictions, airway leakages, and missing airways are depicted in Fig. 3.4. Generally, airways in scans from COVID-19 patients were more extractable compared to those with the other two conditions. The intricate patterns characteristic of fibrotic lung disease, such as honeycombing, led to diminished performance in most modules due to increased predictions of missing airways. Notably, a majority of airway leakages were attributed to absent annotations. Specifically, nnUNet demonstrated proficiency with COVID-19 data yet struggled with missing airways in both BAS and fibrosis datasets. WingsNet showed promising results in detecting small-sized branches while overlooking the main trachea. While Rose was comparable on COVID-19 and BAS datasets, it exhibited substantial missing airways in fibrosis data, evidenced by a significant area of false negatives in its fibrosis predictions. The proposed method is observed to have superior robustness and perfor-

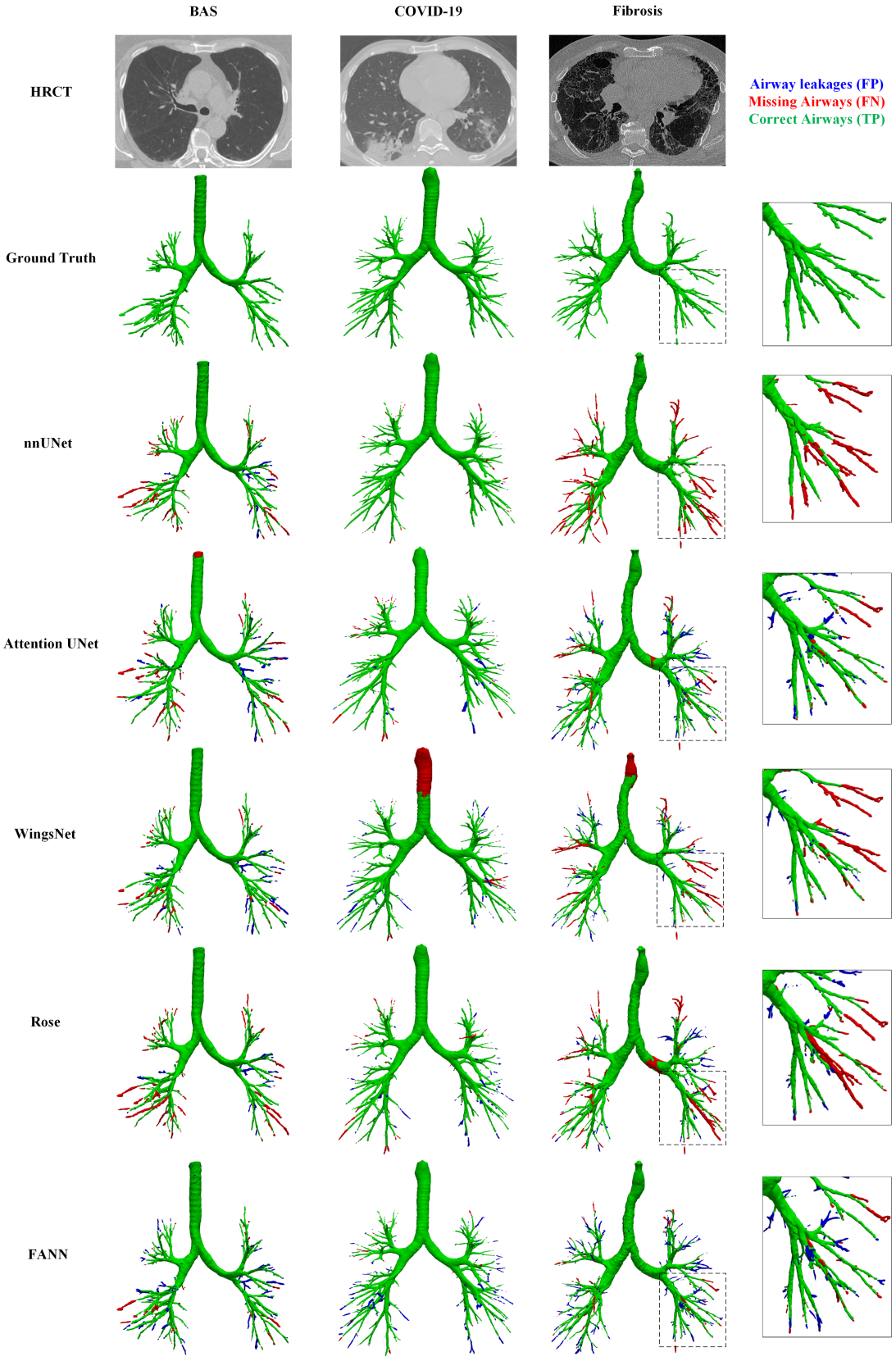


Figure 3.4: Visualization results of the competitive module, including nnUNet, Attention UNet, WingsNet, Rose and the proposed FANN on fibrosis data (first row), COVID-19 data (second row), and BAS data (third row), with the green, red, blue representing the correct airways (TP), missing airways (FN), and airway leakages (FP), respectively.

mance, achieving the fewest missing airways and the highest accuracy in airway extraction.

3.1.5 Discussion

This section delves into the comprehensive performance analysis, underscores the significance of data sampling, elucidates on the issues of airway leakage and omission in predictions, emphasizes the importance of evaluation metrics, and discusses the detection rate across branches of various sizes.

Overall performance analysis. The empirical results from both public and our in-house datasets underscore the outstanding capacity of the proposed technique for airway segmentation. V-Net and BronchiNet, in particular, manifested the least favorable outcomes due to discontinuities in airway predictions, despite all models suffering post-processing to retain only the largest 3D component. Notably, BronchiNet registered the most inadequate performance among the preceding models for airway extraction, with diminished IoU scores and branch metrics. The methods introduced by Juarez et al. and Wang et al. displayed similar efficacy. WingsNet outperformed NaviAirway, showcasing nearly a 10% enhancement in branch metrics. Among methods adapted from different imaging modalities, Rose asserted superior robustness against counterparts such as ERNet and CENet. Herein, we principally compare four models against our proposed framework, comprising attention UNet, nnUNet, Rose, and WingsNet. Across all sub-datasets, nnUNet, Rose, and WingsNet depicted comparable precision against attention-based models. Nonetheless, WingsNet's AMR was approximately double that of nnUNet and Rose, whereas its DBR and DLR were substantially superior, signifying less accurate predictions on the principal trachea but enhanced detection of small to medium-sized branches. Among the models, attention UNet demonstrated superior performance across all datasets, showing the efficacy of the attention mechanism.

FANN exhibited significantly enhanced performance in terms of DLR, DBR, and CCF scores on the public BAS dataset when compared to Rose, nnUnet, attention UNet, and WingsNet. Although the IoU for FANN (0.8738) did not show a significant difference from that of nnUNet

(0.8805) and attention UNet (0.8762), it was superior to that of WingsNet (0.8445). In the case of COVID-19 data, FANN surpassed Rose, nnUnet, and attention UNet in both DLR and DBR metrics. Despite a relatively lower IoU score (0.8282), WingsNet managed to attain a DLR and DBR comparable to FANN, as per the Wilcoxon signed-rank test with a p-value greater than 0.05. On datasets from patients with fibrotic lung disease, nnUNet outperformed FANN in IoU and precision but exhibited significant discontinuities, as evidenced by its lower DLR (58.15%) and DBR (50.18%). WingsNet, although achieving the highest precision (0.9401), also had the largest AMR (21.98%), indicative of substantial false-negative predictions. FANN consistently achieved the best CCF score, DLR, DBR, and AMR across all comparisons, maintaining competitive IoU and precision scores.

Fuzzy theory-based method. While some studies have combined fuzzy theory with deep neural networks, only a handful are applicable to image segmentation. The modification of 2D fuzzy layers in LFSCNN [83] to 3D ones yielded comparable results to other approaches. Specifically, the fuzzy module within LFSCNN was positioned at the residual skip connection, bridging the encoder and decoder blocks for feature fusion. However, on BAS, COVID-19, and Fibrosis datasets, LFSCNN's performance was a bit weak, with a branch metrics gap of nearly 20% compared to FANN. In further exploration of fuzzy integration in [83], ablation studies showed an enhancement in LFSCNN without the fuzzy fusion module, achieving an IoU of 0.8462, a DLR of 76.14%, and a DBR of 66.36% for BAS data, and an IoU of 0.8507, a DLR of 69.63%, and a DBR of 59.60% for COVID-19 data. Conversely, a decrease in performance was observed on fibrosis data, with an IoU of 0.7808, a DLR of 51.43%, and a DBR of 42.46%. This suggests that the fuzzy feature integration module did not confer advantages on datasets with mild lung pathology but did show improvement with more complex pathological changes, such as honeycombing.

Data sampling strategy. Data sampling strategy is crucial for patch-based approaches, especially in imbalanced datasets. This study introduced a smart data sampling strategy for patch extraction. Experiments on the open BAS dataset with varying data sampling strategies highlighted the importance of this aspect. We defined three extraction criteria: (1) Sequential: patches are extracted sequentially by sliding windows within the ground truth's minimum

bounding box. (2) Drop: similar to Sequential, but either the main trachea (Drop-A) or pure negative patches (Drop-B) are discarded to balance the data. (3) Smart patch sampling. We also examined three patch sizes: (1) a uniform $128 \times 128 \times 128$, (2) proportions close to the median ground truth volume ($160 \times 96 \times 160$), and (3) proportions close to the mean ground truth volume ($128 \times 96 \times 144$).

Table 3.5: Comparison experiments of data sampling strategy

	Extraction	Patch size	IoU	Precision	DLR(%)	DBR(%)	AMR(%)
Cut-1	Sequential	uniform	0.8782±0.0641	0.9341±0.0409	85.17±14.28	76.41±15.27	7.64±7.22
Cut-2	Sequential	median	0.8854±0.0470	0.9408±0.0300	85.79±8.12	78.77±10.06	6.03±4.63
Cut-3	Sequential	mean	0.8862±0.0422	0.9394±0.0275	86.66±8.88	79.39±10.80	6.06±4.61
Cut-4	Drop-A	mean	0.8069±0.0826	0.9437±0.0316	84.64±10.00	77.57±12.70	14.87±8.13
Cut-5	Drop-B	mean	**0.8883±0.0407**	**0.9451±0.0286**	87.73±8.01	80.92±10.51	5.95±4.25
Cut-6	Smart	mean	0.8749±0.0385	0.9190±0.0302	**87.81±8.74**	**81.74±11.15**	**5.11±4.10**

The 'uniform' patch size refers to 128×128×128, while the 'median' and 'mean' are 160×96×160 and 128×96×144, respectively.

Although little increase in IoU was observed across Cut-1, Cut-2, and Cut-3, a marked enhancement in DLR and DBR was recorded. Moreover, the results of Cut-3 and Cut-4 (as shown in Table 3.5) reveal that segmentation performance can be amplified by omitting pure negative patches, as evidenced by a 1.6% increase in DBR and a 1% rise in DLR. However, excluding the main trachea proved counterproductive, as the model failed to predict it after training predominantly on smaller branch features (as demonstrated in Cut-4). The model trained with Cut-5 exhibited a competitive DBR, improving by approximately 4.5% relative to Cut-1. Ultimately, the SPS strategy we proposed attained the highest DLR, DBR, and the lowest AMR.

Airway leakages and Airway neglect. Airway leakages, or false-positive predictions, are reflected in the precision score, while airway neglect, or false-negative predictions, are measured by the AMR. Fig 3.4 illustrates that some leakages may be attributed to incorrectly unannotated airway branches. nnUNet, while achieving a high precision score indicating limited airway leakages, suffers from relative AMR with notable airway neglects, particularly in fibrosis cases. WingsNet shows significant airway neglect of the main trachea, leading to a lower IoU score despite comparable DBR and DLR. Rose, WingsNet, attention UNet, and FANN all exhibit some leakage at terminal branches, pointing to incomplete airway annotations. Among all models studied, FANN records the smallest AMR, with most of its leakages linked to incomplete annotations. Hence, we advocate for a limited number of airway leakages to improve model

prediction of terminal bronchioles.

Evaluation metric. Conventional metrics like the IoU or dice score are insufficient in identifying the optimal model during training. As depicted in Tables 3.3, 3.4, the model with the highest IoU score does not necessarily possess superior segmentation capabilities. For instance, nnUNet scored highest in IoU on the BAS dataset but performed poorly in fibrosis evaluations. Focusing solely on branch performance while neglecting overlap metrics can also result in airway leakages. For example, the BL+fuzzy+JCAM model showed better DLR and DBR but a lower CCF score than the final FAAN combination in ablation studies (Table I). Evaluating BL+fuzzy+JCAM on fibrosis and COVID-19 datasets revealed IoU scores of 0.7961 (0.9099), DLR of 74.66% (93.81%), DBR of 68.85% (91.87%), AMR of 10.29% (2.36%), and CCF scores of 0.7687 (0.9215) on fibrosis (COVID-19) data, respectively, indicating a notable disparity compared to FAAN. This underscores the effectiveness of the proposed CCF score in aiding module selection.

Table 3.6: Detection rate of different-sized branches

	TB	SB	MB	LB
BAS	81.55%	95.76%	98.59%	99.38%
COVID-19	82.96%	97.05%	99.67%	100%
Fibrosis	47.15%	87.60%	98.32%	98.50%

The detection rate of different-sized branches. To assess the segmentation efficacy across airway branches of varying sizes, we calculated the detection rates based on their radii. A branch is considered correctly detected if its IoU surpasses 0.8. We categorized airway branches into four size groups: TB (terminal branches with diameters from 0 to 2mm), SB (small branches from 2mm to 4mm), MB (middle branches from 4mm to 8mm), and LB (large branches exceeding 8mm in diameter). The performance of the proposed FAAN model on BAS, COVID-19, and fibrosis datasets is detailed in Table 3.6. It reveals that FAAN outperforms in segmenting airways in both BAS and COVID-19 datasets. However, the detection rate of terminal branches in fibrotic lung disease is comparatively lower. This reduction is attributed to the complex honeycombing patterns in fibrotic lungs, which present similar structure characteristics to airways.

Table 3.7: Computational complexity of different methods

	# Parameter (M)	FLOPs (G)	Avg speed (s)
WingsNet	1.47	259.49	20.39‡
ERNet	5.11	341.26	26.62‡
LFSCNN	29.36	649.25	31.93‡
FANN	14.12	308.86	35.17
3D U-Net	13.85	298.70	35.77
Attention U-Net	14.03	305.47	36.82‡
CENet	74.68	1530	50.08‡
Bronchinet	5.88	122.10	92.71‡
nnUNet*	31.19	532.96	169.8‡
Rose	20.13	10940	177.08‡

*The employed nnUNet was built followed by a "3d full resolution" configuration, where the details can be found at https://github.com/MIC-DKFZ/nnUNet. ‡ indicates significant differences (p-value <0.05). "# parameters" refers to the number of parameters.

Computational complexity. We report and rank the computational complexities of all methods in Table 3.7 based on the average inference time. The number of network parameters (# Parameter), floating point operations (FLOPs), and Average inference speed (Avg speed) are presented to illustrate each model's comprehensive computational efficiency. The parameter count influences GPU memory consumption during inference, while FLOPs are indicative of model complexity and are crucial for inference speed. As shown in Table 3.7, WingsNet is the most lightweight model with 1.47M parameters and an inference speed of 20.39s, followed by ERNet, which has 5.11M parameters and a time cost of 26.62s. FANN and 3D U-Net exhibit similar average inference speeds, with no significant differences (p-value >0.05) between them. nnUNet, with its large number of parameters and FLOPs, demonstrates a high computational complexity, especially in image pre/post-processing. Despite its relatively few parameters (5.88M), BronchiNet's extensive pre/post-processing results in longer processing times (nearly 3 minutes). Additionally, CENet and Rose, originally designed for 2D images, become more complex when adapted to 3D architectures. Rose, despite having only 20.12M trainable parameters, results in a substantial computational cost with 10.94T FLOPs, leading to an average inference speed of 177.08s.

Potential studies. This research underscores the potential of integrating fuzzy logic into deep attention neural networks. There are several promising avenues for future investigation, such as

incorporating alternative membership functions (MFs) like a generalized bell or sigmoid MFs.

3.1.6 Conclusion

Airway segmentation plays a pivotal role in aiding clinicians with disease assessment and prognosis analysis. However, existing approaches struggle with discontinuous predictions, particularly for small-sized airway branches. This study highlights the criticality of data sampling strategies and evaluation metrics. The efficacy of fuzzy attention neural networks, Jaccard continuity and accumulation mapping loss, and the CCF score in 3D airway segmentation have been substantiated. Our proposed method has exhibited outstanding performance across both open-access datasets and our in-house datasets, including scans from patients with cancer, COVID-19, fibrotic lung disease, and mild lung disorders.

The observed suboptimal detection rate for terminal branches in fibrosis data signals is exposed for future research. Future studies might explore innovative fuzzy membership functions, adversarial learning techniques, and diffusion modules to bolster terminal branch extraction in cases of fibrotic lung disease.

3.2 Adversarial Transformer for Airway Purification

3.2.1 Introduction

Airway segmentation plays a vital role in diagnosing and prognosticating pulmonary diseases through chest computed tomography (CT) scans. The manual segmentation process, undertaken by radiologists, is notably labor-intensive and prone to errors, due to the extensive volume of CT data and the complex pattern of the airway trees. To alleviate the burden of manual labelling for radiology experts, a range of automated and semi-automated algorithms are being developed. The efficacy of conventional segmentation techniques, such as thresholding [71] and region-growing [67], was evaluated in the EXACT'09 challenge [72], where no algorithm succeeded in extracting more than an average of 74% of the total tree length. In recent years, convolutional neural networks (CNNs) have demonstrated promising outcomes in airway segmentation. Notably, models such as U-Net [106], 3-D U-Net [79], V-Net [73], and their variants [107] [108] [89] have been applied to this task.

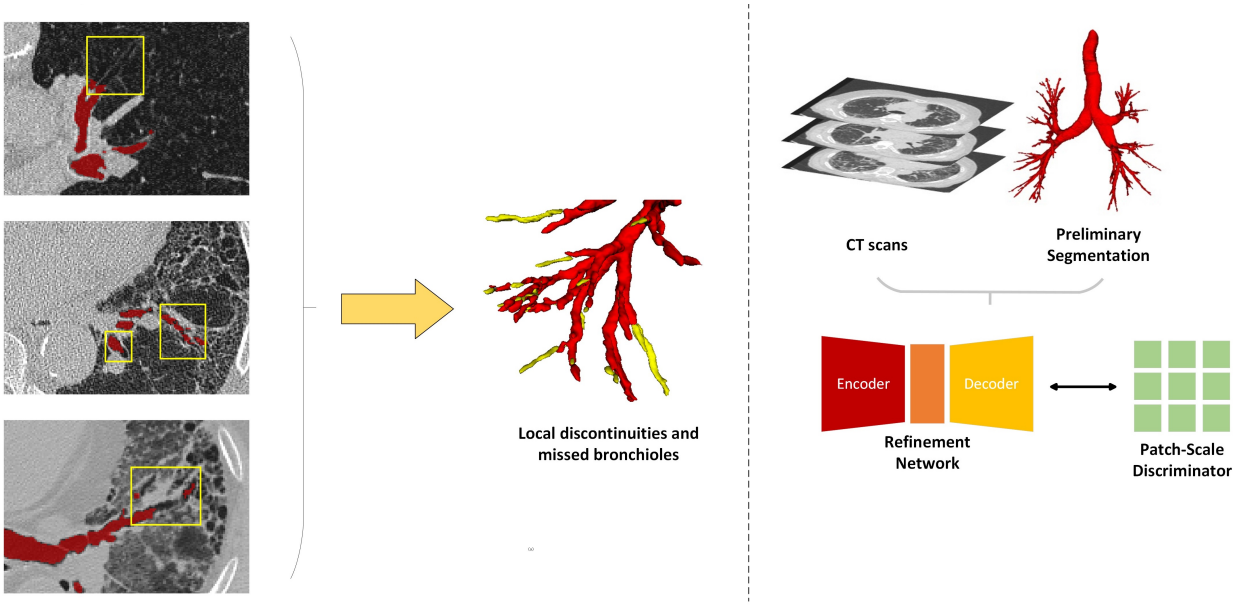


Figure 3.5: Challenges and solutions in airway modelling. (left) demonstrates issues like unclear boundaries, disease-related changes, and low image quality (marked in yellow boxes) causing segmentation breaks and overlooked branches (shown in yellow). (right) introduces our solution: an adversarial-based refinement network aimed at thorough and reliable segmentation. We use patch-scale discriminators to address local structural inaccuracies.

Despite notable advances in the Dice score attained by CNN-based models for airway segmen-

tation, local discontinuity persists as a significant challenge, as depicted in Figure 3.5. This issue is especially pronounced in the lung's peripheral region, characterized by an imbalance between the distal small airways (the target) and the background. Additionally, due to high computational demands, these deep neural networks are trained on image patches, making it difficult to detect minor errors until the patches are combined to form the full image. Recent studies by [91], [86], and [87] have endeavored to mitigate discontinuity through the development of complex architectures and loss functions. Nonetheless, these approaches do not fully accommodate the variability in airway structure attributed to pathological conditions like honeycombing, bronchial wall thickening, and bronchiectasis. Such variability can diminish model robustness, leading to omissions of bronchi or bronchioles in the segmentation results. Furthermore, the heterogeneity of data from various medical centres and the diversity in airway branching patterns among individuals exacerbate these challenges.

To address these challenges, this study introduces a straightforward yet effective adversarial-based patch-scale refinement network. This network aims to enhance the connectivity of preliminary segmentation in both normal and pathological cases. We hold that refining existing segmentation is a feasible approach, given the difficulty of devising a single model that adequately addresses the task. Distinguishing our work from other GAN-based medical segmentation models, we implement several novel enhancements. Specifically, the *tanh* activation function is utilized in the generator's final layer to adjust pixels in preliminary labels, steered by two centreline objective functions. The refined labels first undergo dilation with a fixed-size kernel, followed by multiplying with the CT image before the discriminator. This dilation is designed to broaden the discriminator's visual field, potentially improving its ability to discern real from fake pairs. The discriminator evaluates image patches rather than the entire image, enabling localized penalization of inaccurate segmentations and enhanced capture of high-frequency details, such as terminal brunches. Our model was evaluated on three datasets: the BAS dataset (n=90) and two in-house datasets (n=50), each consisting of 25 cases of pulmonary fibrosis and COVID-19. We utilized the detected length ratio, detected branch ratio, and false negative rate as evaluation metrics. The model demonstrated significant improvements, exceeding 15% in all three metrics. Additionally, our refinement pipeline also effectively enhanced the preliminary

segmentation from other models, evidenced by metric increases of over 10%. An ablation study was conducted to dissect the contribution of each element within our refinement framework.

3.2.2 Related works

Transformer-based medical segmentation. Transformers were first suggested for NLP tasks because of their capacity to encapsulate long-range relationships. However, recent research [109] has shown that this design may also be effective for processing images. Consequently, there has been a significant increase in applying transformers in medical imaging studies. For 3D segmentation, Wang et al. [110] introduced TransBTS, which incorporates a transformer module between the encoder and decoder as the bottleneck. Hatamizadeh et al. [111] developed the U-Net Transformer (UNETR), utilizing a transformer as the U-Net's encoder. Li et al. [112] proposed the Squeeze-and-Expansion Transformer, featuring a squeezed attention block for regulating the self-attention module and an expansion block for learning diversified representations. Liu et al. [113] suggested a hierarchical transformer that integrates a shifted window scheme into the multi-head attention (MSA) module. Unlike these approaches, our method uniquely employs the transformer module as a discriminator rather than a generator.

Adversarial-based medical segmentation. Adversarial-aided segmentation has rapidly gained traction in medical imaging. Zhang et al. [114] utilized a modified 2-D dense U-Net for initial airway prediction, followed by morphological operations for small airway extraction and adversarial training with a cGAN using a 3-D dense U-Net. Zhao et al. [115] introduced a generative adversarial learning model with dilated residual blocks as the generator and a multi-layer CNN with residual blocks as the discriminator. Similarly, Guo et al. [116] employed a dense U-Net with an inception module as the generator, and Park et al. [117] designed an M-GAN with a multi-layer CNN with residual blocks as the discriminator. Different from these studies, the proposed method uniquely adopts a patch-scale discriminator.

3.2.3 Method

This section demonstrates our innovative refinement approach, encompassing a generator $\mathcal{G}$ and a discriminator $\mathcal{D}$ as illustrated in Figure 3.6. The generator $\mathcal{G}$ receives the preliminary segmentation x_i alongside its corresponding CT image y_i, subsequently generating a refined label $x_r = \mathcal{G}(y_i, x_i)$. The discriminator $\mathcal{D}$ endeavors to differentiate between $y_i \cdot (x_r \oplus k)$ and $y_i \cdot (x_t \oplus k)$, where k represents a kernel sized $5 \times 5 \times 5$, and x_t denotes the ground truth.

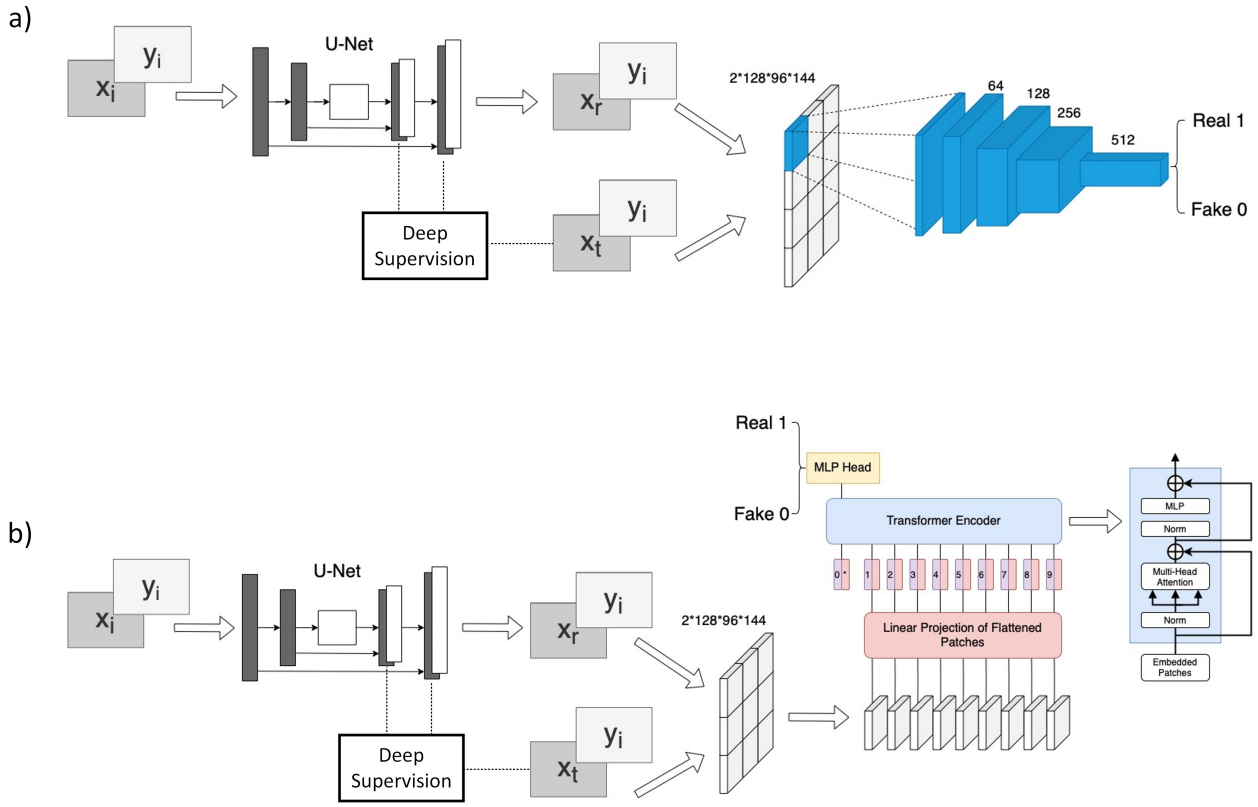


Figure 3.6: The refined airway segmentation model employs a U-Net generator with multi-scale supervision and a patch-scale discriminator. The discriminator is either (a) a PatchGAN with five convolution layers for feature extraction or (b) a Vision Transformer (ViT) that divides the image into patches and transforms them into fixed-size vectors with positional embeddings.

Generator

The generator is a 3-D U-Net [79] with multi-scale deep supervision [118], tailored for airway structure segmentation. It takes the fusion of CT images and preliminary labels as input, and outputs patch masks with the size of $128 \times 96 \times 144$. These preliminary labels are generated using a compact version of U-Net [77]. The initial segmentation's efficacy was assessed using

various metrics, delineated in Table 3.8. In refining the segmentation, we replaced the sigmoid activation with *tanh*, which outputs in the range of $[-1, 1]$. We set the threshold at 0, classifying outputs below this as -1 (to remove false positives) and above as 1 (to eliminate false negatives). Additionally, specific loss functions were employed to maintain segmentation connectivity, to be elaborated in a later section.

Discriminator

In order to explore novel discriminators in medical segmentation, we tested PatchGAN [119] and vision transformer [109]. PatchGAN aims to classify if each image patch (N×N) is real or fake. Conversely, a vision transformer considers an image as a sequence of patches, focusing on long-range dependencies. These patches are encoded into vectors while retaining positional information via embedded positions. Our hypothesis posits that a well-trained discriminator could enhance the generator's focus on finer structures, reducing discontinuities in peripheral bronchioles. The refined and ground truth masks are dilated using a $5 \times 5 \times 5$ kernel and then concatenated with their corresponding images. These concatenations are then fed into a 5-layer Markovian discriminator, classifying each $70 \times 70 \times 70$ patch as either real or fake. Additionally, we adapted a ViT-Base (number of layers = 12, Hidden Size = 768, MLP size = 3072, heads = 12) as our discriminator, replacing *tanh* in the MLP head with LeakyReLU (negative slope = 0.2) and setting the patch size to $32 \times 38 \times 36$.

Loss Functions

To effectively train our refinement model, we utilize a composite loss function integrating various loss components. The adversarial loss is combined with the Dice loss [73], the cl-Dice loss [120], and the Continuity and Completeness F-score (CCF) [121]. The mathematical formulations of these loss functions are as follows:

$$\text{Dice} = \frac{2\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} g_i^2} \tag{3.18}$$

where $p_i \in P$ represents the predicted binary segmentation and $g_i \in G$ the ground truth binary volume, with the summation over N voxels. For preserving the topological integrity of predicted

airway labels, we adopt clDice defined by:

$$\text{clDice} = \frac{2 \times T_{\text{prec}} \times T_{\text{sens}}}{T_{\text{prec}} + T_{\text{sens}}} \tag{3.19}$$

where Topology precision $T_{\text{prec}} = \frac{|S_P \cap V_L|}{|S_P|}$ and Topology sensitivity $T_{\text{sens}} = \frac{|S_L \cap V_P|}{|S_L|}$. Here, V_L and V_P denote the ground truth and predicted masks, respectively, and S represents the skeletonized version of V. The combined Dice and clDice loss is:

$$L_D = (1 - \alpha)(1 - \text{Dice}) + \alpha(1 - \text{clDice}) \tag{3.20}$$

with weight parameter $\alpha \in [0, 0.5]$. This aims to force the model to focus on the completeness (assessing all voxels) and continuity (assessing the centrelines) of the predictions.
To further boost centreline detection, we incorporate a loss focused on continuity and completeness:

$$\begin{aligned} L_{CCF} &= 1 - C \\ &= 1 - \frac{\sum X \cdot Y_{CL}}{\sum Y_{CL}} \end{aligned} \tag{3.21}$$

where X is the predicted airway label and Y_{CL} the ground truth centreline. In addition, L1 loss is employed to maintain overall segmentation accuracy. These loss functions are aggregated for each supervised layer j:

$$L_j = \psi_1 L_1 + \psi_2 L_{CCF} + \psi_3 L_D + \psi_4 L_{cGAN} \tag{3.22}$$

where ψ_1 to ψ_4 are weight parameters within the range $[0, 1.0]$, and L_{cGAN} represents the GAN adversarial loss:

$$L_{cGAN}(G, D) = \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_{x,z}[\log G(x, z)] \tag{3.23}$$

The overall loss is calculated across all layers by:

$$L = \phi_1 L_{\text{layer1}} + \phi_2 L_{\text{layer2}} + \phi_3 L_{\text{layer3}} + \phi_4 L_{\text{final}} \tag{3.24}$$

with ϕ_1 to ϕ_4 being the weight parameters for each layer.

3.2.4 Experiments and Results

Dataset

We trained and evaluated our model on the Binary Airway Segmentation (BAS) dataset, comprising 90 CT scans, with 20 from the EXACT'09[2] training set [72] and 70 from LIDC-IDRI [100]. The original LIDC-IDRI dataset includes 1018 cases but lacks airway annotations. [86] and [87] selected 70 cases with a slice thickness of ≤ 0.625 mm and meticulously annotated them[3]. We divided the 90 scans into 72 for training and 18 for testing. Additionally, our model was tested on in-house fibrosis datasets (25 cases) and COVID-19 datasets (25 cases).

3.2.5 Implementation Details

The model was implemented using PyTorch and trained for 150 epochs on an NVIDIA GeForce RTX 3090. An *Adam* optimizer with a learning rate of 0.001 was used, with the rate decaying by half at the 50th, 80th, 100th, and 120th epochs for the generator. The discriminator also utilized an *Adam* solver with a learning rate of 0.001. Backpropagation was performed first on the discriminator, followed by the generator. Data augmentation included random horizontal and vertical flips with a 0.5 probability during training. The hyperparameter α was set to 0.5, as recommended by [120]. The values of ψ_1 to ψ_4 were set at 0.1, 0.6, 0.3, and 0.01, respectively, and ϕ_1 to ϕ_4 were empirically set at 0.1, 0.1, 0.3, and 0.5, respectively.

In the inference stage on the test dataset, input patches were extracted using a sliding window approach (window size = $128 \times 96 \times 144$) with 50% overlap in the x, y, and z directions. Post-processing involved stitching binary prediction patches back into a full image and retaining the largest connected component to minimize noise.

[2] http://image.diku.dk/exact/

[3] https://geronsushi.github.io/lung.html

Evaluation Metrics

In this study, we use the same evaluation metrics as we demonstrated in work 2.2.

Segmentation Results

For a thorough evaluation of our model's performance, we employed seven metrics: intersection over union (IoU), dice coefficient, detected length ratio (DLR), detected branch ratio (DBR), precision, leakage, and false negative rates (AMR). Additionally, we conducted a Wilcoxon signed-rank test with a significance level of $\alpha = 0.01$ to perform statistical analysis.

Table 3.8: Preliminary segmentation generated for training

	Model	IoU	Dice	DLR	DBR	Precision	Leakages	AMR
Valid	small U-Net	0.8372±0.0752	0.9093±0.0505	0.7346±0.1419	0.6425±0.1555	0.9344±0.0314	0.0621±0.0269	0.1117±0.0762
Fibrosis	small U-Net	0.7911±0.0532	0.8823±0.0349	0.5616±0.1076	0.4829±0.1218	0.9296±0.0294	0.0657±0.0362	0.1568±0.0637
COVID	small U-Net	0.8637±0.1017	0.9228±0.0756	0.7173±0.1507	0.6302±0.1507	0.9566±0.0134	0.0407±0.0137	0.1008±0.1079

Table 3.9: Comparison experiments on different datasets

	Model	IoU↓	Dice↓	DLR↑	DBR↑	Precision↓	Leakage↑	AMR↓
Valid	Wang et al. [91]†	0.7330±0.0786‡	0.8434±0.0554‡	0.8505±0.1227‡	0.7858±0.1420‡	0.7636±0.0737‡	0.3070±0.1412‡	0.0507±0.0619
	WingsNet [87]*	**0.8544±0.0673**‡	**0.9120±0.0422**‡	0.8698±0.1175‡	0.8166±0.1305‡	**0.9458±0.0271**‡	**0.0529±0.0300**‡	0.1002±0.0769‡
	NaviAirway [101]*	0.8348±0.0335	0.9096±0.0200	0.8734±0.0715	0.8099±0.0950‡	0.8672±0.0406	0.1500±0.0536	**0.0413±0.0304**
	U-Net+PatchGAN	0.8150±0.0519	0.8971±0.0330	0.8828±0.1012	0.8337±0.1263	0.8556±0.0393	0.1620±0.0522	0.0544±0.0525
	U-Net+ViT	0.8132±0.0518	0.8961±0.0334	**0.8902±0.0967**	**0.8439±0.1261**	0.8600±0.0401	0.1549±0.0539	0.0623±0.0504
Fibrosis	Wang et al. [91]†	0.6979±0.0647‡	0.8203±0.0462‡	0.6961±0.0924‡	0.6261±0.1117‡	0.7468±0.0773‡	0.3272±0.1445‡	0.0823±0.0388
	WingsNet [87]*	0.8052±0.0539‡	0.8910±0.0440‡	0.6951±0.0977‡	0.6198±0.1175‡	**0.9505±0.0116**‡	**0.0438±0.0112**‡	0.1595±0.0576‡
	NaviAirway [101]*	**0.8074±0.0533**‡	**0.8924±0.0440**‡	0.5994±0.1440‡	0.5148±0.1490‡	0.9247±0.0165‡	0.0714±0.0188‡	0.1345±0.0645‡
	U-Net+PatchGAN	0.7481±0.0678	0.8541±0.0464	0.7157±0.1067	0.6437±0.1195	0.8035±0.0775	0.2374±0.1286	**0.0823±0.0332**
	U-Net+ViT	0.7272±0.0631	0.8405±0.0436	**0.7242±0.1096**	**0.6550±0.1266**	0.7879±0.0816	0.2606±0.1401	0.0916±0.0325
COVID	Wang et al. [91]†	0.7433±0.1010‡	0.8481±0.0798‡	0.8487±0.1320‡	0.7993±0.1409‡	0.7749±0.0736‡	0.2843±0.1500‡	0.0541±0.1035‡
	WingsNet [87]*	**0.9176±0.0363**‡	**0.9566±0.0209**‡	0.9073±0.0647	0.8682±0.0831	**0.9683±0.0204**‡	**0.0311±0.0203**‡	0.0544±0.0283‡
	NaviAirway [101]*	0.8861±0.0298‡	0.9394±0.0171‡	0.8729±0.0792‡	0.8060±0.1034‡	0.9100±0.0245‡	0.0970±0.0316‡	0.0286±0.0239
	U-Net+PatchGAN	0.8292±0.0335	0.9063±0.0203	0.8911±0.0827	0.8443±0.1044	0.8506±0.0413	0.1741±0.0575	**0.0281±0.0258**
	U-Net+ViT	0.8090±0.0483	0.8936±0.0307	**0.9104±0.0751**	**0.8772±0.0987**	0.8354±0.0458	0.1933±0.0705	0.0378±0.0231

* refers to results obtained from open-source implementations with model weights provided.
† refers to reproduced results.
‡ represents statistical significance (with Wilcoxon signed-rank test $p<0.01$) compared with the proposed method.

Comparison on BAS dataset: Our proposed method demonstrated superior performance on the BAS dataset (refer to Table 3.9 and Figures 3.7, 3.8), achieving a DLR of 0.8902, a DBR of 0.8439, and an AMR of 0.05441. Among the competing models, WingsNet recorded the highest IoU at 0.8544 and voxel-wise precision at 0.9458. NaviAirway exhibited the lowest average surface distance (ASD) of 0.0413, while the model by Wang et al. encountered a high leakage rate of 0.3070. Both NaviAirway and Wang et al.'s models demonstrated significant results in terms of DLR and DBR.

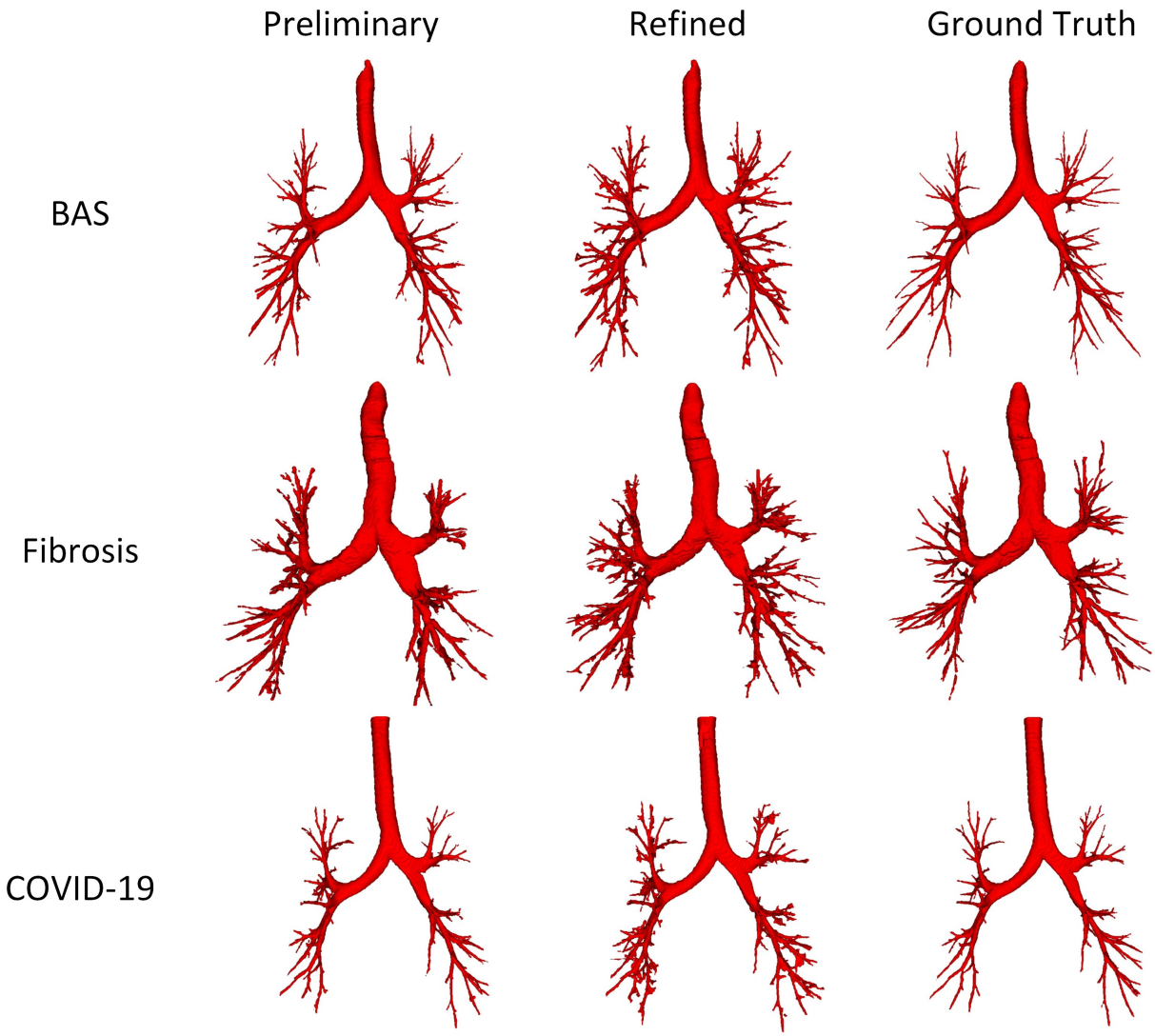


Figure 3.7: Visualization of the preliminary segmentation, refined segmentation and the ground truth on the BAS dataset (CASE 02) as well as our in-house dataset (CASE 162_01 in fibrosis and CASE RM451 in COVID-19). While the overall structure appears similar, differences are mainly observed at a fine scale. Therefore, zoomed-in views are provided in figures 3.8 to 3.11.

Comparison on in-house fibrosis dataset: Performance notably declined across all models in the fibrosis cases (as shown in Table 3.9 and Figures 3.7, 3.8). Nonetheless, our model still outperformed others, with the highest DLR of 0.7242, DBR of 0.6550, and the lowest AMR of 0.0823. WingsNet and the model by Wang et al. achieved comparable results in DLR and DBR, whereas WingsNet secured a better IoU score. Surprisingly, NaviAirway, despite attaining the highest IoU, underperformed in maintaining continuity, as reflected in its DLR of 0.5994 and DBR of 0.5148.

Comparison on in-house COVID-19 dataset: All models perform better on the COVID-

19 dataset compared to the BAS dataset (Table 3.9 and Figures 3.7, 3.8). The proposed method achieved the best DLR (0.9104), DBR (0.8772), and AMR (0.0281) in COVID-19 cases. WingsNet followed closely with the second-highest DLR (0.9073) and DBR (0.8682), and NaviAirway exhibited a DLR of 0.9394 and a DBR of 0.8729. Additionally, WingsNet attained the highest IoU (0.9176), Dice (0.9566), and precision (0.9683).

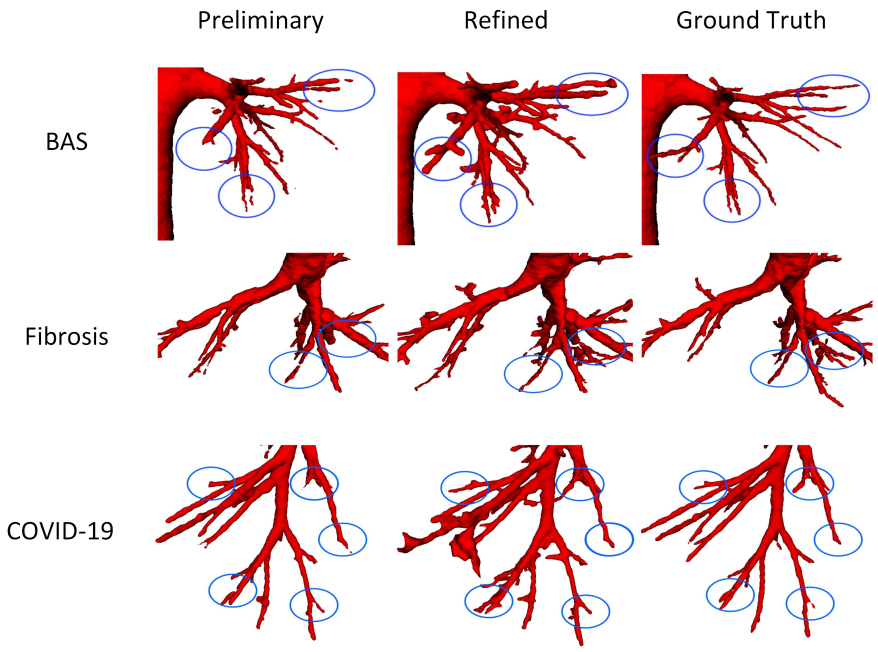


Figure 3.8: Visualization of a zoomed-in region in the preliminary segmentation, refined segmentation, and ground truth on the BAS dataset (CASE 02) and our in-house datasets (CASE 162_01 in fibrosis and CASE RM451 in COVID-19).

Refined segmentation on other three models: To validate the refinement model, we applied it to previous studies including WingsNet, NaviAirway, and the model developed by Wang et al. Across all datasets (BAS, fibrosis, and COVID-19), we observed significant enhancements in performance post-refinement, with DLR and DBR witnessing an average increase of over 10% (refer to Table 3.10). A slight trade-off was noted in IoU and Dice scores, although some reductions were not statistically significant. Furthermore, the false negative ratio showed a notable decrease in all scenarios. It is important to highlight that while these improvements were substantial, they were not as pronounced as those achieved with the refinement of our own preliminary results. Figures 3.9 to 3.11 offer a detailed visual representation of the refinement's impact.

Table 3.10: Refined segmentation results

	Model	IoU↓	Dice↓	DLR↑	DBR↑	Precision↓	Leakages↑	AMR↓
Valid	Wang et al. [91]†	0.7330±0.0786‡	0.84348±0.0554‡	0.8505±0.1227‡	0.7858±0.1420‡	0.7636±0.0737‡	0.3070±0.1412‡	0.0507±0.0619‡
	Refined Wang et al.	0.7136±0.1000	0.8284±0.0763	0.9093±0.1263	0.8700±0.1380	0.7341±0.0793	0.3552±0.1281	0.0423±0.0960
	WingsNet [87]*	0.8544±0.0673	0.9200±0.0422	0.8698±0.1175‡	0.8166±0.1305‡	0.9458±0.0271‡	0.0529±0.0300‡	0.1002±0.0769‡
	Refined WingsNet	0.8504±0.0651	0.9177±0.0409	0.9050±0.1032	0.8621±0.1230	0.9331±0.0303	0.0668±0.0347	0.0927±0.0752
	NaviAirway [101]*	0.8348±0.0335‡	0.9096±0.0200‡	0.8734±0.0715‡	0.8099±0.0950‡	0.8672±0.0406‡	0.1500±0.0536‡	0.0413±0.0304‡
	Refined NaviAirway	0.8246±0.0354	0.9034±0.0213	0.9064±0.0595	0.8590±0.0857	0.8516±0.0432	0.1716±0.0584	0.0357±0.0261
Fibrosis	Wang et al. [91]†	0.6979±0.0647‡	0.8203±0.0462‡	0.6961±0.0924‡	0.6261±0.1117‡	0.7468±0.0773‡	0.32727±0.1445‡	0.0822±0.0388
	Refined Wang et al.	0.6528±0.0925	0.7859±0.0722	0.8007±0.0847	0.7399±0.1042	0.6736±0.0951	0.4953±0.2456	0.0456±0.0277
	WingsNet [87]*	0.8052±0.0539‡	0.8910±0.0340‡	0.6951±0.0977‡	0.6198±0.1175‡	0.9505±0.0116‡	0.0438±0.0113‡	0.1595±0.0576‡
	Refined WingsNet	0.8035±0.0515	0.8901±0.0326	0.7318±0.1012‡	0.6638±0.1221‡	0.9341±0.0159	0.0604±0.0163	0.1480±0.0565
	NaviAirway [101]*	0.8074±0.0533‡	0.8924±0.0340‡	0.5994±0.1440‡	0.5148±0.1490‡	0.9247±0.0165‡	0.0714±0.0188‡	0.1345±0.0645‡
	Refined NaviAirway	0.8061±0.0514	0.8917±0.0329	0.6534±0.1513	0.5725±0.1617	0.9090±0.0189	0.0891±0.0226	0.1215±0.0657
COVID	Wang et al. [91]†	0.7433±0.1010	0.8481±0.0798	0.8487±0.1320‡	0.7993±0.1409‡	0.7749±0.0736‡	0.2843±0.1500‡	0.0541±0.1035‡
	Refined Wang et al.	0.7408±0.0566	0.8498±0.0392	0.9538±0.0435	0.9316±0.0603	0.7462±0.0562	0.3453±0.1152	0.0098±0.0106
	WingsNet [87]*	0.9176±0.0363‡	0.9566±0.0209‡	0.9073±0.0647‡	0.8682±0.0831‡	0.9683±0.0204‡	0.0311±0.0203‡	0.0544±0.0283‡
	Refined WingsNet	0.8971±0.0345	0.9454±0.0200	0.9571±0.0575	0.9409±0.0778	0.9424±0.0266	0.0588±0.0297	0.0510±0.0259
	NaviAirway [101]*	0.8861±0.0298‡	0.9394±0.0171‡	0.8729±0.0792‡	0.8060±0.1034‡	0.9100±0.0245‡	0.0970±0.0316‡	0.0286±0.0239‡
	Refined NaviAirway	0.8725±0.0315	0.9316±0.0186	0.9153±0.0724	0.8732±0.0997	0.8904±0.0295	0.1218±0.0407	0.0223±0.0214

* refers to results obtained from open-source implementations with model weights provided.
† refers to reproduced results.
‡ represents statistical significance (with Wilcoxon signed-rank test p<0.01) compared with the refined results.

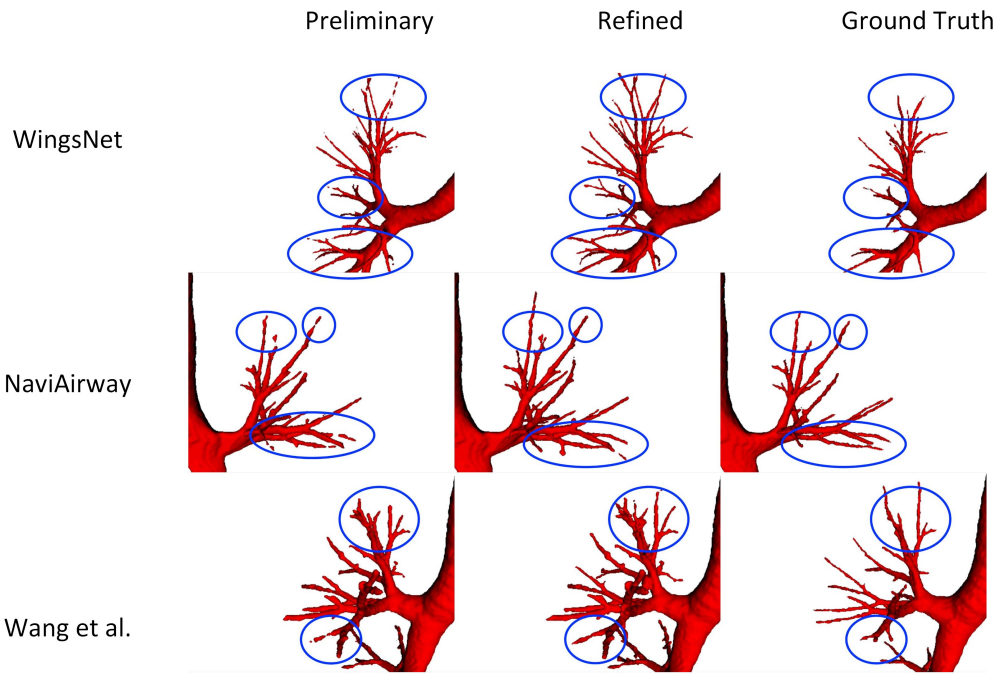


Figure 3.9: Close-up of bronchioles in CASE02 from the EXACT'09 test set. Top to bottom rows show results from WingsNet, NaviAirway, and Wang et al. respectively.

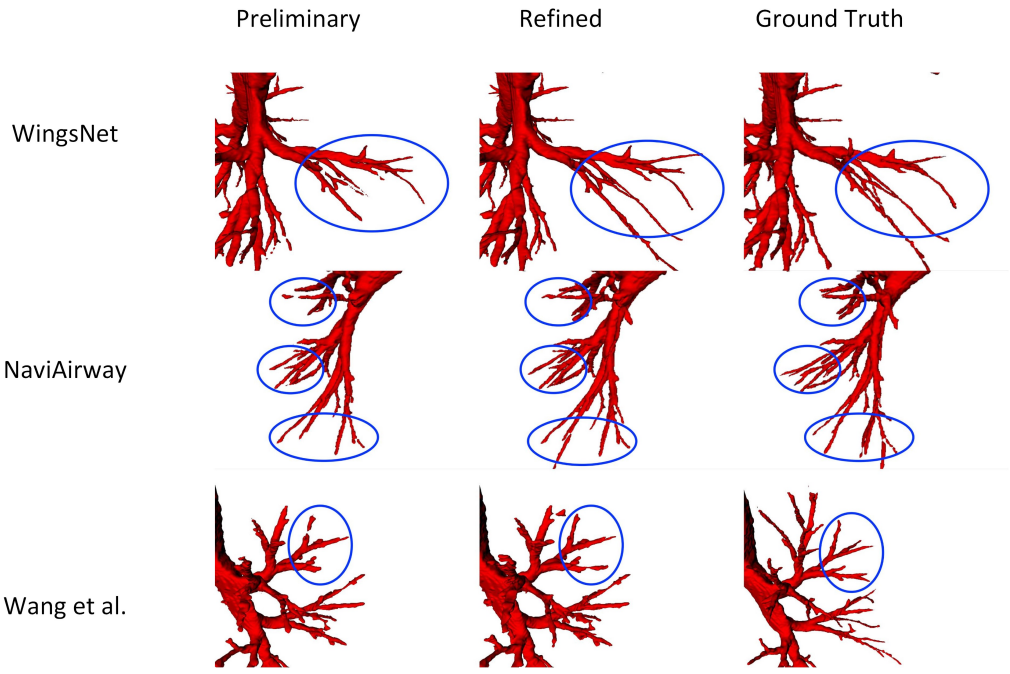


Figure 3.10: Zoomed-in images of bronchioles in CASE 162_01 from the fibrosis test set. The images are organized with results from WingsNet, NaviAirway, and Wang et al., in top-to-bottom order.

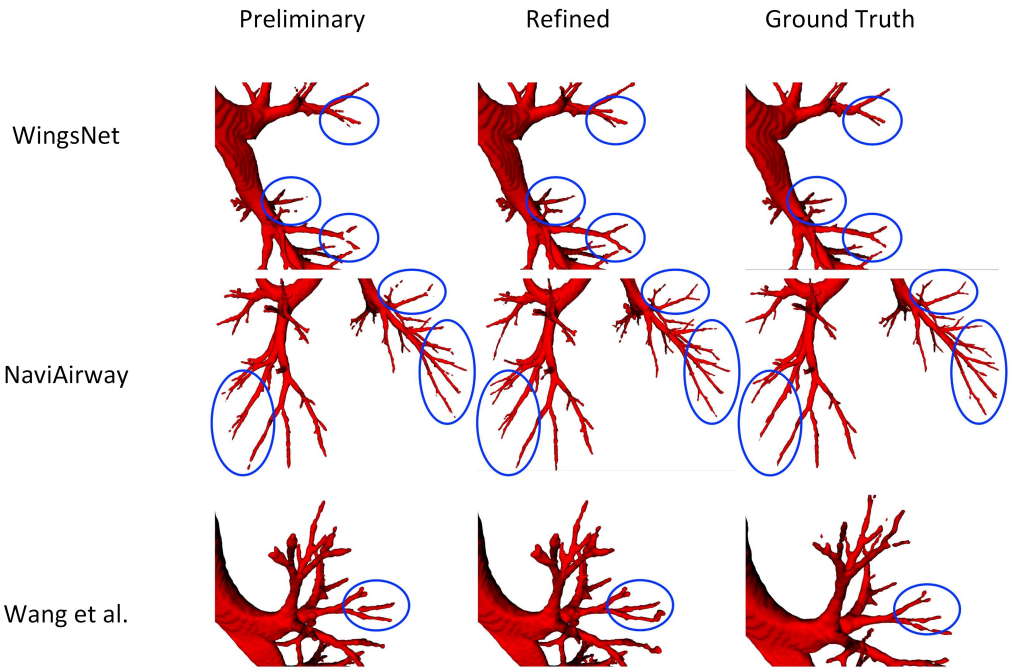


Figure 3.11: A zoomed-in view of bronchioles in CASE RM451 from the COVID test set, with results from WingsNet, NaviAirway, and Wang et al. presented from top to bottom, respectively.

3.2.6 Ablation Study

Table 3.11: Ablation study of the proposed model.

Model	IoU	Dice	DLR	DBR	Precision	Leakages	AMR
U-Net+PatchGAN+dilation(BL)	0.8274±0.0563	0.9044±0.0355	0.7936±0.1303	0.7088±0.1557	0.8866±0.0405	0.1215±0.0504	0.0730±0.0622
BL+clDice	0.8325±0.0465	0.9079±0.0294	0.8059±0.1285	0.7324±0.1525	0.8895±0.0313	0.1178±0.0400	0.0693±0.0608
BL+ccf	0.7133±0.0854	0.8295±0.0631	0.8526±0.1148	0.7947±0.1474	0.7441±0.0880	0.3519±0.2151	**0.0516±0.0541**
BL+ccf+multi-scale	0.7701±0.0522	0.8691±0.0343	0.8420±0.1277	0.7788±0.1542	0.8101±0.0484	0.2266±0.0761	0.0573±0.0600
BL+clDice+multi-scale	**0.8342±0.0670**	**0.9080±0.0427**	0.8352±0.1156	0.7671±0.1421	**0.8969±0.0478**	**0.1082±0.0578**	0.0781±0.0565
BL+ccf+ clDice + multi-scale	0.8150±0.0519	0.8971±0.0330	0.8828±0.1012	0.8337±0.1263	0.8556±0.0393	0.1620±0.0522	0.0544±0.0525
BL(ViT) + ccf + cl_dice + multi-scale	0.8133±0.0518	0.8961±0.0334	**0.8902±0.0967**	**0.8439±0.1261**	0.8600±0.0401	0.1549±0.0539	0.0623±0.0505

An ablation study was carried out to evaluate the individual contributions of each component in our refinement model. The outcomes are presented in Table 3.11, where the baseline model, a 3-D U-Net combined with PatchGAN and dilation (BL), is compared against six variants: BL + clDIce, BL + CCF, BL + CCF with multi-scale supervision, BL + clDice with multi-scale supervision, BL + clDice + CCF with multi-scale supervision, and BL (where PatchGAN is replaced with ViT) + clDice + CCF with multi-scale supervision. The integration of cl-Dice resulted in a modest enhancement of approximately 1.0% across all seven metrics. Substituting cl-Dice for CCF, however, led to a decrease in IoU, Dice, and leakage scores, while the other four metrics improved markedly by an average of over 5.0%. Adding multi-scale supervision to models utilizing CCF and cl-Dice bolstered their performance across all metrics. The optimal results in DLR (0.8902) and DBR (0.8439) were attained by integrating CCF, cl-Dice, and multi-scale supervision with ViT in our model.

3.2.7 Discussion and Conclusion

In this study, we introduced an innovative adversarial-based refinement model employing *tanh* as the final activation function. The network was trained with objective functions clDice and CCF, emphasizing airway continuity. This model adeptly corrects segmentation breakages and integrates missing branches in preliminary outputs from networks like U-Net.

Evaluated on three distinct datasets—the BAS dataset and two in-house datasets each consisting of 25 cases of pulmonary fibrosis and COVID-19—our refined method showed superior performance compared to the preliminary approach, as indicated in Tables 3.8 and 3.9. Notably, it significantly minimized false negatives and enhanced the length and number of detected

branches across all test datasets. We highlight instances in Figures 3.7 through 3.11, using blue circles, where our refinement successfully identifies breakages and bronchioles missed in the preliminary segmentation. The completeness of airway segmentation is essential for clinical application, offering biomarkers for assessing lung disease severity such as traction bronchiectasis [9] and airway tapering [122]. Our model's proficiency in preserving local connectivity positions it as a valuable clinical tool, particularly since disconnected branches are typically removed in airway evaluation processes.

However, our method has limitations. It relies on preliminary segmentation produced by a specific small U-Net, leading to greater enhancements when refining segmentation from the same network compared to those from other models like Wang et al., WingsNet, and NaviAirway. The quality of initial results also influences the refinement's efficacy, as better preliminary results generally lead to more effective refinement (Table 3.10). To address this, future work will involve training the model on preliminary segmentations generated by randomly masking ground truth. Additionally, while our approach ensures airway continuity, it may occasionally over-segment peripheral bronchioles, resulting in higher leakage rates, particularly noticeable in the fibrosis dataset. In some cases, our predictions identified airway branches not present in the ground truth, partially explaining the marginal decreases in IoU and Dice scores alongside increased leakage post-refinement.

In summary, this research demonstrates that patch-scale discriminators can significantly enhance airway segmentation by improving connectivity and reducing false negatives. Furthermore, this refinement pipeline holds potential for adaptation to other models and segmentation tasks in medical imaging.

Chapter 4

Unsupervised Learning in Medical AI

Tissue segmentation plays a pivotal role in pathological examinations, whereas manual delineation proves to be excessively burdensome. To facilitate this time-consuming and subjective task, researchers have developed automated methods for segmenting structures in pathological images. Recent trends in tissue segmentation research have been dominated by automated machine learning and deep learning approaches. However, most of these methods are supervised and require extensive training samples, where pixel-wise annotations are expensive and sometimes unattainable. This study introduces a novel unsupervised learning approach, integrating an end-to-end deep mixture model with a constrained indicator for precise semantic tissue segmentation. This constraint is designed to centralize the components of deep mixture models during the optimization function calculation, thereby significantly reducing the common issues of redundant or empty classes in current unsupervised learning methods. Validated on both public and in-house datasets, our proposed deep-constrained Gaussian network demonstrates significantly better performance (confirmed by the Wilcoxon signed-rank test) with average Dice scores of 0.737 and 0.735, respectively.

4.1 Introduction

Given an image, a segmentation algorithm's objective is to assign labels to pixels based on their feature representations. Tissue segmentation plays a crucial role in automated pathological examination, diagnosis, and prognosis, yet manual delineation is time-consuming, laborious, and lacks reproducibility. To mitigate these challenges, researchers have delved into traditional methods for automatic organ or structure segmentation, including watershed [123], contour detection [124], clustering [125, 126], and random field [127], among others. These approaches, however, are often unreliable and heavily dependent on thresholds or preset parameters. Recent advancements in machine learning and deep learning have led to significant progress in computational pathology [128–131]. For instance, Mahbod et al. [131] introduced a progressive sequential causal GAN for synthesizing late gadolinium enhancement imaging to enhance the segmentation of diagnosis-related structures. Liu et al. [132] combined CycleGAN with an adaptive Mask RCNN for unsupervised nuclei segmentation in histopathology images, leveraging knowledge from fluorescence microscopy images. Nonetheless, most learning-based methods are either fully supervised, necessitating manual labelling, or unsupervised, requiring intricate training procedures. Particularly, complex pathological structures significantly increase the difficulty of pixel-level annotation, underscoring the need for developing segmentation methods that demand minimal or no manual annotation.

One approach to address limited annotations is (deep) semi-supervised learning, which constructs models using minimal annotations or existing knowledge about the targets. Self-training, a prevalent technique, involves training the model with a limited number of annotated labels and subsequently refining it using pseudo labels generated by the model itself. For example, Liang et al. [133] developed an iterative learning scheme for segmenting gastric tumours using a partially labelled dataset. In addition to self-training, incorporating prior knowledge from conventional methods or empirical constraints, such as target labels, is another strategy for training networks. This can involve the use of coarse masks from image processing algorithms, pre-trained weights from related datasets, or image-level annotations from domain experts. Hu et al. [134] employed activation maps for detecting COVID-19 infections without

requiring pixel-level annotation. Atlason et al. [135] utilized coarse masks from an automated labelling system as attention maps to guide the network's focus on specific regions.

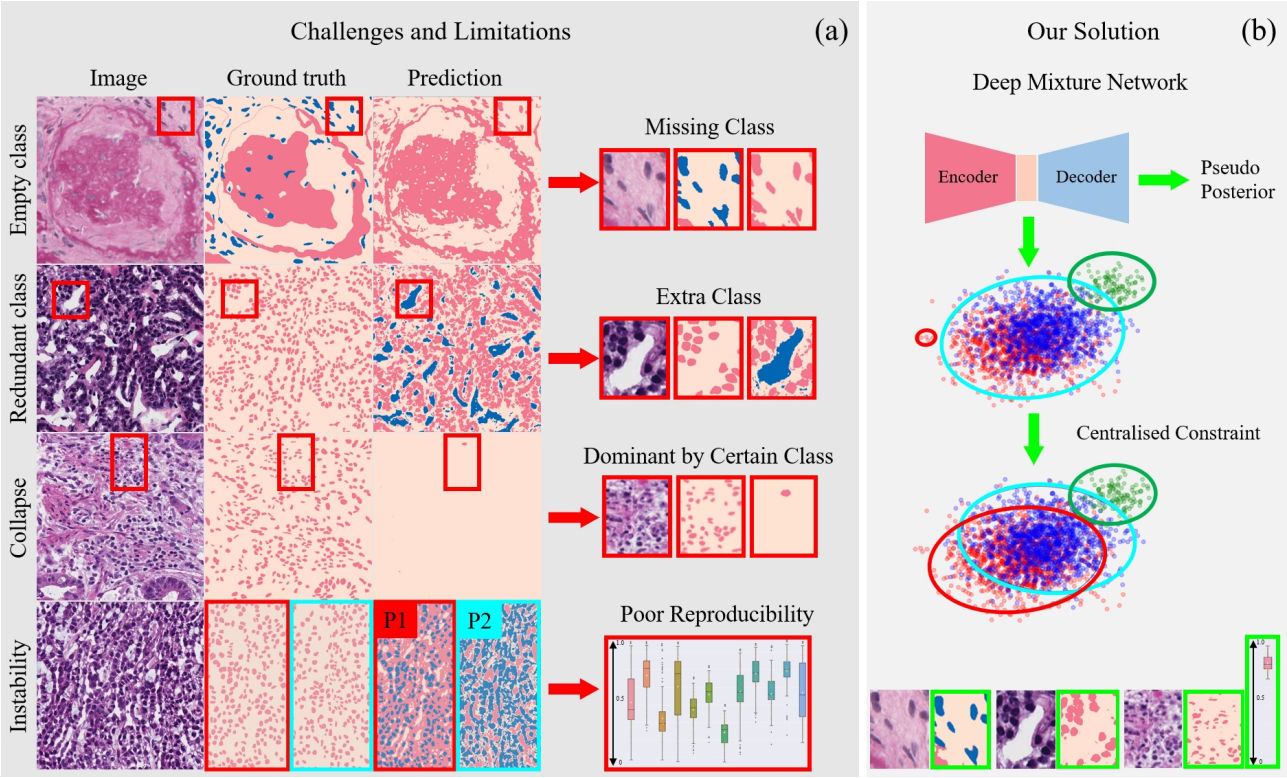


Figure 4.1: Challenges in Unsupervised Tissue Segmentation (a) and Our Solutions (b). (a) Displays common issues: empty classes (first row), redundant classes (second row), collapse (third row), and instability (fourth row). Red boxes mark three sections of the raw image, ground truth, and predictions using current methods. P1 and P2 are first and second predictions from experiments (last row). (b) Introduces our centralized constraint deep mixture network for unsupervised segmentation. Our model's effectiveness is shown in green boxes (last row), demonstrating improvements in handling empty classes, redundancies, collapse, and instability. Dice scores for all examples range between [0, 1].

Another viable approach is (deep) unsupervised learning, which generates general semantic predictions like 'background' and 'foreground' without relying on manual annotations. Kanezaki et al. [136] utilized Simple Linear Iterative Clustering [137] to achieve super-pixel level segmentation in combination with convolutional neural networks for segmenting natural images. Shen et al. [138] introduced a coupled "deep-image-prior" module for background and foreground segmentation. However, the efficacy of these methods for pathological images remains uncertain, as they primarily focus on natural images. Moreover, variations in image quality, such as differences in brightness, contrast, noise, and shading levels in pathological images, may challenge the generalizability of models originally developed for natural images. The random

initializations inherent in some unsupervised learning methods might lead to inconsistent performance and limited reproducibility. Specifically, several degenerative issues exist, including empty class, redundant class, collapse, and instability (see Fig. 4.1). The empty class problem arises when a model incorrectly merges one class with another, leading to fewer predicted classes than predefined (e.g., predicting only two classes despite defining three, as shown in the first row of 4.1(a)). Redundant class refers to the need for an extra class to handle challenging samples, characterized by pixel intensities that diverge significantly from the mean intensity of their true/false class (e.g., the white regions in the second row of Fig. 4.1(a)). Collapse issues occur when one class dominates the majority of an image's predictions, with other classes appearing sporadically (as depicted in the third row of Fig. 1(a)). Instability denotes fluctuating performance across repeated training iterations (illustrated in the fourth row of 4.1 (a)).

To overcome the aforementioned limitations, this study introduces a novel unsupervised approach, the Deep Constrained Gaussian Network (DCGN), which combines a deep neural network with log-likelihood maximisation and a centralised constraint (illustrated in 4.1(b)). DCGN differs from previous methods by processing raw images directly to generate pixel-wise predictions for tissue structures. A key feature of DCGN is its centralised constraint, designed to enhance the model's robustness and performance by narrowing the gap between the estimated mean values of the components and the actual data centroids. Our comprehensive experimental studies were conducted on the multicentre open-access MoNuSeg dataset from the TCGA archive and our own in-house dataset. Repeated experiments were performed to assess the stability of various approaches. The proposed DCGN method sets a new benchmark in unsupervised segmentation for pathological images, achieving Dice scores of 0.743 and 0.737 on MoNuSeg and our in-house dataset, respectively. This performance significantly surpasses all comparison models (Wilcoxon signed-rank test p-value <0.001). The rest of this work is organized as follows: Section 4.3 summarizes related studies on unsupervised segmentation, Section 4.4 details the proposed method, Section 4.5 describes the experimental settings including dataset and training parameters, and Sections 4.6 and 4.7 present the discussion and conclusion of the study.

4.2 Related Works

4.2.1 Conventional Unsupervised Segmentation

Unsupervised segmentation is commonly conceptualized as a clustering task. To process a three-channel RGB image, clustering algorithms initially flatten the 3D array into a 2D vector. Subsequently, each pixel group, encompassing the R, G, and B channels, is treated as a multi-dimensional sample for clustering. Various methods, such as graph/normalised cuts [139, 140], Markov random field [140], minibatch K-means [141], Gaussian mixture model (GMM) [142], and mean shift [143], have been extensively employed in medical image analysis tasks including registration [144], lesion detection [145], and segmentation [142]. Beyond clustering, differentiating between feature representations also facilitates the segmentation of regions of interest in images. For instance, Fan et al. [146] utilized hierarchical image matting for vessel segmentation in fundus images, while Tosun et al. [147] introduced an object-oriented approach with a homogeneity measurement for segmenting biopsy images. However, these methods have weak generalization abilities and low accuracy.

4.2.2 Deep Clustering and Mutual Information

Recent studies of unsupervised learning aim to combine conventional clustering methods with deep neural networks [148–150]. Specifically, these methods use clustering-based objective functions to train a neural network. For instance, DeepCluster [148] jointly updated the parameters of the neural networks and clustering during the training, and used pseudo labels to calculate objective functions. Kim et al. [151] proposed a spatial constraint to the softmax cross-entropy loss (given by pseudo labels and predictions) to keep the spatial continuity of semantic predictions. Wellmann et al. [148] integrated domain knowledge as probabilistic relations and proposed a deep conditional GMM. However, using pseudo labels for training is prone to weak solutions, such as empty clusters, and trivial parametrisation [148].

Maximizing the mutual information of paired predictions is effective [152]. To further alle-

viate degenerative issues, Invariant Information Clustering (IIC) [153] modified co-clustering approaches and proposed mutual information-based objective functions between paired samples to train a segmentation model. Given a pair of variables X, Y and their marginal distribution $p(x)$ and $p(y)$, the mutual information between X and Y, jointly distributed according to $p(x, y)$, is defined as

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{4.1}$$

IIC generated paired images by randomised rotation to assist the network in learning the invariant information and textual representations. More generally, IIC aimed to find common parts of paired samples while ignoring the redundant ones. However, it still suffers from degenerative issues and unstable performance.

4.2.3 Deep Generative Models and Log Likelihood

Deep generative models are designed to learn image representations by reconstructing input images through various generative models, such as generative adversarial networks (GAN), variational auto-encoders (VAE), and encoder-decoders. These representations are then utilized to generate semantic predictions or to compute objective functions [154]. Chen et al. [155] leveraged redrawing concepts to segment foreground and background samples, while Gandelsman et al. [156] introduced a double Deep Image Prior (DIP) method for compositing images into background and foreground samples. However, these techniques are typically restricted to segmenting a limited number of classes, potentially leading to computational inefficiency in multi-class predictions.

An innovative approach is the integration of deep neural networks with Gaussian Mixture Models (GMM). Zong et al. [157] developed a Deep Auto-encoder Gaussian Mixture Model (DAGMM), incorporating GMM into the low-dimensional feature representations of an auto-encoder for unsupervised anomaly detection. Oord et al. [158] employed GMM in the upper layers of hierarchical structures for unsupervised classification. Building on these concepts, Zanjani et al. [159] adapted DGMM for segmentation by classifying each pixel for stain normalization,

proposing three new schemes: GAN-based, VAE-based, and Deep Convolutional Gaussian Mixture Model (DCGMM) approaches. The VAE-based method and DCGMM showed promising applicability in segmentation. Specifically, the VAE-based approach utilized log-likelihood loss and Kullback-Leibler (KL) divergence loss to evaluate raw data reconstruction performance and the relationship between latent variables and the prior distribution, respectively. DCGMM optimized the network by maximizing the log-likelihood objective function. Nevertheless, most of these methods primarily merge expectation maximization with deep neural networks, without specifically tackling the prevalent challenges in unsupervised tissue segmentation.

4.3 Methodology

4.3.1 Overview

To address the limitations inherent in current unsupervised segmentation approaches, we outline essential characteristics for an effective model:

1. The model should exhibit robust reproducibility during the training and validation phases.
2. The model should be lightweight without complex pre-processing or post-processing steps.
3. The model must be capable of mitigating degenerative issues, such as the empty clusters issue.

In light of these criteria, we introduce the Deep Constrained Gaussian Network (DCGN) for unsupervised segmentation of pathological tissue images.

4.3.2 Deep Constrained Gaussian Network

In the domain of biomedical image segmentation, particularly with pathological images, semantic labels are closely linked to colour representations, unlike in natural images. This suggests the potential for effectively integrating a mixture model with a deep neural network for unsupervised segmentation.

Let ω represents the learnable parameters of a deep neural network, and J denotes the objective function. In fully supervised learning, ω is optimized by minimizing J, calculated based on the discrepancies between ground truth labels and model predictions. Therefore, the crux of unsupervised segmentation lies in identifying an optimal objective function that can train deep neural networks without reliance on annotated data (ground truth labels). Alongside maximizing mutual information between paired samples in Eq. (3.1), the maximization of log-likelihood can be integrated into a gradient descent training framework by minimizing the negative log-likelihood.

The proposed Deep Constrained Gaussian Network (DCGN) is composed of a feature extractor, a decoder, and a log-likelihood estimation module. Unlike the precise objective functions in supervised learning that compute the discrepancy between ground truths and predictions, log-likelihood maximization serves as a biased estimator, providing only an approximate 'direction' towards the global optimum [160, 161]. Consequently, we believe that complex architectures are more likely to suffer from overfitting and local optima in the absence of a robust supervised optimization function. To design a light architecture, MobileNet-V2 [162] is employed as the feature extractor, followed by a decoder consisting of Upsampling layers, Convolution layers, Batch normalization layers, and ReLU activations. To align the network's predictions with the pseudo posterior of the latent variable Z in the mixture model, a differentiable softmax layer is applied to the output, yielding a $[W, H, K]$-shaped prediction ϕ (where W and H represent width and height, and K denotes the number of classes). Given input images I with K classes, the network Φ aims to generate semantic probability maps ϕ, which are treated as the pseudo posterior y in traditional GMM, such that

$$\gamma \approx \phi = \Phi(I, \omega) \in \mathbb{R}^{W \times H \times K}. \tag{4.2}$$

Under this framework, the E-step can be performed via forward propagation through the neural network, while the M-step involves optimizing the likelihood function through gradient descent.

Given the pseudo posterior γ_{ik}, the likelihood $\mathcal{L}(\theta|\theta^{(t)})$ of the multivariate Gaussian Mixture Model (GMM) can be estimated as follows:

$$\mathcal{L}(\Theta|\Theta^{(t)}) = \sum_{k=1}^{K}\sum_{i=1}^{N} \gamma_{ik} \left[\log \alpha_k - \frac{D}{2}\log(2\pi) - \log \gamma_{ik} - \frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(X_i - \mu_k)^T \Sigma_k^{-1}(X_i - \mu_k)\right], \tag{4.3}$$

where $\frac{D}{2}\log(2\pi)$ is a constant and thus can be omitted, D represents the dimension of each sample (which is 3 for a flattened image), N is the total number of samples (pixel groups) in the image, and α_k is the weight of the k-th Gaussian mixture, satisfying $\sum_k \alpha_k = 1$. Consequently, by combining Eqs. (3.2) and (3.3), the network Φ can be optimized by minimizing the negative

log-likelihood:

$$\omega^* = \arg\min_{\omega}[-\mathcal{L}(\Phi(X;\omega))]. \tag{4.4}$$

where ω is the parameters of Φ, $\mathcal{L}$ is the loglikelihood. A significant challenge in existing deep Gaussian models is the redundant class issue, caused by small inter-class and substantial intra-class variations. This results in the model inaccurately assigning identical (or different) labels to samples from different (or the same) classes. Outliers or hard samples can also skew the optimization function's estimation, leading to trapping in local optima or unstable training. Moreover, instability is a prevalent issue in existing unsupervised learning algorithms, where randomized initializations necessitate multiple training iterations to achieve optimal performance.

In this work, we introduce a centralised constraint for the log-likelihood objective function to mitigate the degenerative problems inherent in deep Gaussian networks. The objective function of such networks is computed using the estimated parameters Θ and the pseudo posterior y. However, the variance in batch data often complicates the accurate determination of the real parameters μ_{real}. To elucidate our proposed centralised constraint, we present two simplified examples in Fig. 4.2. Initially, we consider a single-class sample group conforming to a Gaussian distribution, as illustrated in Fig. 4.2(a). Given a batch of data X, let μ_{est} be the estimated mean of the mixture model, μ_{obs} the observed mean of minibatch data X, and μ_{real} the actual (ideal) mean of the mixture model. The centralised constraint aims to subtly nudge μ_{est} closer to μ_{obs}, though μ_{obs} is not identical to μ_{est} as it represents the mean of minibatch samples.

For multi-class samples, this centralised constraint can also mitigate the impact of minor inter-class differences, as depicted in Fig. 4.2(b). Assume two classes a and b, with a' and b' representing the estimated classes. The model may erroneously classify the majority of samples from both classes a and b as class a', while outliers from class b are assigned to class b'. Such misclassification is particularly problematic when dealing with samples exhibiting slight inter-class variances.

Consequently, we devise a centralised constraint Δ to guide the estimated mean μ_{est} towards

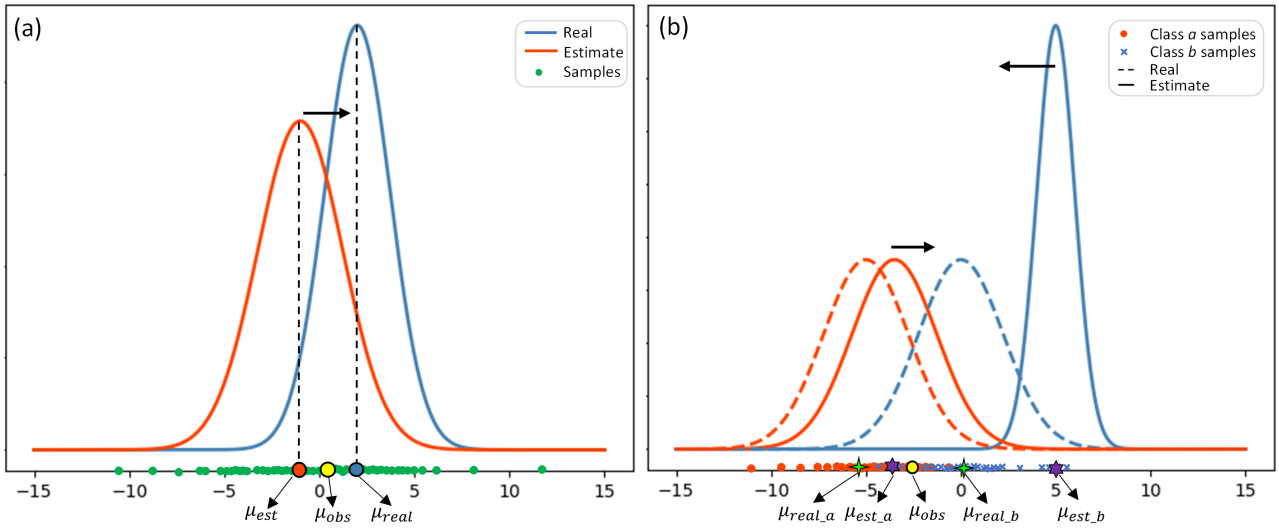


Figure 4.2: Variation in Estimated Parameters: (a) Shows a normal distribution for single-class samples. (b) Depicts a mixture model for multi-class samples with two classes (k=2). Note that μ_{est} represents the estimated mean of the mixture model, μ_{obs} is the observed mean of the minibatch data X, and μ_{real} is the ideal mean value of the mixture model.

μ_{obs}, taking into account the diversity of $\bar{X}$:

$$\Delta = \frac{|\mu_{est} - \bar{X}|}{\sigma_X^2}. \tag{4.5}$$

In scenarios involving challenging samples with minimal inter-class variances, the observed variance is relatively low, yielding a substantial constraint value. This constraint compels the model to realign the estimated mean closer to the observed mean, thereby diminishing degenerative issues. Conversely, for "straightforward" samples (characterized by significant inter-class variances), the observed variance is higher, resulting in a minimal constraint on the objective function that scarcely influences parameter estimation.

Based on the centralised constraint Δ, the objective function $\mathcal{L}_C$ for our DCGN is defined as

$$\mathcal{L}_C = \mathcal{L}(\Theta|\Theta^{(t)}) - \lambda \sum_{k=1}^{K} \sum_{c=1}^{C} \frac{\left|\mu_k^{(t)} - \bar{X}_c\right|^2}{\sigma_c^2}, \tag{4.6}$$

where C denotes the dimension of the input samples (e.g., $C = 3$ for RGB images), σ_c^2 is the variance of minibatch samples in channel c, and $\bar{X}_c$ represents the mean value of minibatch samples in channel c. The objective function $\mathcal{L}_C$ incurs a penalty when the estimated μ_k

deviates significantly from the observed mean μ_{obs}, thereby reducing the influence of outliers or hard samples on the function and enhancing the stability and effectiveness of the training process.

Letting the constraint weight be λ, the centralised mixture parameters can be derived by computing the partial derivatives of μ_k, Σ_k, and α_k in Eq. (3.8):

$$\gamma_k^{(t+1)} = \phi(X, \omega^{(t)}) \tag{4.7}$$

$$\mu_k^{(t+1)} = \begin{cases} \frac{\sum_{i=1}^{N} \gamma_{ik}^{(t+1)} X_i - \lambda \sum_{c=1}^{C} \frac{\Sigma_k^{(t)}}{\sigma_c^2}}{\sum_{i=1}^{N} \gamma_{ik}^{(t+1)}}, & \mu_k \geq \bar{X}_c \\ \frac{\sum_{i=1}^{N} \gamma_{ik}^{(t+1)} X_i + \lambda \sum_{c=1}^{C} \frac{\Sigma_k^{(t)}}{\sigma_c^2}}{\sum_{i=1}^{N} \gamma_{ik}^{(t+1)}}, & \mu_k < \bar{X}_c \end{cases} \tag{4.8}$$

$$\alpha_k^{(t)} = \frac{\sum_{i=1}^{N} \gamma_{ik}^{(t+1)}}{N} \tag{4.9}$$

$$\Sigma_k^{(t)} = \frac{\sum_{i=1}^{N} \gamma_{ik}^{(t+1)} (X_i - \mu_k^{(t+1)})^T (X_i - \mu_k^{(t+1)})}{\sum_{i=1}^{N} \gamma_{ik}^{(t+1)}} \tag{4.10}$$

Note that in Eq. (3.8), the calculation of $\mu_k^{(t+1)}$ relies on $\Sigma_k^{(t)}$; thus, an initialisation of Σ_k is required prior to the training. In this study, we employed a random initialisation from a uniform distribution. The pseudo-code of the entire training procedure for DCGN is shown in Algorithm 2.

Algorithm 2 Pseudo-code for training DCGN

Input: images $X \in \mathbb{R}^{W \times H \times 3}$
Output: trained network parameters ω, semantic prediction Y

randomly initialize $\Sigma_k^{(0)}$, network parameters $\omega^{(0)}$
for t in iterations **do**
  $Y^{(t)} = \phi(X, \omega^{(t)}) \in \mathbb{R}^{W \times H \times K}$
  update $\mu_k^{(t+1)}, \alpha_k^{(t)}$ with $Y_k^{(t)}, \Sigma_k^{(t)}$
  update $\Sigma_k^{(t+1)}$ with $Y_k^{(t)}, \mu_k^{(t+1)}$
  Compute L_c through $\mu_k^{(t+1)}, \Sigma_k^{(t+1)}, \alpha_k^{(t+1)}$
  update ω by $\arg\min_\omega[-L_c(\omega^{(t)})]$
end for

4.3.3 Preprocessing

Each input X was pre-processed by the max-min normalisation for RGB channels,

$$X'_c = \frac{X_c - \max(X_c)}{\max(X_c) - \min(X_c)} \tag{4.11}$$

where X_c is the channel c of the input image X.

4.4 Experiments

This section delineates the experimental setup, including datasets, evaluation metrics, implementation specifics, and results. The effectiveness of the proposed Deep Constrained Gaussian Network (DCGN) is evaluated using a public dataset from The Cancer Genome Atlas (TCGA) repository (MoNuSeg) and our in-house Renal Biopsy Image (RBI) dataset.

4.4.1 Datasets and Training Strategies

MoNuSeg: The MoNuSeg dataset[1] comprises 44 pathological tissue images containing 28,846 manually annotated nuclear boundaries. Extracted from distinct whole slide images (WSIs) scanned at 40× magnification from the TCGA repository, these 1000 × 1000 pixel images represent nine different organs sourced from 44 individuals. Both stromal and epithelial nuclei were manually labeled using Aperio ImageScope, ensuring accuracy in annotation. The diversity in tissue sections enriches the dataset, providing a comprehensive platform for evaluation. The details of the MoNuSeg dataset are presented in Table 4.1.

RBI: The Renal Biopsy Image (RBI) dataset comprises over 10,000 image patches extracted from 400 whole slide images with biopsy-proven results, collated from the National Clinical Research Centre of Kidney Diseases at Jinling Hospital. In adherence to the principles of the Declaration of Helsinki [163], all data were de-identified. Each image in the dataset was

[1] https://monuseg.grand-challenge.org/

Table 4.1: Composition of the MoNuSeg dataset.

Subset	Nuclei	Images	Anatomical Details
Training	21623	30	6 breast, 6 liver, 6 kidney, 6 prostate, 2 bladder, 2 colon, 2 stomach
Testing	7223	14	2 breast, 3 kidney, 2 prostate, 2 bladder, 1 colon, 2 lung, 2 brain

resized to a uniform dimension of 512×512 pixels. For training and validation, we randomly selected 577 and 20 images, respectively. Notably, the glomerular structures in these images were annotated by experienced pathologists with over 20 years of expertise. It is important to mention that the training and validation sets were sourced from distinct whole slide images.

Training Strategies: The parameters of the encoder were initialized using pre-trained weights from ImageNet to enhance feature extraction capabilities. The parameters of the decoder were initialized following the He-normal initialization method. To augment the dataset, we applied randomized transformations including hue adjustment (delta = 0.12), saturation modification (saturation factor ranging from 0.5 to 1.5), and randomized flips (up/down and left/right). The training was performed on an NVIDIA RTX 3090 GPU over 200 epochs, starting with a learning rate of $5e^{-5}$ and implementing a decay rate of 0.98 per epoch.

4.4.2 Experimental Details

Comparisons. To assess the efficacy of the DCGN, we compared it with a range of both deep learning-based and traditional unsupervised segmentation methods. These include mini-batch K-Means (referred to as mKMeans), Gaussian Mixture Model (GMM), Invariant Information Clustering (IIC) [153], Double Deep Image Prior (Double DIP) [138], Deep Clustering via Adaptive GMM Modelling (DCAGMM) [164], Deep Image Clustering (DIC) [165], Kim's approach [151], Kanezaki's method [136], and Deep Convolutional Gaussian Mixture Model (DCGMM) [159]. Notably, DCAGMM was adapted for this study by incorporating its distance-based constraints into the original DCGMM framework, which was initially intended for image classification. The open-source implementations of these comparison methods are available on

Github. For an equitable assessment, we modified the network architecture of the DCGMM to align with our DCGN. Besides these unsupervised approaches, we also benchmarked against a fully supervised U-Net adapted for the cell segmentation task. This modified version of U-Net includes additional batch normalization and dropout layers, diverging from the original U-Net model [106].

Cell Segmentation on MoNuSeg. A common challenge in many existing unsupervised learning methods is the variability in performance due to random initializations. To address this, our study involved repeated experiments to evaluate the stability and reproducibility of all the comparison algorithms. Each method was trained for 150 epochs, and this process was replicated 10 times without altering any parameters or training samples. The 'upper bound' performance is characterized as the best outcome from these 10 iterations. Although cell segmentation inherently is a binary classification task, we assessed all compared methods with varying numbers of classes (k=2 or 3) to demonstrate their maximum potential performance. Additionally, as a benchmark in supervised learning, a fully supervised U-Net model was trained and included in the comparative analysis.

Glomeruli Decomposition on RBI. Beyond binary segmentation effectiveness, we also conducted a glomeruli decomposition task using the RBI dataset. In this analysis, the glomerular structures were segregated into three distinct parts (k=3), including (1) the mesangial matrix and basement membrane; (2) intra-glomerular cells such as mesangial, endothelial, and podocytes, along with macula densa; and (3) other regions encompassing glomerular capillaries, Bowman's space, and exudates. It is important to note that the Double Deep Image Prior (Double DIP) method was excluded from this task, as it was originally designed solely for binary segmentation.

Degeneration Assessment. To explore degenerative issues, we conducted an analysis using the following methodology: (1) Analysis was based on 140 predictions on the MoNuSeg datasets and 100 predictions on the RBI datasets. (2) Predictions were obtained from repeated experiments, conducted 10 times for MoNuSeg and 5 times for RBI. (3) We assessed collapse on both MoNuSeg and RBI datasets. Collapse is defined as a scenario where a certain class dominates

the major region of an image, with a set threshold of 97%. (4) The presence of a redundant class was evaluated on the MoNuSeg dataset. This occurs when segmentation performance improves by adding an extra class without semantic meaning. (5) Empty class assessment was conducted on the RBI dataset, referring to instances where a certain class is missing or has an extremely low ratio (below 1%) in the prediction. (6) Instability was evaluated on both MoNuSeg and RBI datasets. It is considered significant when the standard deviation of the average performance among repeated experiments exceeds 8%.

Evaluation Metrics. In addition to the commonly utilized Dice coefficient score, we also reported pixel-wise precision and recall. For statistical evaluation of performance, we adopted the Wilcoxon signed-rank test to compare the evaluation results derived using DCGN with other comparative methods. A P-value of less than 0.05 (or 0.001) was considered indicative of significant (or highly significant) differences between the two compared methods. The Aggregated Jaccard Index (AJI) was applied to the MoNuSeg dataset to verify instance-level segmentation performance, as given by:

$$AJI = \frac{\sum_{i=1}^{N} G_i \cap P_i}{\sum_{i=1}^{N} G_i \cup P_i + \varepsilon}, \tag{4.12}$$

where i represents the number of cells, ε is a smoothing parameter, and G_i and P_i are the ground truth and prediction for the i-th cell, respectively. In the case of glomeruli segmentation, we applied normalised mutual information (NMI) to evaluate the mutual dependence between two samples, defined as:

$$NMI(Y, C) = \frac{2I(Y; C)}{H(Y) + H(C)}, \tag{4.13}$$

where Y and C denote the ground truth labels and predictions, respectively, I is the mutual information between Y and C, and $H(\cdot)$ represents the entropy. It is important to note that all ground truth labels were only used during the evaluation phase and were not revealed during the training process.

4.4.3 Experimental Results

Unsupervised Cell Segmentation on MoNuSeg. The performance of repeated experiments is presented in Table 4.2, indicating mean ± standard deviation, with upper-bound

results of each method in brackets. The results demonstrate that some unsupervised methods, originally developed for natural images, underperform on pathological images, showing a significantly lower average Dice score (3-39% lower) compared to our proposed DCGN, as seen in Figure 4.3 and Table 4.2. For instance, double DIP [138] recorded a mere 0.344 average Dice score in cell segmentation. In contrast, the conventional GMM (k=3) achieved a commendable 0.695 average Dice score, comparable to DCGMM's 0.707.

Table 4.2: Performance of the cell segmentation (MoNuSeg dataset).

Methods	Precision	Recall	Dice	AJI
mKMeans*	0.657±0.175 (0.679)*	0.792±0.174 (0.773)‡	0.678±0.094 (0.682)‡	0.305±0.140 (0.338)‡
GMM*	0.631±0.150 (0.664)‡	0.822±0.109 (0.819)	0.695±0.085 (0.717)‡	0.290±0.151 (0.319)‡
IIC*	0.467±0.092 (0.516)‡	0.725±0.121 (0.796)‡	0.560±0.087 (0.618)‡	0.056±0.030 (0.072)‡
Kim et al.*	0.575±0.249 (0.698)‡	0.824±0.189 (0.772)	0.606±0.171 (0.694)‡	0.220±0.176 (0.323)‡
Double DIP	0.221±0.051 (0.221)‡	0.820±0.109 (0.851)	0.344±0.067 (0.350)‡	0.013±0.006 (0.013)‡
Kanezaki et al.*	0.629±0.195 (0.725)‡	0.822±0.162 (0.783)	0.669±0.119 (0.727)‡	0.260±0.166 (0.351)‡
DCGMM*	0.693±0.135 (0.698)*	0.786±0.171 (0.801)‡	0.707±0.064 (0.719)‡	0.314±0.124 (0.345)‡
DIC*	0.511±0.249 (0.595)‡	**0.848±0.170 (0.832)***	0.571±0.165 (0.644)‡	0.147±0.169 (0.193)‡
DCAGMM	0.619±0.137 (0.691)‡	0.767±0.131 (0.763)‡	0.664±0.079 (0.706)‡	0.300±0.126 (0.365)‡
DCGN	0.685±0.113 (0.716)	0.834±0.115 (0.808)	0.737±0.043 (0.743)	0.352±0.113 (0.379)
U-Net†	**0.695±0.095 (0.740)**	**0.849±0.083 (0.848)**	**0.755±0.045 (0.782)**	**0.370±0.093 (0.436)***

* indicates models with a redundant class (k=3). † represents a fully supervised learning baseline using a modified U-Net. Bold values signify the top average performance among unsupervised methods, excluding the supervised U-Net. ** show significant or highly significant differences compared to DCGN, as per the Wilcoxon signed-rank test ($P<0.05$ or $P<0.01$). Results are presented as 'mean ± standard deviation (upper-bound results)'.

Statistical assessments using the Wilcoxon signed-rank test were performed on the results of 10 repeated experiments. Reviewing the upper bound of segmentation performance in Table 4.2, our proposed DCGN achieved the highest Dice score of 0.743 among unsupervised learning methods, followed by Kanezaki's method at 0.727 and DCGMM at 0.719. Additionally, DCGN recorded the best AJI score (0.379) among all unsupervised approaches. Moreover, DCGN significantly outperformed other unsupervised segmentation approaches in terms of both Dice coefficient and AJI scores ($P<0.001$). Notably, while DCGMM showed superior Precision to DCGN ($P=0.036$), it lagged in Recall, Dice, and AJI scores ($P<0.001$). DIC achieved the highest Recall but had relatively low Precision, indicating a prevalence of false-positive predictions. Double DIP, although achieving high Recall, recorded the lowest Precision and consequently a very low Dice score. For a clearer demonstration of the performance of competitive approaches (Dice>0.65), three images were randomly selected from the test set to visualize the upper-bound segmentation performance. It is important to note that in Fig 4.4, predictions of the

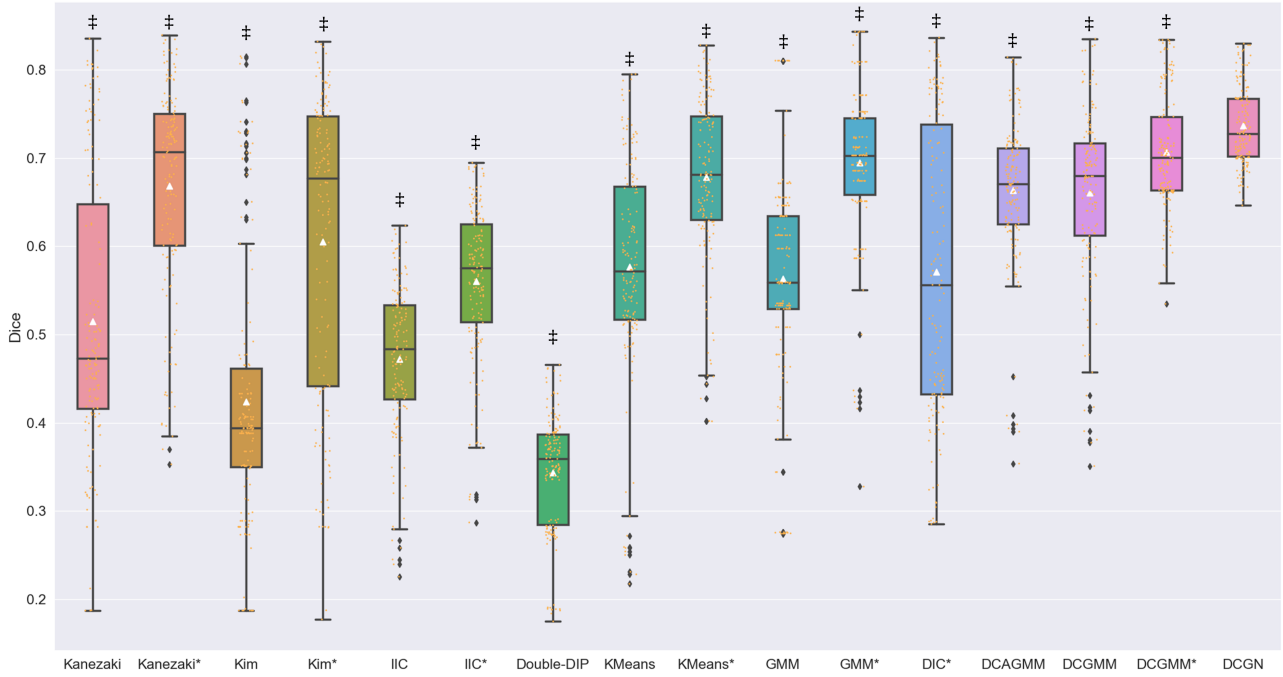


Figure 4.3: Dice Score Box Plot from Repeated Experiments. The asterisk (*) marks models with a redundant class (pre-defined classes k=3 for cell segmentation). The dagger (‡) denotes results with highly significant differences (Wilcoxon signed-rank test, P¡0.001) compared to DCGN. Black dots represent outliers, white triangles show mean values and small orange dots are individual samples.

redundant class have been removed, as some methods achieved upper-bound performance by adding a redundant class (i.e., k=3).

Table 4.3: Performance of the glomerulus segmentation (RBI dataset).

Methods	NMI	Dice
mKMeans	0.200±0.040(0.200)‡	0.555±0.037(0.559)‡
GMM	0.328±0.051(0.328)‡	0.640±0.061(0.640)‡
DCGMM	0.207±0.042(0.229)‡	0.567±0.047(0.579)‡
Kanezaki	0.186±0.095(0.207)‡	0.502±0.109(0.537)‡
IIC	0.090±0.068(0.124)‡	0.501±0.054(0.534)‡
Kim	0.187±0.089(0.195)‡	0.500±0.106(0.511)‡
DIC	0.192±0.117(0.234)‡	0.516±0.127(0.558)‡
DCAGMM	0.207±0.041(0.212)‡	0.578±0.048(0.582)‡
DCGN	**0.377±0.053(0.384)**	**0.735±0.050(0.746)**

‡ indicates highly significant differences results (P¡0.001 in Wilcoxon signed-rank test) compared with the proposed DCGN. Bold values demonstrate the best performance among comparison methods. The results are shown as "mean±standard deviation (upper-bound results)".

Unsupervised Glomeruli Decomposition on RBI. The average performance of our comparative study on the RBI dataset is summarized in Table 4.3 and Figure 4.5, assessed using

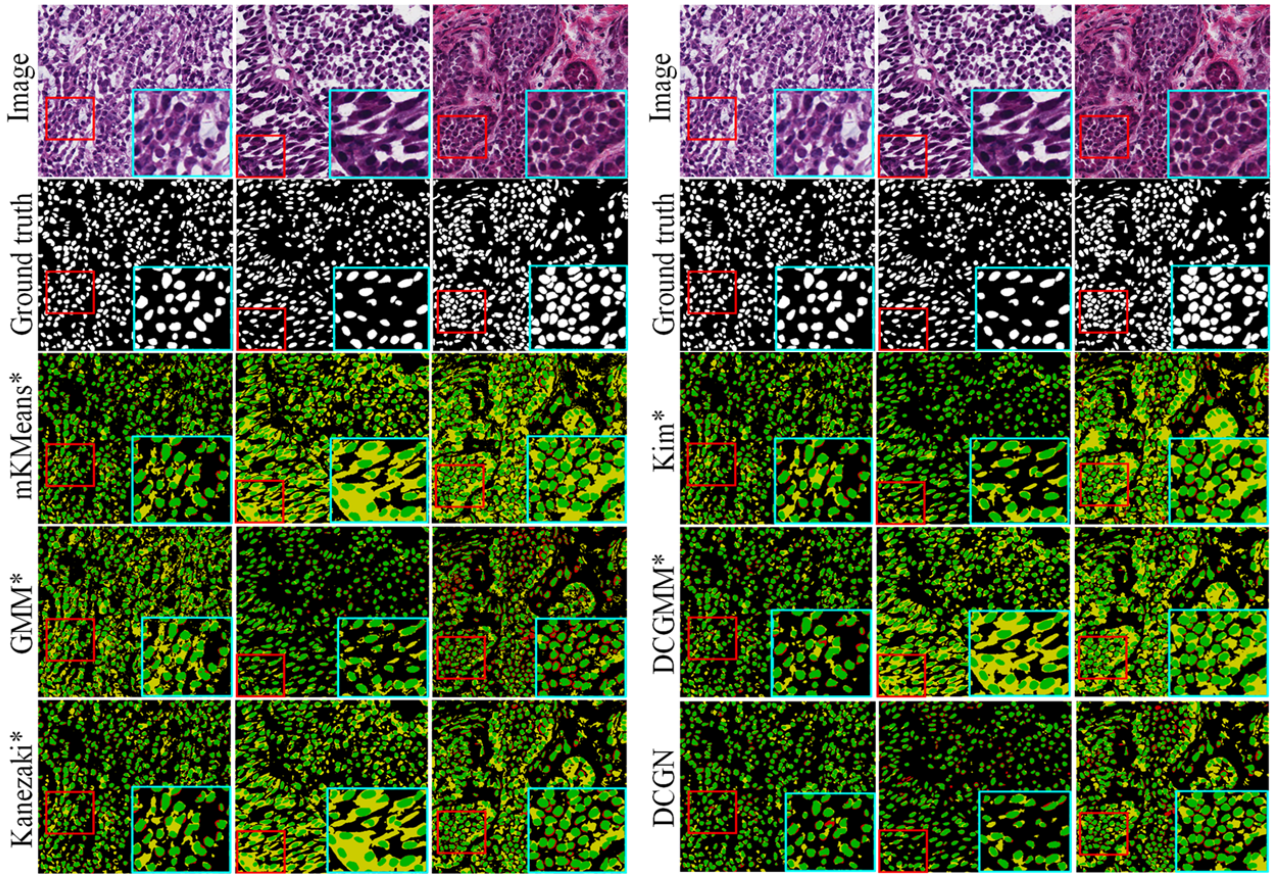


Figure 4.4: Unsupervised cell segmentation comparisons. * marks models with a redundant class (k=3). Green, yellow, and red colours show true positive, false positive, and false negative predictions, respectively. Red and cyan boxes emphasize regions of interest before and after zooming in.

the Normalized Mutual Information (NMI) and Dice coefficient score. All comparative studies used $k = 3$ for segmenting three semantic labels, as detailed in 4.5.2. As shown in Table 4.3, the proposed DCGN significantly outperformed state-of-the-art methods in glomeruli composition ($P<0.001$), achieving an average Dice score of 0.735 and an NMI of 0.377. This was followed by GMM, with an average Dice score of 0.640 and NMI of 0.328, and DCAGMM, with an average of 0.578 Dice and 0.207 NMI.

Degeneration Assessment. The results of the degeneration assessment are presented in Table 4.4. Notably, Double DIP was excluded from this assessment due to its relatively weak performance. As indicated in Table 4.4, the methods proposed by Kanezaki and Kim exhibited significant susceptibility to all identified degenerative issues. Similarly, the empty class issue was prevalent in DIC. DCGMM and DCAGMM occasionally encountered the empty class issue, and GMM showed instability during repeated experiments. Both mKMeans and IIC demonstrated

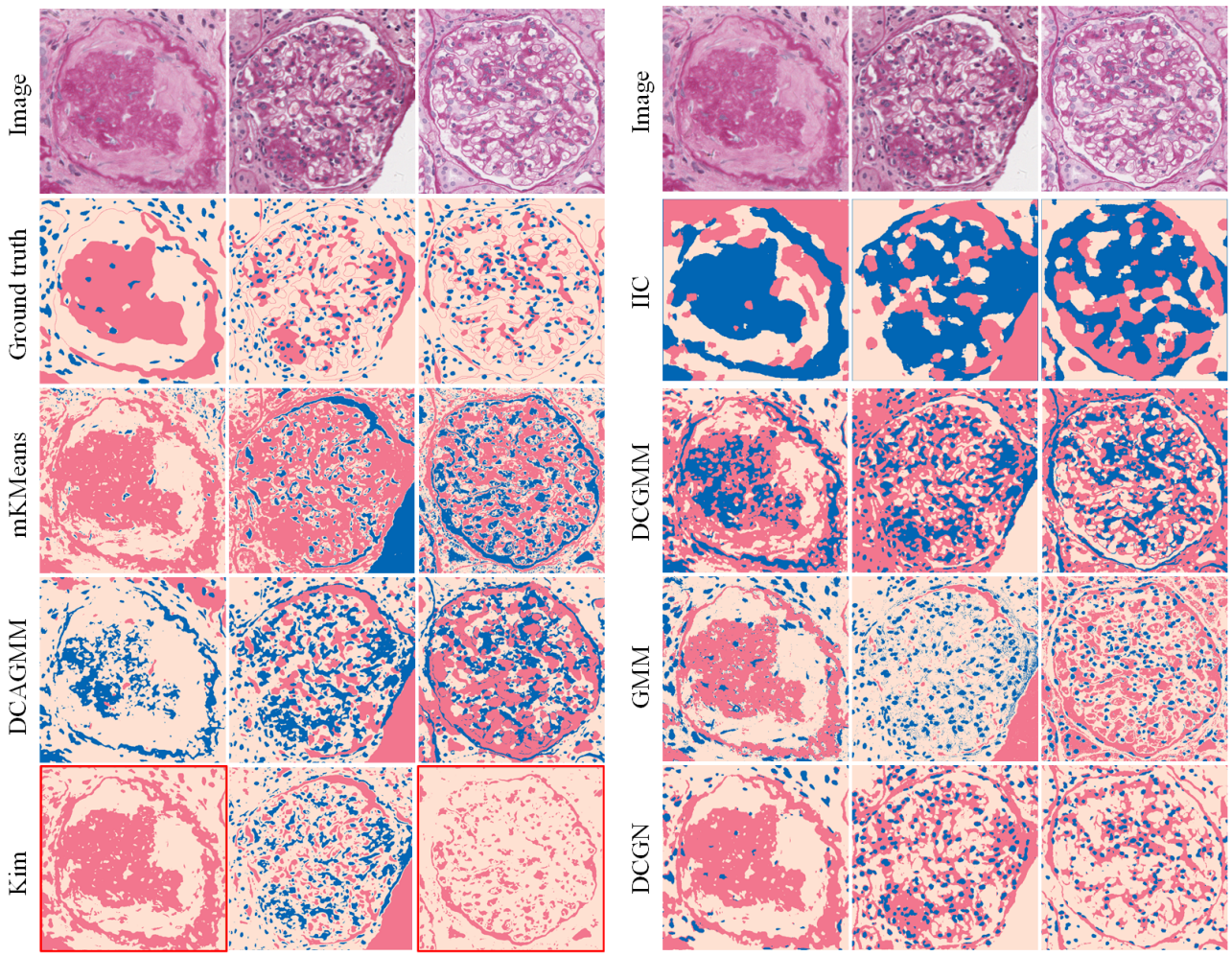


Figure 4.5: Comparison of unsupervised glomeruli segmentation. Red bounding boxes indicate empty class problems. In the ground truth, red, blue, and almond colors represent (1) mesangial matrix and basement membrane, (2) intra-glomerular and macula densa cells, and (3) other areas like glomerular capillaries, respectively.

Table 4.4: Degeneration assessment (MoNuSeG and RBI dataset)

Methods	**Collapse**	**Empty Class**	**Stability**
GMM	0/240	0/100	×
Kanezaki	7/240	23/100	×
Kim	17/240	28/100	×
mKMeans	0/240	0/100	×
IIC	0/240	0/100	×
DIC	0/240	25/100	×
DCAGMM	0/240	1/100	✓
DCGMM	0/240	2/100	✓
DCGN	0/240	0/100	✓

The red blocks indicate the occurrence of degenerative issues.

instability in the MoNuSeg dataset.

4.5 Discussion

In this study, we have developed a novel unsupervised segmentation method that integrates deep neural networks with a constrained Gaussian Mixture Model (GMM). This method was extensively evaluated on pathological images using the public MoNuSeg dataset and an in-house RBI dataset. Our approach demonstrated significant improvements over existing unsupervised segmentation methods, effectively addressing degenerative issues that pose challenges in the delineation of pathological tissue images. Additionally, our method yielded results comparable to some widely adopted semi-supervised and fully supervised learning methods.

Performance Analysis. The comprehensive comparison results, as detailed in Tables 4.2 and 4.3 and illustrated in Figures 4.4 and 4.5, underscore the superior segmentation capabilities of the proposed DCGN. DCGN exhibited robustness against small inter-class variations, a common challenge in unsupervised segmentation. For example, as shown in Figure 4.5, all unsupervised methods, except for DCGN, misclassified white regions as a single class, neglecting the exudation/stroma regions (light pink in the raw images).

Conventional methods like mKMeans and GMM have proven effective in tissue segmentation. GMM, in particular, outperformed mKMeans in kidney tissue segmentation, achieving a higher average Dice score of 0.08 and a higher NMI score by 0.12. However, methods proposed by Kanezaki et al. and Kim et al., while producing reasonable results in cell segmentation, encountered significant issues with collapse and empty classes, as evidenced by large variances in Fig 4.3 and numerous failures summarized in Table 4.4. Both methods exhibited poor performance in kidney tissue segmentation, as seen in Table 4.3 and Fig 4.5. DIC showed a high recall but low precision in cell segmentation, and performed poorly in glomeruli segmentation. Double DIP yielded coarse predictions similar to DIC, indicating its unsuitability for tissue segmentation, despite potential applicability in natural image segmentation. IIC's coarse predictions highlighted its inapplicability to pathological images. While DCGMM showed performance comparable to DCGN in cell segmentation, it significantly underperformed in kidney tissue segmentation and frequently generated empty classes. Moreover, DCGMM struggled with samples having small inter-class variations, as indicated by poor cell segmentation results

from dark background areas (Fig 4.4). DCAGMM, similar in performance to DCGMM, faced difficulties in segmenting classes with high intra-class variations due to its normalized distance constraint, designed to increase the distance between Gaussian centers.

Comparison with Fully Supervised Segmentation Methods. A critical aspect of evaluating unsupervised segmentation methods is their performance relative to fully supervised segmentation algorithms. Alongside the U-Net baseline presented in Table 4.2, we compared our proposed DCGN with various previously published supervised studies, as summarized in Table 4.5. It is important to note that all these comparisons were conducted using the same test data from the MoNuSeg dataset. As indicated in Table 4.5, the proposed DCGN achieved a Dice coefficient score comparable to that of the fully supervised U-Net-based method, with no significant differences observed in Precision, Recall, and Dice scores. While DCGN outperformed other unsupervised segmentation methods significantly, it exhibited a lower AJI score compared to both fully supervised and semi-supervised segmentation methods (Table 4.5). This discrepancy is primarily attributed to the adhesion of adjacent cells, a challenge that tends to be more effectively addressed by supervised or semi-supervised methods.

Table 4.5: Performance of supervised methods (MoNuSeG Dataset)

Methods	Avg F1 (Dice)	Avg AJI
Our DCGN	0.7432	0.3790
U-Net [106]	0.7582	0.4357
Mask RCNN [36]*	0.7991	0.5128
Dual U-Net [166]*	0.7913	0.5899
Tian et al. [167]†,*	0.7638	0.4927
Qu et al. [168]†,*	0.7566	0.5160
CNN [2]*	0.7623	0.5083

* refers to patch-based training, † indicates semi-supervised learning.

Analysis of Semi-Supervised Approaches and Patch-Based Methods. Semi-supervised approaches, such as those proposed by Tian et al. [167]] and Qu et al. [168], have incorporated prior knowledge of cell central points, resulting in competitive AJI scores of 0.4927 and 0.5160, respectively. Moreover, a notable improvement in average AJI was observed with patch-based methods (denoted with * in Table 4.5) compared to raw image-based learning strategies. This underscores the significance of the patch learning strategy in tissue segmentation tasks, where

patches are extracted from original raw images during both training and testing processes. However, unsupervised segmentation approaches often struggle to produce precise pixel-level predictions, especially for densely packed and small objects.

Distinguishing Samples with Small Inter-Class Variations. The ability to distinguish samples with small inter-class variations is crucial for accurate tissue segmentation. We explored this capability by analyzing the class intensity maps of the top four methods in cell and kidney tissue segmentation. As illustrated in Figure 4.6(a), most unsupervised methods failed to segment background samples clearly and required a redundant class for more challenging samples. In contrast, DCGN effectively differentiated between background and foreground samples without necessitating an additional redundant class. Figure 4.6(b) shows that the mKMeans method produced hard boundaries due to its reliance on Euclidean distance measurement, whereas other methods yielded smoother boundaries. Notably, DCGN generated class intensity maps most similar to those from ground truth, highlighting its effective centralized function.

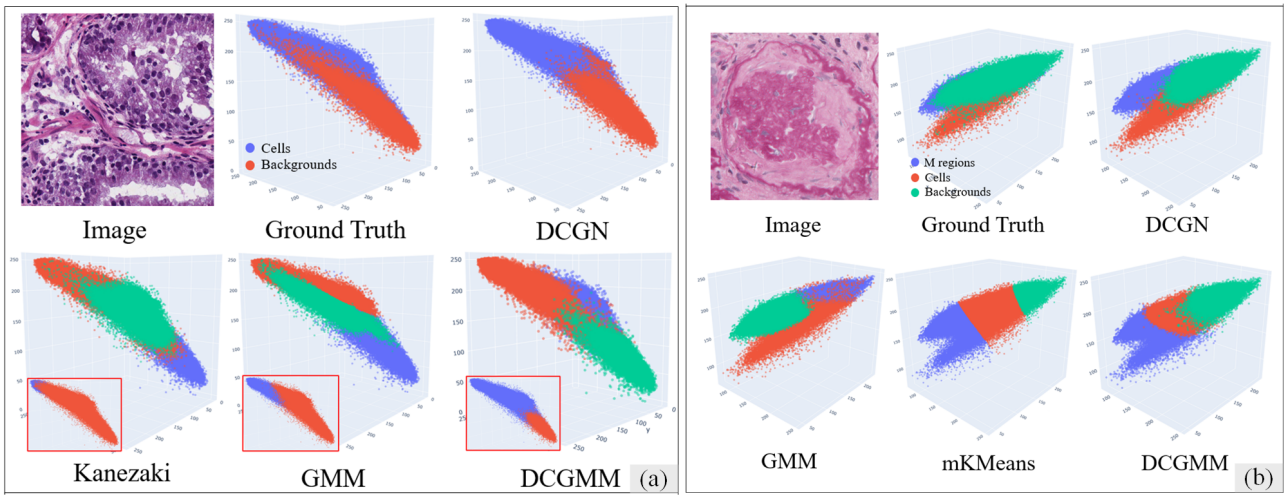


Figure 4.6: Unsupervised glomeruli segmentation comparison. Red bounding boxes highlight areas with empty class issues. In the ground truth images, red, blue, and almond colors signify (1) mesangial matrix and basement membrane, (2) intra-glomerular and macula densa cells, and (3) other regions such as glomerular capillaries.

Issues with Redundant Class and Stability. Experimental results indicated that many unsupervised segmentation methods grappled with the redundant class issue. As shown in Figure 3, most methods demonstrated significant improvements in binary segmentation performance when increasing the number of classes from 2 to 3. This suggests difficulties in distinguish-

ing samples with small inter-class variations. Fluctuations in performance were observed, as hard samples were variably labeled across different experiments. However, unlike other methods that required a redundant class to address these challenges, DCGN effectively segmented tissues without an additional redundant class, showcasing both efficiency and effectiveness. Regarding stability, IIC, mKMeans, DCGMM, and DCGN exhibited good stability in repeated experiments, as seen in Figure 4.3 and Table 4.5. Conventional methods like GMM, as well as Kanezaki's and Kim's methods, showed considerable variance in performance. Although the stability of previous methods improved with the introduction of a redundant class, their segmentation performance still lagged significantly behind DCGN.

Reproducibility and Empty Class Issues. Methods not suited for large-scale studies often face challenges with reproducibility. For instance, the conventional Gaussian Mixture Model (GMM) without minibatch learning is limited to a small dataset, which restricts the development of generalized segmentation models. Additionally, certain methods, such as those proposed by Kim and Kanezaki, can only process a single image during training, leading to low reproducibility in repeated experiments and inconsistent semantic label assignment. The empty class issue poses another significant challenge in unsupervised segmentation. Methods like Kim's, Kanezaki's, and DCGMM frequently encounter this problem during evaluation, mainly due to their inability to differentiate hard samples (i.e., pixels with similar intensities but belonging to different categories). In contrast, the proposed DCGN effectively avoids the empty class issue and demonstrates higher reproducibility in large-scale training scenarios.

Ablation Studies of Penalty Weights. The impact of the centralised constraint was examined by varying the penalty weights λ in Equation (3.5). The outcomes of 10 repeated experiments for each λ setting are summarized in Table V. It was observed that while the upper-bound performance of models with different λ values remained relatively consistent (0.740, 0.743, and 0.745 for $\lambda = 0.05$, $\lambda = 0.005$, and $\lambda = 0.0005$, respectively), the standard deviation of the Dice score exhibited significant differences. As Table 4.6 indicates, a larger weight for the centralised constraint promotes faster convergence but can also lead to an unstable training process, potentially due to local optimum trapping. Conversely, a smaller weight, while necessitating more training epochs for convergence, ensures a more stable training process.

Table 4.6: Ablation studies of constrained weights

λ	Dice	Avg Epochs
0.05	0.637 ± 0.076 (0.740)	37
0.005	0.737 ± 0.043 (0.743)	62
0.0005	0.734 ± 0.005 (0.745)	89

"Avg Epochs" indicates the average number of epochs for convergence.

Capacity on Whole Slide Images. The performance of DCGN on whole slide images, particularly when predictions are made across patches (tiles), was evaluated using a renal whole slide image. Our method demonstrated promising results in cell segmentation (two classes) on renal images with homogenous features. However, instances of false positives were noted in some vessel regions, suggesting potential avenues for future research, such as enhancing the utilization of textural features to improve module capacity.

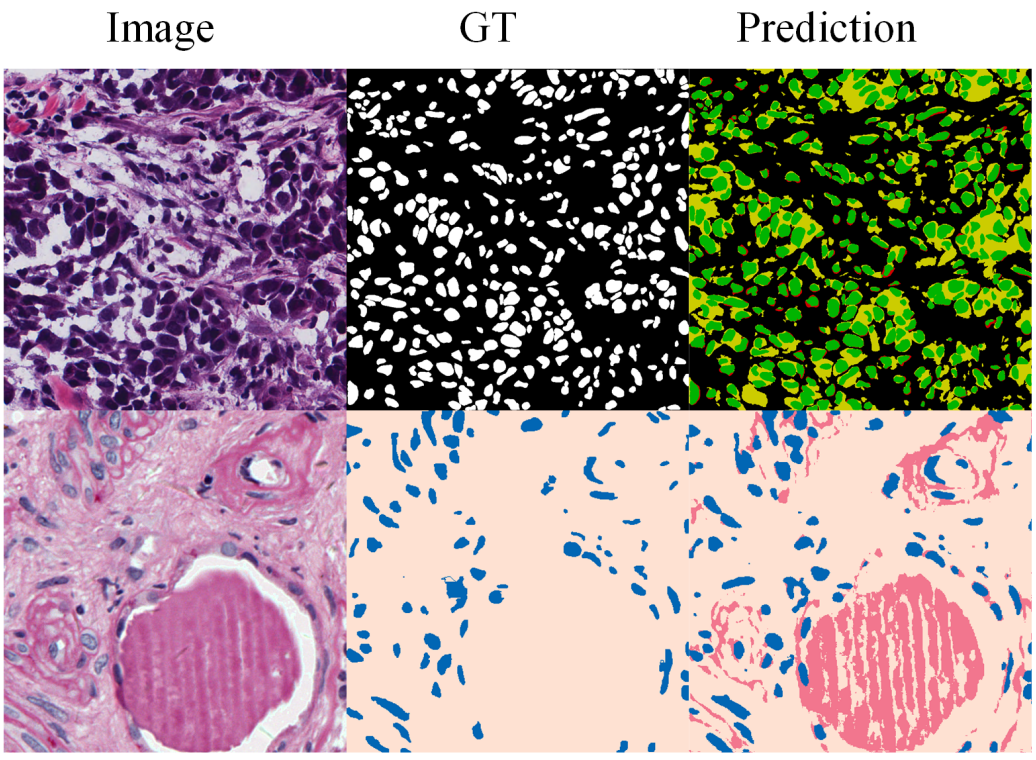


Figure 4.7: Failed predictions of glomerular structures and cells.

Limitations. A fundamental limitation of unsupervised learning is its reliance on allocating identical labels to samples of the same class, often without precise segmentation predictions due to the lack of prior knowledge or annotation. In comparison with existing pathological image segmentation studies [169, 170], our method may fall short in producing satisfactory instance

segmentation results, particularly in cases where cells tend to adhere, potentially limiting its clinical applicability for single-cell analysis.

Most unsupervised learning methods focus on pixel intensities, neglecting textual features. While integrating deep neural networks with clustering or mixture models can enhance textural feature utilization, the reliance on pixel intensity-based objective functions is still evident. Weak predictions are observable in cell segmentation (as shown in the first row of Fig. 4.7) and glomerular structure segmentation (second row of Fig. 4.7). This issue mainly stems from the discrepancy between hypothesized Gaussian distributions and real data distributions. Although DCGN may struggle with complex images having numerous categories or many outliers, it excels in general tasks like foreground/background segmentation. The centralised constraint of DCGN helps establish better classification boundaries for classes with small inter-class variations, a significant technical contribution of our method. However, while it can reduce false predictions, it cannot entirely eliminate them.

Alleviating Degenerative Issues with DCGN. To provide a clearer understanding of how DCGN addresses degenerative issues, we created schematic illustrations using simplified 2D examples (since real 3D clusters are complex to demonstrate and comprehend). For instance, the missing class issue often arises when modules fail to correctly handle outliers, such as classifying outliers as a unique class while merging different categories (e.g., blue and red dots) into a single class, as depicted in Fig.4.8(a). This problem is more common in iterative methods that depend on pseudo labels and is also occasionally seen in existing deep Gaussian networks. The centralised constraint in DCGN encourages the mixture module to align closer to the data sample centroids, thereby mitigating the occurrence of the missing class issue.

Additionally, the redundant class issue, often an artificial construct to enhance performance in unsupervised methods, arises from the discrete distribution of certain classes. Some methods necessitate an additional class to accommodate specific samples (e.g., blue samples within green dotted circles in Figure 4.8(b)). While avoiding redundant classes by setting an appropriate number of classes is feasible, modules without centralised constraints struggle to perform well, as evidenced in Fig 4.3.

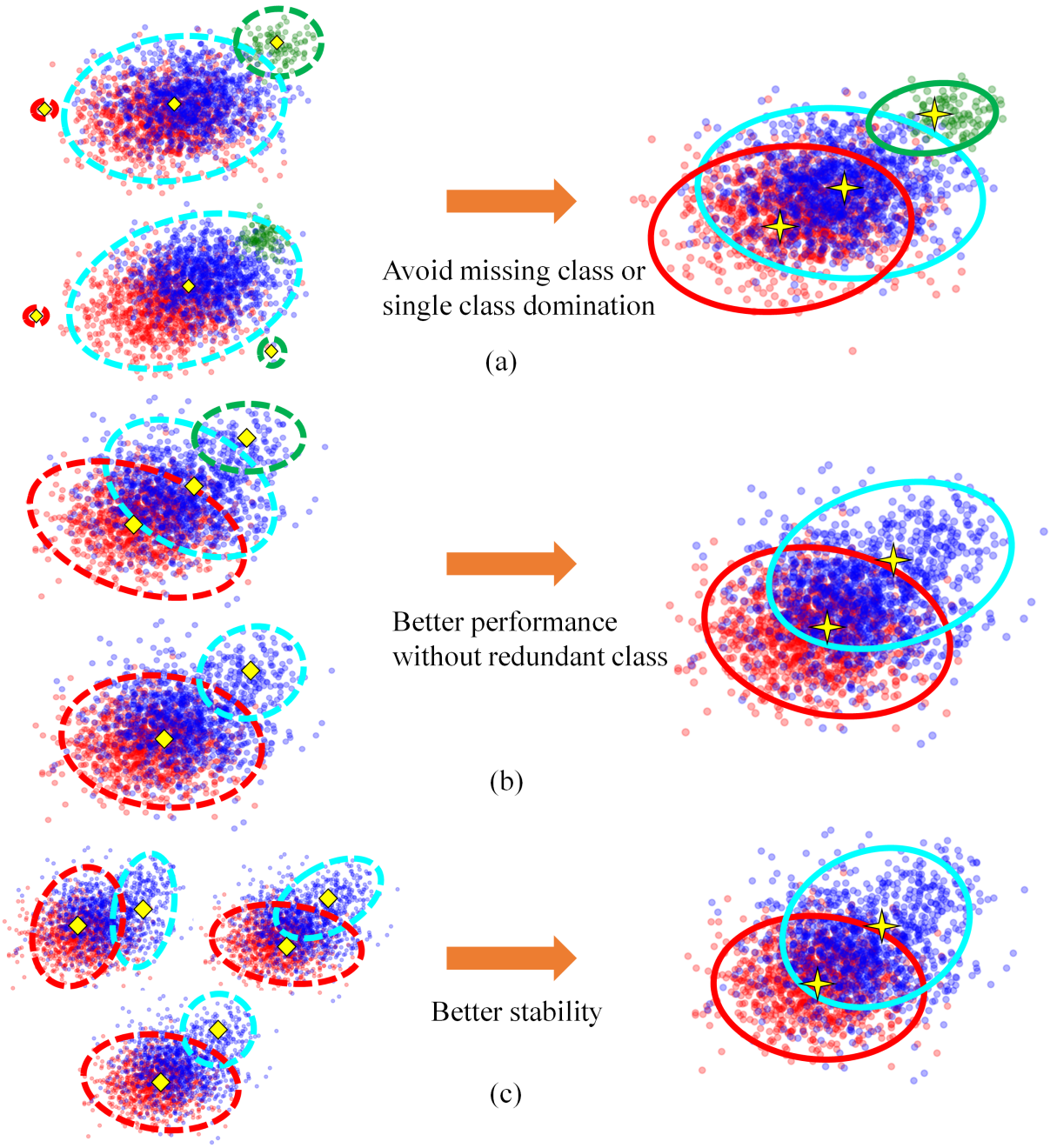


Figure 4.8: Illustration of How Our Centralized Constraint Solves Issues: (a) shows handling of missing class and single class domination (collapse), (b) demonstrates addressing redundant class, and (c) deals with instability. Predictions by methods lacking this constraint are marked with dotted circles (left column). Class centroids are represented by yellow diamonds (from methods without the constraint) and yellow stars (from our proposed method).

Moreover, instability or low reproducibility stems from random initialization. The centralised constraint in DCGN mitigates the randomness caused by initialization, guiding the module to learn parameters that approximate the data centroid. This approach has resulted in the lowest variance of evaluation metrics, as detailed in Table 4.2.

Suggested Criteria and Future Directions. Our study reminds us to use specific evaluation criteria for unsupervised segmentation methods: 1) Conduct repeated experiments to assess stability and reproducibility, and 2) Discuss degenerative issues in detail to evaluate method robustness. Potential future research directions in unsupervised segmentation include examining the performance of DCGN on various image modalities, such as segmenting tumours from brain magnetic resonance scans or organs from computerized tomography images. Additionally, exploring uncertainty estimation in semantic predictions for unsupervised segmentation could be valuable. Integrating a self-supervised paradigm with unsupervised learning, particularly using 'confident' predictions, may lead to enhanced performance. Furthermore, developing methods that can handle images with multiple classes remains a challenge, as most unsupervised segmentation approaches currently excel only in relatively simple semantic predictions. Lastly, creating a robust model capable of better-addressing outliers is a crucial goal for future advancements.

4.6 Conclusion

Tissue segmentation is a critical component in computational pathology. Traditionally, most existing methods require extensive manual annotations, posing significant time and resource challenges. This study introduces an effective approach for unsupervised tissue segmentation using the innovative DCGN method. The proposed DCGN method stands out in its ability to accurately segment tissue structures without relying on any manual annotations or prior knowledge. This characteristic of DCGN holds the potential to significantly reduce annotation costs in computational pathology, thereby streamlining the process and enhancing efficiency.

Chapter 5

Algorithms Benchmarking in Medical Imaging

5.1 introduction

Pulmonary fibrosis, characterized by symptoms like dyspnoea, ultimately leads to mortality. Despite the severity of the condition, there is no cure for pulmonary fibrosis, and available treatments primarily aim to manage symptoms and slow progression. Idiopathic Pulmonary Fibrosis (IPF) is notably the most serious chronic form of lung fibrosis, causing progressive scarring of the lungs and increasingly difficult breathing, with no known cause [171]. The gold standard for IPF diagnosis is invasive lung biopsy, but this is often rejected by patients due to its high risk (2% mortality rate within 100 days post-biopsy) and cost [172,173]. Consequently, there is a pressing need for non-invasive clinical methods to diagnose IPF in its early stages. The first sign of fibrotic lung disease is typically the distension of airway structures, traditionally assessed by human experts. Manual airway modelling is, however, laborious and time-consuming, taking nearly 5 hours per case, and suffers from subjectivity and potential oversight of subtle changes, especially in early-stage disease. In contrast, automatic quantification of airways offers an objective, stable, and cost-effective alternative to manual assessments.

Despite several challenges and publicly available datasets focusing on airway modelling, lung

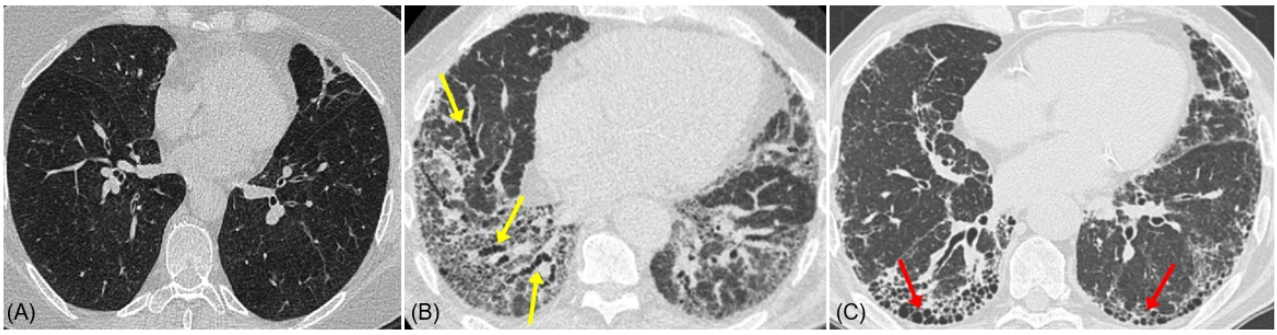


Figure 5.1: High-Resolution CT (HRCT) Images: (a) shows patients with nodules from the LIDC dataset, (b) features patients with traction bronchiectasis (indicated by yellow arrows), and (c) highlights patients with honeycombing (marked by red arrows). Honeycombing complicates the extraction of airway trees, resulting in suboptimal airway modelling.

fibrosis has received limited attention. Current open challenges, such as EXACT'09 [72] and ATM22 [174], mainly involve cases from healthy individuals or patients with common lung diseases, with ATM22 including some COVID-19 cases for additional assessment. However, no datasets specifically cater to complex fibrotic lung disease. Fibrotic lung disease patients experience distinct pathologies, such as volume loss, bronchiectasis, distortion of the secondary lobule, and loss of smooth air-to-interstitium interfaces, differing notably from other pulmonary diseases. For instance, unlike patients with other lung conditions shown in Fig. 5.1(a), those with fibrosis often exhibit bronchiectasis with significant distension of terminal/small branches, as depicted in Fig. 5.1(b). Furthermore, the presence of honeycombing, a UIP-like (Usual Interstitial Pneumonia) pattern in fibrosis cases, as shown in Fig. 5.1(c), complicates automatic airway modelling. Honeycombing, characterized by small cystic airspaces in the lung periphery resembling bronchial patterns, leads to significant challenges in airway leakages and mis-predictions in computational models.

Developing reliable AI systems for fibrotic lung disease involves overcoming challenges associated with abnormal morphological changes in lung tissues and airway trees. The primary challenges include: 1) **Airway Leakages:** Initially mentioned in [72], airway leakages refer to false-positive predictions (voxel-wise). This issue arises from the resemblance between airway trees (particularly terminal branches) and lung periphery tissues. In fibrotic patients, honeycombing patterns intensely confuse AI models, often misclassifying these honeycombs as airway branches, leading to extensive airway leakages. While the EXACT'09 organizers reported leak-

age counts and volumes, they were not fully effective in assessing leakage conditions due to varying lung tissue scales in individuals. 2) **Missing and Discontinuity Airways:** Various imaging conditions can render airway intensity values similar to those of lung tissues, leading AI models to falsely classify airway voxels as background, resulting in missing or discontinuous airways. 3) **Weak Robustness in Fibrosis Cases:** AI models developed for normal lung diseases often lack robustness in fibrosis cases due to complex abnormal changes. An existing study [121] highlights a significant performance gap in fibrosis cases compared to COVID-19 cases, although the models performed adequately in the latter. 4) **Limited Research on AI Prognostication:** Despite manual-derived biomarkers showing associations with mortality, the potential of automatically derived airway imaging biomarkers for prognostication in fibrotic lung disease remains unexplored.

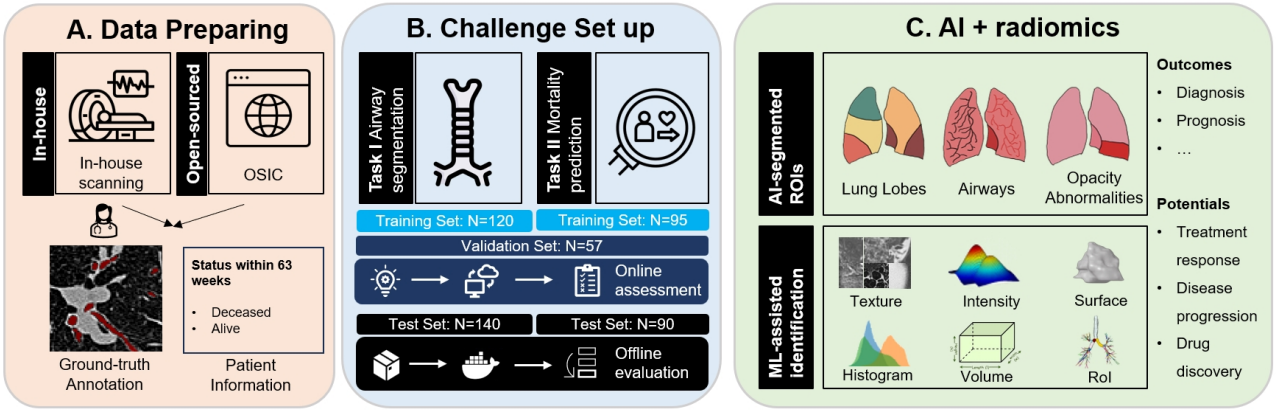


Figure 5.2: High-Resolution CT (HRCT) Images: (a) shows patients with nodules from the LIDC dataset, (b) features patients with traction bronchiectasis (indicated by yellow arrows), and (c) highlights patients with honeycombing (marked by red arrows). Honeycombing complicates the extraction of airway trees, resulting in suboptimal airway modelling.

The 2023 Airway-Informed Quantitative CT Imaging Biomarker for Fibrotic Lung Disease Challenge (AIIB23), held in conjunction with MICCAI-2023 (Fig. 5.2), aimed to accelerate progress in computational approaches for complex airway extraction, investigate imaging biomarkers for prognostication, and highlight ongoing challenges and research directions in this field. Hosted on the CodaLab platform [175], AIIB23 comprised two subtasks: airway segmentation (Task I) and mortality prediction (Task II). The challenge attracted 148 team registrations from over 300 participants. Teams were evaluated based on overlapped metrics and branch scores across 52 cases, with the top 10 teams ranked by both accuracy and efficiency in 140 offline test cases.

This study's contributions are threefold:

- AIIB23 establishes a new benchmark for AI-based airway modelling in fibrotic lung disease, releasing 120 HRCT scans with rare expert annotations, marking the first airway reference for this condition.
- The challenge incorporates a prognostication task, promoting participants to develop models that predict the progression of fibrosis, fostering a comprehensive approach to patient care.
- It showcases state-of-the-art airway modelling approaches and underscores remaining challenges for AI models in airway segmentation, investigating independent airway-derived biomarkers for mortality predictions in fibrotic lung disease.

This study is organized as follows: Section 4.2 introduces related works in airway segmentation, AI-based prognostication in fibrotic lung disease, and publicly available datasets for airway segmentation. Section 4.3 delves into the AIIB23 challenges, detailing the criteria for airway segmentation and mortality prediction tasks. Section 4.4 reports the results of the AIIB23 challenges, with additional offline evaluations further analyzing the collected results.

5.2 Related work

5.2.1 Airway segmentation

Airway segmentation, crucial for deriving imaging biomarkers, involves modelling 3D binary masks of airway trees from High-Resolution Computed Tomography (HRCT) scans. Given an HRCT scan X, the airway segmentation model Φ aims to extract airway tree masks $M_A = \Phi(X)$. This technique has been applied to various pulmonary diseases, including lung cancers [78, 87] and common pulmonary conditions [69, 70, 72, 75].

The first significant contribution to airway segmentation dates back to 1994, with Sonka et al.'s knowledge-based method for segmenting intrathoracic airway trees [176]. Subsequent develop-

ments included image processing techniques like morphology operations, region growing [67], intensity thresholds [71], and fuzzy connectedness [70]. For example, Xu et al. enhanced fuzzy connectedness segmentation with a fuzzy affinity relationship and utilized a spatially constrained Markov random walk for airway wall thickness estimation [70]. Mori et al. implemented region growing to extract bronchial regions from 3D CT images and employed the centreline for anatomical labelling [177]. However, these traditional methods often suffered from weak robustness and reproducibility due to variations in airway voxel intensities [80].

Since 2015, deep learning has revolutionized airway segmentation. Charbonnier et al. used a 2D CNN to classify airway patches (identified by conventional methods) into true or false positives, reducing airway leakages [75]. Yun et al. developed a 2.5D CNN, incorporating spatial information from orthogonal slices [76]. Additionally, 3D CNNs have shown superiority by integrating more spatial context. To overcome GPU memory limitations with 3D data, patch-based strategies and lung region cropping from HRCT scans have been adopted [77, 78]. The 3D U-Net, particularly its variant nnUNet [74], has become a common baseline for 3D airway segmentation due to its robustness and reproducibility [174]. Other notable advancements include graph refinement [108], connectivity awareness [78], centreline-aided learning [88], hard samples mining [87], fuzzy attention mechanisms [121], and transformer-based 3D U-Nets [178]. For instance, Nadeem et al. [90] and Qin et al. [86] modified the standard 3D U-Net by incorporating voxel-connectivity awareness and freeze-and-grow propagation, addressing airway leakage issues.

Despite these advancements, existing models struggle with fibrotic lung cases characterized by complex abnormalities like bronchus deformation, ground glass/consolidation, and honeycombing, even when trained on numerous other pulmonary disease cases [121]. Performance gaps in fibrosis cases, when compared to COVID-19 cases, are evident, with significant declines in Intersection over Union (IoU) score, Detected Length Ratio (DLR), and Detected Branch Ratio (DBR) [121]. The effectiveness of these models when trained specifically on fibrotic lung cases, and their performance on COVID-19 and other pulmonary diseases, remains an area for further exploration.

5.2.2 AI-driven prognostication of lung fibrosis

Radiology plays a crucial role in diagnosing and managing pulmonary diseases, particularly lung fibrosis. Chest CT scans showing characteristic features, such as those indicative of Idiopathic Pulmonary Fibrosis (IPF), can be used for diagnosis without the need for pathologic confirmation [179]. The guidelines from the American Thoracic Society/European Respiratory Society/Japanese Respiratory Society/Latin American Thoracic Association (ATS/ERS/JRS/ALAT) include specific imaging patterns in CT scans, like honeycombing, as key prognostic factors [180] However, the accurate quantification of pulmonary fibrosis from CT scans is still a challenging task, limiting their current prognostic utility. As a result, physiologic indicators like Forced Vital Capacity (FVC) and the Composite Physiologic Index (CPI) continue to be primary prognostic biomarkers despite certain limitations.

To overcome these challenges, researchers have explored the prognostic values of imaging-derived biomarkers. The initial step towards High-Resolution CT (HRCT)-derived biomarkers was proposed by Best et al. [181], who found a correlation between the kurtosis of lung HRCT histograms and mortality. Another significant advancement was CALIPER (computer-aided lung informatics for pathology evaluation and ratings), an AI-driven method for lung fibrosis prognosis. CALIPER uses histogram characteristics to classify and quantify the disease, predicting outcomes in various interstitial lung diseases (ILDs) [182]. Nam et al. developed a cascaded AI model, comprising two U-Nets, to segment various lung abnormalities [183]. Their findings demonstrated that the proportions of normal and fibrotic lung volumes were independent prognostic factors for overall survival in IPF patients. Humphries et al. [184] proposed an HRCT-derived metric for IPF diagnosis and prognosis, examining the relationship between image-derived biomarkers (quantitative measurement of fibrosis regions) and physiological measurements like FVC and Diffusing Capacity of the Lung for Carbon Monoxide (DLCO). They reported that the minimal clinically important difference for HRCT-derived biomarkers was 3.4%. Despite these efforts, the potential of airway-derived biomarkers for the prognosis of lung fibrosis diseases remains an area of ongoing research.

5.2.3 Publicly available datasets and open challenges.

This section reviews existing publicly available datasets and open challenges in airway segmentation and lung fibrosis prognostication, highlighting their differences from the AIIB23 challenge.

EXACT'09

EXACT'09, the inaugural open challenge for airway segmentation, featured 40 cases to evaluate 15 different airway segmentation algorithms. The cases, sourced from eight institutions, exhibited a variety of pathological abnormalities, collected under different acquisition protocols and devices, either at peak inspiration or full expiration. The dataset was divided into training and test sets, with 20 cases each. Notably, only the training cases were publicly released. The slice thickness in BAS ranged from 0.6 mm to 1.25 mm, and the reconstruction kernels included B50f, FC12, FC10, B30f, and B70f. The anomalies covered healthy volunteers, patients with small lung nodules, ground glass, and diffuse emphysema. Techniques applied to EXACT'09 primarily involved image processing methods, such as morphology approaches [185], and variations of region-growing [186, 187].

ISICDM21

ISICDM21 consisted of 12 cases with expert annotations, each slice being 512x512 pixels with slice thicknesses varying from 0.625 mm to 1.5 mm [178]. However, this dataset is currently not accessible.

Binary Airway Segmentation (BAS)

BAS, a publicly available dataset, includes 90 HRCT cases. Twenty of these cases were from EXACT'09, and the remaining 70 were sourced from the Lung Image Database Consortium image collection (LIDC) [100]. The slice thickness for these 70 cases ranged from 0.45 mm to 0.625 mm, primarily featuring lung nodules. The number of slices in BAS data varied significantly, from 157 to 764, indicative of diverse scanning protocols.

ATM22

ATM22, a multicentre airway segmentation dataset, comprises 500 cases (300 for training, 50

for validation, and 150 for testing), releasing 300 cases with semantic annotations to the public. Of these 300 cases, 90 were from the BAS dataset, while the remaining 210 were collected from LIDC-IDRI data and Shanghai Chest Hospital. The number of slices per case ranged from 157 to 1125, with slice thicknesses varying from 0.45 mm to 1.00 mm.

5.3 Method

5.3.1 The AIIB23 Dataset

The AIIB23 challenge incorporated lung CT scans from patients with fibrosis and COVID-19, collected from OSIC, AIPFR, and UHP (demographics shown in Table 5.1). Data from the OSIC repository were randomly split into training and validation sets for public release, while AIPFR and RBH data formed the offline test set (Fig. 5.2).

Table 5.1: Demographics of data used in AIIB23 challenge.

Attribute	OSIC (n=172)	AUS (n=90)	RBH-COVID (n=50)
Age (years)	68 (57, 75)	70 (67, 76)	65 (45, 82)
BMI (kg/m^3)	-	28.39 (26.27, 30.92)	-
Gender			
Male	101 (58.7%)	54 (61.1%)	24 (48%)
Female	71 (41.3%)	36 (38.9%)	26 (52%) *
Diagnosis			
ILD	99 (57.6%)	0	0
IPF	62 (36.0%)	71 (78.9%)	0
COVID-19	0	0	44 (88%)
Other abnormalities	11 (6.4%)	19 (21.1%)	6 (12%)

Continuous variables are shown as Median (Q1, Q3).

- **OSIC:** The Open Source Imaging Consortium (OSIC) is an open-access global dataset comprising multidimensional medical images (e.g., 3D HRCT, 3D MRI) from patients with IPF, fibrotic ILD, and other respiratory diseases. It includes clinical metadata like physiologic indicators (FVC percentage, DLCO), body measurements (height, age, gender), mortality status, and clinical diagnoses. The dataset adheres to GDPR/HIPAA privacy standards to ensure diversity and maximal dataset size.

- **AIPFR:** The Australian IPF Repository (AIPFR) [188], ethically approved by the Sydney Local Health District (protocol no. X14-0264), contains diagnoses including IPF, probable IPF, alternative diagnoses, and other fibrotic patterns following the 2018 ATS/-ERS/JRS/ALAT IPF guideline statement [179].

- **UHP:** The University Hospital of Parma (UHP) dataset follows the local Ethics Committee's approval (code 934/2021/OSS/AOUPR - 11.01.2022). It includes patients admitted with COVID-19 symptoms/signs and confirmed by PCR tests.

All HRCT scans, initially in DICOM format, were converted to Nifti format using the SimpleITK 2.1.0 package in Python 3.10.0. Preliminary annotations were provided by an AI model trained on the BAS dataset [121], followed by detailed revisions by three junior radiologists and then refined by senior radiologists. Annotations were made using a HUION Kamvas 16 Tablet via ITK-SNAP 4.0.1 [189]. The challenge dataset was carefully curated from the three databases, focusing on high-quality cases. For Task I (Task II), 120 (95) cases formed the training data, 52 cases were used for validation, and 140 cases (90 fibrosis and 50 COVID-19) constituted the test set (Fig. 5.2).

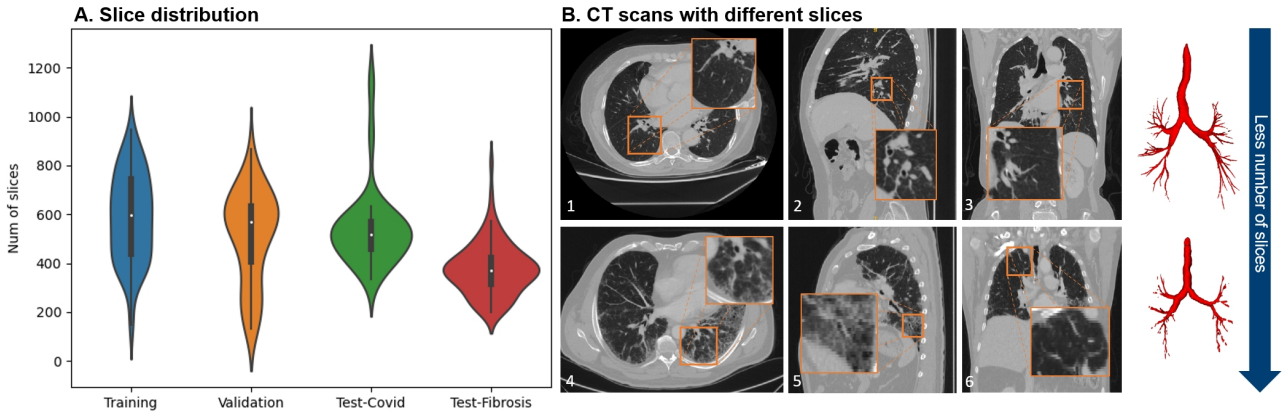


Figure 5.3: Differences in Slice Count in AIIB Datasets. A) Violin plot showing the distribution of slice counts per case in training, validation, and two test sets. B) Comparative visualization of two CT scans with varying slice counts and their respective annotations. Top row: CT scan comprising 1145 slices. Bottom row: CT scan with 140 slices (B1-B6).

AIIB23 showcased diverse imaging qualities, primarily differentiated by slice counts. An analysis of the slice number distribution in training, validation, and test sets (Fig. 5.3A) revealed that training and validation data (OSIC) generally had higher quality than test fibrosis data

(AIPFR), as indicated by a greater number of slices ($p<0.05$). This difference in slice count affected image quality, especially in coronal and sagittal planes of low-slice CT scans (Fig. 5.3B5 and B6) compared to high-slice scans (Fig. 5.3B2 and B3), with the former showing more visual blurriness and jaggedness. This discrepancy also impacted expert annotations due to the loss of detailed information between slices.

5.3.2 Challenge setup

Subtasks The AIIB23 challenge consisted of two main subtasks:

1. **Task I - Airway Segmentation:** Participants were tasked with segmenting airway trees from HRCT scans, with a specific focus on extracting small airway branches.
2. **Task II - Mortality Prediction:** This task involved predicting the mortality status (survival or not) of patients at 63 weeks after the initial HRCT scan and diagnosis.

Schedules The data annotation process spanned from January to April 2023, following the challenge's proposal submission to MICCAI 2023. The official challenge page was launched on 15th May 2023, with TASK I and TASK II, marking the opening of challenge registration and the release of training data. The validation cases were released on 30th May, with online validation submissions starting on the same day. The test phase commenced on 30th July, with a deadline for docker file submissions set for 25th August.

Reproducibility: For the test phase, teams were required to submit a docker image and a short paper detailing their methods. Validation phase scores were re-evaluated using the submitted docker image. Participants were also encouraged to publicly share their original code.

Fairness: The use of external data for model training was prohibited, although pretrained weights (not related to airway datasets) were permissible. Pre-training and domain-adaptation approaches were allowed. All competitors were provided with an identical test environment, featuring 1 GTX3090 TI (24GB) and 64 GB RAM.

Submission Restrictions: Participants had to register using official email addresses to prevent multiple submissions by the same team under different accounts. Each team was limited to 50 submissions during the validation phase and only one successful submission in the test phase. A successful submission was defined as a docker image that could be executed successfully and generate predictions.

Evaluation metrics

(1) Task I-Airway segmentation. The evaluation metrics mainly included IoU score, DLR, DBR, precision, ALR, and inference time, which have been demonstrated in section 2.2.3. The final results were ranked based on both the overall accuracy ($OvAcc$) of airway modelling and the efficiency of the algorithms (inference time), with

$$OvAcc = (\text{IoU} + \text{Precision} + \text{DBR} + \text{DLR}) \times 0.25, \tag{5.1}$$

$$R = R[OvAcc] \times 0.7 + R(t) \times 0.3, \tag{5.2}$$

where R is the ranking $R(x) \in \{0, 1, 2, \ldots, N\}$ and t refers to the inference time (seconds).
(2) Task II-mortality prediction. This task is a binary classification problem focused on predicting the mortality status of patients at 63 weeks after their initial HRCT scan. A prediction of 0 indicates mortality and vice versa. The evaluation metric is comprehensive, encompassing Accuracy, AUC (Area Under the Curve), Sensitivity, Specificity, and F1 Score. These metrics collectively assess the model's ability to accurately predict patient mortality at 63 weeks, evaluating both its effectiveness in correctly classifying deceased and alive patients and balancing the trade-off between sensitivity and specificity. Models with high specificity are especially valuable for precisely identifying patients at high risk of mortality within the 63-week period.

5.3.3 Participants

As an official challenge of MICCAI 2023, AIIB23 attracted a considerable number of applicants for both of its tasks. Specifically, there were 214 applications for registration, with 112 appli-

Table 5.2: Details of teams that participated in the validation phase and test phase (bolded).

Team ID (A to Z)	Affiliation	Location
DJ_92	LITIS laboratory, INSA Rouen Normandy	France
Dolphins	National Heart and Lung Institute, Imperial College London	UK
EarthIsFlat	Department of Psychiatry, University of Cambridge	UK
Gexing	Tianjin University	**China**
Icuneo	Istituto Italiano di Tecnologia, Genova	Italy
IMR	Institute of Medical Robotics, Shanghai JiaoTong University	China
Insimaging2023	Shanghai JiaoTong University	China
Junqiangmler	Shanghai MediWorks Precision Instruments Co.,Ltd	China
Likehood	Zhejiang University	China
MedibotTeam	Shanghai Microport MedBot(Group) Co., Ltd	China
Medxmu	Xiamen University	China
Neu204	College of Medicine and Biological Information Engineering, Northeastern University	China
Newtab	School of Biomedical Engineering, University of Science and Technology of China	China
Optima	European Institute for Systems Biology and Medicine	France
Polaris	School of Biomedical Engineering, Shanghai JiaoTong University	China
Riipl	Atrium Health Wake Forest Baptist	USA
Sanmed AI	Sanmed Biotech Inc.	China
Tastefish	Fudan University	China
Telecomparis	Institut Polytechnique de Paris-Telecom Paris	France
Twen	Infervision Medical Technology Co., Ltd	China
uAI-Team	Shanghai United Imaging Intelligence Co., Ltd.	China

cations for Task I (airway segmentation) and 102 for Task II (mortality prediction). Out of these applications, 53 were rejected based on the challenge's terms and conditions. For Task I, validation results were received from 20 teams, whereas for Task II, results came from 8 teams (detailed information can be found in 5.2). The top-performing approaches for both tasks are highlighted in this study: the top 10 methods for Task I and the top 5 for Task II. These include only the participants who consented to the publication of their solutions and results.

5.4 Results

5.4.1 Performance of airway modelling

The methodologies proposed by different teams in airway modelling demonstrated some commonalities, as detailed in Table 5.3. Notably, all submitted models were based on deep learning.

Each method employed a patch-learning strategy due to GPU memory limitations. This involved training models on sampled patches and then aggregating predictions from all patches during the inference phase. More than half of the teams (6 out of 10) performed lung extraction

Table 5.3: Summary of solutions proposed by top-10 teams in Task I.

Teams	WindowT	RoIEx	SpacingT	Norm	Data sampling	DA	ME	Loss
MedibotTeam	✗	✓	Median	z-score	Random crop	✓	✓	L_G, L_{Dice}, L_C
IMR	✗	✓	Median	max-min	Targeting crop	✓	✓	L_G, L_{CE}, L_{CA}
Twen	[-1000,500]	✓	✗	max-min	Hard-mining	✓	✗	L_G, L_{Dice}
Sanmed AI	[-1400, 200]	✓	✗	z-score	Crop	✓	✗	L_{Dice}, L_{FL}
Gexing	✗	✗	Median	z-score	Random crop	✓	✓	L_{Dice}, L_{CE}
DJ 92	✗	✗	Median	z-score	Crop	✓	✗	L_{Dice}, L_{CE}
Riipl	[-1400, 600]	✗	Median	z-score	Random crop	✓	✓	L_{Dice}, L_{CE}
earthIsFlat	✗	✗	✗	z-score	Random Crop	✓	✗	L_{Dice}
Dolphins	✗	✓	Median	z-score	Random Crop	✓	✗	L_{Dice}, L_{CE}
Junqiangmler	[-1000,-300]	✓	Mean	z-score	Random Crop	✓	✗	L_{Dice}, L_{CE}

Notes: WindowT refers to window transform (within certain HU values), SpacingT is spacing transform, RoIEx indicates RoI extraction, Norm refers to normalization, ME represents model ensembling, and DA is data augmentation, L_G: general union loss, L_C: airway continuity loss, L_{Dice}: dice coefficient loss, L_{CA}: connectivity-aware loss, L_{FL}: focal loss, L_{CE}: cross entropy loss.

prior to data sampling to minimize artefacts outside the lung region. Techniques included using open-source lung extraction algorithms [190] and employing a support vector machine (SVM) for classifying different connected components. Most participants applied a spacing transform to mitigate the anisotropy of raw CT data, achieved by resampling voxel spacing to the median of the entire training set. Interestingly, only four teams implemented window transforms on the raw CT data; others opted for HU value normalization strategies, such as max-min or z-score normalization. The top-3 methods utilized weighted general union loss L_G [87] and connectivity loss (e.g., L_C and L_CA) [121], alongside multi-stage training protocols. For data sampling, in addition to random cropping, all competitive methods (top-3 approaches) employed specific strategies like hard mining (sampling patches based on mispredicted airway skeletons) and targeted cropping (sampling patches around airway skeletons). These strategies allowed for more focused sampling around regions of interest (RoIs), addressing the issue of data imbalance. Interestingly, model ensembling also improved airway modelling performance, though it could reduce time efficiency without multi-thread techniques.

Table 5.4 shows the overall performance of the top-10 teams, ranked by model accuracy and inference time cost. MedibotTeam achieved the best results in the test phase, considering both inference speed (43.79 seconds per scan) and model accuracy (0.9185), which was 30.4% faster than the IMR team (62.98 seconds per scan). The IMR team achieved the best branch metrics

Table 5.4: Top-10 winners in AIIB23 Task I (n=140).

Teams	IoU	DLR	DBR	Precision	ALR	AMR	OvAcc	Time
MedibotTeam	0.9049	0.9365	0.9051	0.9276	0.0786	**0.0259**	**0.9185**	**43.79**
IMR	0.8770	**0.9510**	**0.9312**	0.9014	0.1089	0.0299	0.9152	62.98
Twen	0.9016	0.9201	0.8825	0.9399	0.0639	0.0425	0.9110	87.93
Sanmed AI	0.8903	0.8838	0.8354	0.9203	0.0854	0.0404	0.8825	92.78
Gexing	**0.9087**	0.8603	0.7765	0.9518	0.0499	0.0466	0.8743	356.04
DJ_92	0.8967	0.8466	0.7504	**0.9619**	**0.0397**	0.0686	0.8639	223.66
Riipl	0.9003	0.8560	0.7672	0.9584	0.0423	0.0620	0.8705	531.52
earth1isflatten	0.8221	0.6986	0.6043	0.9325	0.0906	0.1538	0.7644	155.36
dolphins	0.8838	0.8556	0.7627	0.9365	0.0682	0.0580	0.8597	1500.12
Junqiangmler	0.7488	0.7853	0.6853	0.8200	0.2188	0.1313	0.7599	316.89

among all competitors, with a Detected Length Ratio (DLR) of 0.9510 and a Detected Branch Ratio (DBR) of 0.9312. However, their model exhibited some weaknesses, including a high Airway Leakage Ratio (ALR) of 0.1089 and a precision score of 0.9014. Team Gexing attained the highest overlapped metric with an Intersection over Union (IoU) score of 0.9087. Despite this, their method overemphasized the IoU score, resulting in relatively lower branch scores (0.8603 DLR and 0.7765 DBR) compared to the top teams. Team DJ_92's model achieved the highest precision (0.9619) and the lowest ALR (0.0397), suggesting potential under-segmentation of terminal branches, which minimally affect overlapped metrics. Team dolphins, while achieving a comparable overall accuracy of 0.8597, required significantly longer processing time, approximately 1500 seconds per CT scan. Notably, all models demonstrated improved performance on out-of-distribution (OOD) data (COVID-19 cases), with a nearly 1-2% increase in IoU scores. Specifically, there was a notable increase in branch metric scores for models from DJ_92 (11.7% DLR and 16.1% DBR), earthis1flatten (16.8% DLR and 18.4% DBR), and Riipl (12.3% DLR and 17.3% DBR). This indicates that models trained on fibrosis cases can be effectively adapted to COVID-19 cases.

5.4.2 Performance of mortality prediction

In mortality prediction studies, two predominant strategies were observed (Table 5.5):

Radiomics with Machine Learning: This approach involved using Pyradiomics for extracting radiomic features. Teams like uAI expanded on this by incorporating hand-crafted

Table 5.5: Summary of solutions proposed by top-5 teams in Task II.

Teams	CT Windowing	Method Type	Semantic input	ME	Classifier	Overall score
uAI-Team	[-1350, 150]	ROI-based radiomics	✓lung lobes ✓pneumonia	✓	LGBM, LR, SVM , and XGBoost	**0.7059**
Tastefish	[-1000,400]	End-to-end CNN	✗	✗	3D CNN	0.7049
junqiangmler	[-1000, 800]	End-to-end CNN	✓lung region	✗	3D VNet	0.6222
Earth1isflatt1	✗	Image-based radiomics	✗	✗	XGBoost	0.6186
DJ_92	✗	ROI-based radiomics	✓trachea	✗	SVM	0.5787

LGBM: light gradient-boosting machine, LR: logistic regression, SVM: support vector machine, XGBoost: extreme gradient boosting, ME: model ensembling.

features, such as lesion volume and surface details. These extracted features were then utilized in machine learning models, such as XGBoost and SVM, for binary classification. This method circumvents the extensive training typically required for convolutional neural networks (CNNs).

CNN-Based Strategy: In contrast to the first approach, this strategy eschewed manually engineered features in favour of the inherent capabilities of CNNs. These solutions employed 3D neural networks, trained directly on the entire volumetric High-Resolution Computed Tomography (HRCT) scans, to perform binary classification tasks. This approach was noted for its speed in delivering predictions without relying on pre-defined feature sets.

These two strategies highlight the diverse methodologies employed in mortality prediction, each with its unique strengths and focus areas.

Table 5.6: Top-5 winners in AIIB23 Task II (n=90).

Teams	Overall score	ACC	Sensitivity	Specificity	F1 score
uAI-Team	**0.7059**	0.7528	0.7848	0.5000	**0.8493**
Tastefish	0.7049	0.6404	0.6203	0.8000	0.7538
junqiangmler	0.6222	0.5955	0.5949	0.6000	0.7231
earth1is1flatten	0.6186	0.6629	0.6962	0.4000	0.7857
DJ_92	0.5787	0.4045	0.3291	**1.0000**	0.4952

Table 5.6 demonstrates the performance of AIIB23 Task II of the top five participating teams. The uAI Team leads the ranking with the highest overall score of 0.7059, alongside an impressive F1 score of 0.8493. Following closely is Tastefish, distinguished by its high specificity at 0.8000. Junqiangmler and earth1isflatten occupy the third and fourth positions, with the latter notable for its F1 score of 0.7857, the second-highest in the ranking. DJ_92, while placing fifth, is unique in achieving a perfect specificity of 1.0000, but it contrasts with lower accuracy (0.4045) and sensitivity (0.3291) scores. This tableau underscores the varying strengths and capabilities

of each model across different evaluative metrics.

5.5 Discussion

This section delves into the research findings and provides a detailed analysis of the AIIB23 challenges. Initially, we summarize the common strategies deployed in airway modelling and mortality prediction, setting the stage for further exploration. This is followed by an in-depth analysis of the prediction errors made by different participating teams. A significant aspect of our discussion also involves examining the effectiveness of various models in handling unharmonized data, a critical consideration for clinical application. Furthermore, we investigate the reliability of radiomic features by conducting a comprehensive survival analysis. This includes not only the biomarkers proposed by the challenge participants but also extends to those identified in existing studies and our own proposed solutions. Such an analysis is instrumental in understanding the real-world implications and potential of these biomarkers in clinical settings.

Overall Performance of Airway Modelling in Task I

Task I demonstrated varied performances with intriguing findings (Table 5.4). The Overall Accuracy (OvAcc) among the top-2 teams was similar, at 0.9185 and 0.9152, with no significant differences ($p > 0.05$). This suggests comparable overall effectiveness, despite differing strengths and weaknesses. The top-3 methods, however, showed significantly higher OvAcc (0.03-0.16 improvement) compared to others (Wilcoxon signed-rank test, $p < 0.05$). Notably, these leading methods all implemented weighted general union loss, highlighting the value of specific optimization functions. Additionally, the first and second teams significantly outperformed the third in detection rates (Wilcoxon signed-rank test, $p < 0.05$), underscoring the benefit of optimizations emphasizing airway branch connectivity. This resulted in AI models better capturing the anatomical structure of airway trees, potentially leading to more clinically relevant outcomes.

Regarding airway modelling capacity, all models performed better on COVID-19 than fibrosis cases (Wilcoxon signed-rank test, $p < 0.05$). A significant difference in Detected Length Ra-

tio (DLR) and Detected Branch Ratio (DBR) between COVID-19 and fibrosis datasets was observed (Wilcoxon signed-rank test, $p < 0.05$), likely due to the distinct pathological features of each disease. The higher DBR and DLR in COVID-19 datasets may result from a greater prevalence of airway leakages and missing segments, making models more error-prone.

Overall Performance of Mortality Prediction in Task II

Task II, a binary classification task, aimed to predict patient mortality status after 63 weeks. The validation set was balanced, with 26 samples each for alive (label 0) and deceased (label 1) categories. However, the test dataset was imbalanced, with 87% survivors beyond the 63-week threshold. This imbalance was notably highlighted by team DJ_92, who ranked first in the validation stage but fourth in the test set.

Table 5.6 shows a consistent preference for radiomics features in quantitative imaging biomarker extraction, primarily due to their interpretability and lower GPU memory requirements. However, performance inconsistencies were evident, attributable to the reliance on ROI-based radiomics by most participants. This reliance can introduce variability due to differing region selection approaches. CNN-based methods also faced limitations, particularly in handling large volumetric CT scans. Teams Tastefish and junqiangmler addressed this with down-sampling and cropping, respectively, to reduce GPU memory demand.

ROC curves of the top-5 submissions (Fig. 5.4) illustrated the diagnostic capabilities of the binary classifiers at varying thresholds. The uAI-Team's model demonstrated the best performance (0.8 AUC score) in discriminating patient mortality status after 63 weeks. However, it should be noted that the uAI-Team's model required information on lesion characteristics.

Delving into Mispredictions in Airway Segmentation

Fig 5.5 offers a visual comparison of the airway modelling performance by the top eight teams on the fibrosis test set (N=90). Heatmaps are used to illustrate discrepancies between the actual ground truth and predicted outputs, with $H = (Y - X)$. The accumulation maps of predictions and ground truth were generated by summing slices along the z-axis, followed by calculating the residual map between these accumulations. The heatmaps utilize colour coding: warm colours indicate areas of false negatives (i.e., missed airway predictions), and cold colours

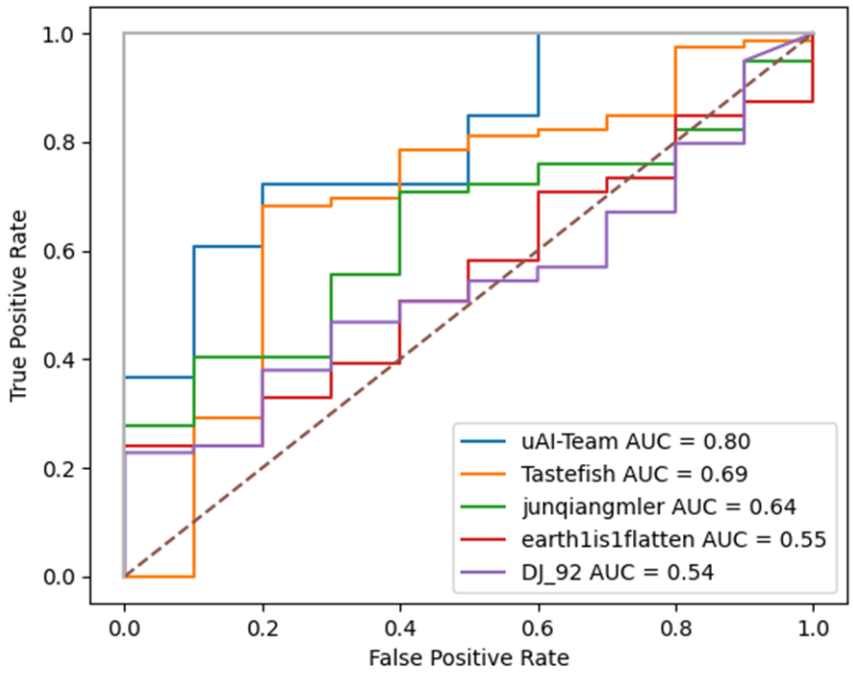


Figure 5.4: ROC of the top-5 teams in Task II.

represent false positives (i.e., incorrectly predicted airways).

This visual analysis reveals certain trends in the predictive abilities of different teams. For instance, MedibotTeam and IMR predominantly had false positive mispredictions in smaller-sized branches, contributing to their higher airway leakage rates of 0.08 and 0.11, respectively. The heatmaps elucidate IMR's scoring patterns, shedding light on their high DBR and DLR scores but lower IoU scores. White bounding boxes in IMR's heatmap indicate their proficiency in identifying terminal and small branches, but a reduced accuracy in medium-sized branches and the trachea. Twen's heatmap, with fewer cold regions within lung lobes, signals fewer false negatives, aligning with their notably low airway leakage rate (0.601). Sanmed AI's heatmap displays a mix of warm and cold colours, with deep colouration around the lobar bronchus (black bounding box) indicating comparable airway leakages, while the trachea area (white bounding box) shows significant neglect. An interesting observation is the false positive region (black arrow in the lower left corner) denoting misprediction of scanning artefacts, suggesting an absence of a post-processing algorithm to retain the largest connective component. Conversely, Gexing's model reflects false positives in the trachea and medium-sized branches, while overlooking some small and terminal branches. DJ_92 and Riipl exhibit similar patterns with

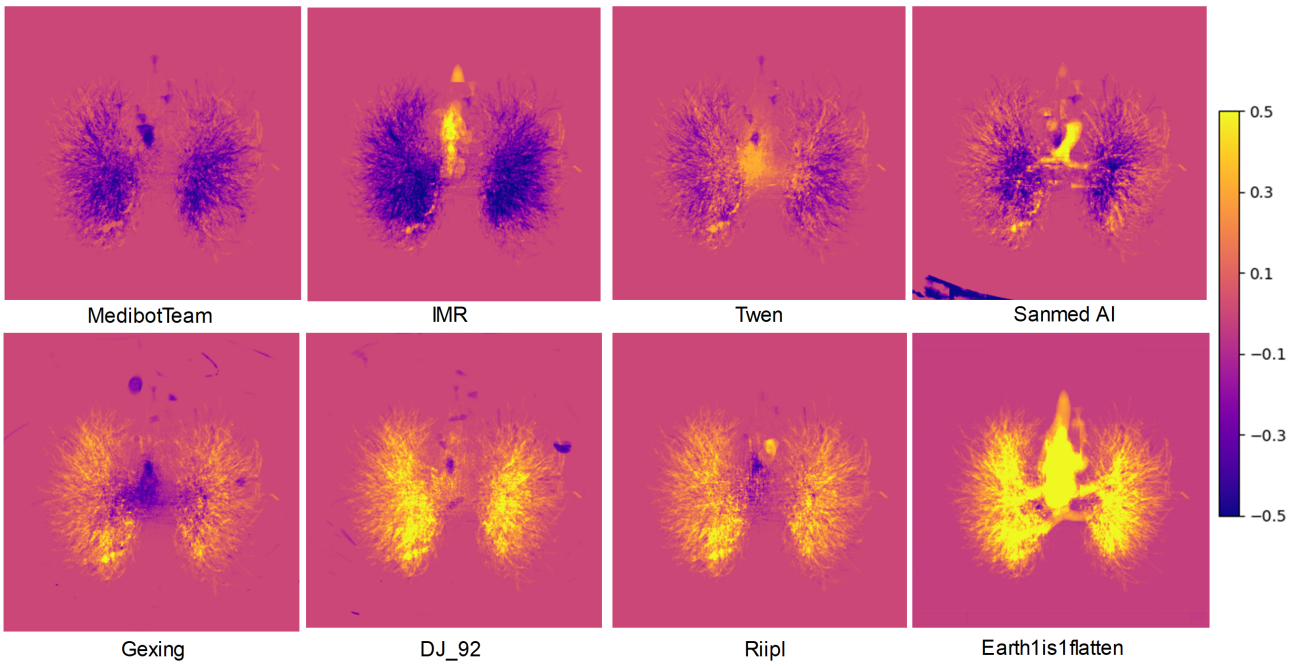


Figure 5.5: Heatmap Showing the Difference Between Ground Truth and Predicted Outcomes on the Accumulated Axis Plane. Warm colors highlight false negative predictions (missed segments), while cold colors point out false positive predictions (extraneous segments).

fewer false positives, correlating with their high precision and low branch scores. Towards the lower end of performance, Earth1is1flatten's heatmaps are dominated by false negatives, with some blocked false positives.

These heatmaps provide a visual narrative that complements the numeric data in Table 4.5, offering a deeper understanding of the strengths and weaknesses of the various models in predicting airway structures.

Rethinking Airway Modelling: Performance on Different-Sized Branches

Airway trees can be categorized into different groups based on their anatomical sizes, such as terminal (TB), small (SB), medium (MB), and large branches (LB). To examine model performances in detail, airway trees were first classified into these groups. The detection rate of each category was assessed according to the methodology outlined in [121] (Table 5.7). Specifically, terminal branches are characterized by a radius of 0-2mm, small branches by 2-4mm, medium branches by 4-8mm, and large branches by a radius greater than 8mm.

Table 5.7 reveals that most models performed better (detection rate >0.95) in identifying medium and large branches compared to small and terminal ones (Wilcoxon signed-rank test,

Table 5.7: Detection rate of different sized branches

Teams	Terminal	Small	Medium	Large
MedibotTeam	0.8351	**0.9820**	**0.9981**	1.0
IMR	**0.8678**	0.9809	0.9925	0.9952
Twen	0.7897	0.9717	0.9936	0.9990
Sanmed AI	0.7078	0.9424	0.9773	0.9829
Gexing	0.5941	0.9585	0.9916	1.0
DJ_92	0.5371	0.9397	0.9903	0.9990
Riipl	0.5705	0.9507	0.9909	0.9963
earth1is1flatten	0.3822	0.8649	0.9429	0.9238
Dolphins	0.5626	0.9523	0.9890	0.9990
Junqiangmler	0.5105	0.9022	0.9688	0.9768

p <0.05). The detection rate for small branches was relatively consistent, ranging from 0.86 to 0.98, while the rate for terminal branches varied significantly, as depicted in Fig. 5.6. For example, IMR attained the highest detection rate for terminal branches at 0.8678, whereas the lowest rate among the top-10 teams was merely 0.3822. These findings suggest that while most deep learning-based methods effectively segment the majority of airway trees in lung fibrosis patients, the identification of terminal branches requires further enhancement. Additionally, despite IMR's high detection rate for the trachea, the model also missed parts of the main bronchus, leading to high false negative predictions in the trachea region, as shown in Fig 5.5.

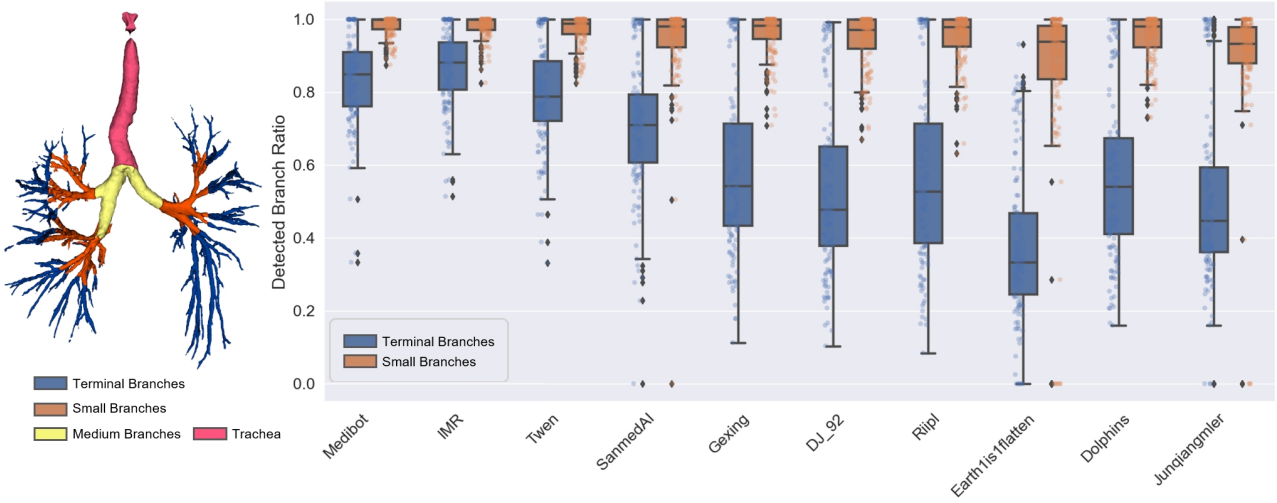


Figure 5.6: Boxplot of the detected branch rate for terminal and small branches.

Model Performance on Unharmonized Data

In clinical practice, High-Resolution Computed Tomography (HRCT) scans are collected across various hospitals, each employing different acquisition protocols, devices, and reconstruction

kernels [14]. To explore model performance on unharmonized data, we tested the top-5 AIIB23 approaches on manually flipped HRCT scans from the fibrosis test set (n=90), flipped along the z, y, and x axes (Table 5.8).

Table 5.8: Model performance on unharmonised data. Colour Indicators for Significant Changes (Wilcoxon signed-rank test, p<0.05) in Performance Compared to Original Test Cases. Green represents positive trends, while red signifies negative trends.

Teams	Data Process	IoU	DLR	DBR	ALR	AMR
MedibotTeam	X-Flip	0.8971	0.9109	0.8700	0.0779	0.0339
	Y-Flip	0.8965	0.9066	0.8635	0.0743	0.0378
	Z-Flip	0.8970	0.8964	0.8509	0.0713	0.0398
IMR	X-Flip	0.8698	0.9255	0.8952	0.1063	0.0389
	Y-Flip	0.8665	0.9283	0.9003	0.1074	0.0418
	Z-Flip	0.8717	0.9079	0.8717	0.0965	0.0454
Twen	X-Flip	0.8944	0.8855	0.8352	0.0605	0.0522
	Y-Flip	0.8906	0.8830	0.8322	0.0571	0.0591
	Z-Flip	0.8919	0.8793	0.8289	0.0562	0.0584
Sanmed_AI	X-Flip	0.8828	0.8418	0.7833	0.0742	0.0526
	Y-Flip	0.3529	0.4931	0.4632	0.2260	0.6275
	Z-Flip	0.5196	0.5893	0.5366	0.0532	0.4571
Gexing	X-Flip	0.8994	0.7897	0.6871	0.0426	0.0625
	Y-Flip	0.8871	0.7918	0.6923	0.0383	0.0791
	Z-Flip	0.8887	0.7851	0.6853	0.0383	0.7750

Among these approaches, team Twen's model demonstrated remarkable robustness, with no significant performance difference between unharmonized and original test data. The Intersection over Union (IoU) scores of the top three teams remained stable, while those of teams Sanmed_AI and Gexing showed fluctuations. Contrary to expectations, the Detected Length Ratio (DLR) and Detected Branch Ratio (DBR) scores for MedibotTeam declined notably, whereas team IMR's scores only decreased significantly in Z-Flipped data. Most models, including MedibotTeam, IMR, Sanmed AI, and Gexing, exhibited a notable reduction in airway leakages on unharmonized data. This suggests potential benefits of test-time augmentation. Sanmed AI, however, experienced a substantial decline in performance on Y-Flip and Z-Flip datasets, with steep drops in IoU (30-50%), DBR, and DLR (around 40%). Team Twen's consistent performance across unharmonized data indicates that appropriate data augmentation techniques could effectively counteract biases arising from data orientation.

Model Performance on Low-Quality Data

Low-quality data, often laden with noise, inconsistencies, missing values, or bias, poses a significant challenge to predictive model accuracy and reliability. Such datasets can lead models to overfit noise, capturing irrelevant patterns that fail to generalize to real-world scenarios. To examine how the top-5 AIIB23 models handle low-quality data, we introduced manual noises and downsampled data, as detailed in Table 5.9. Gaussian noise was randomly added to the original HRCT scans to create noise images, and data was randomly downsampled along the z-axis (slice number reduction ratio $\in [0.5, 1]$) to generate inconsistent data.

Table 5.9: Model performance on low-quality data. Color Indicators for Significant Changes (Wilcoxon signed-rank test, p<0.05) in Performance Compared to Original Test Cases. Green represents positive trends, while red signifies negative trends.

Teams	Data Transfer	IoU	DLR	DBR	ALR	AMR
MedibotTeam	Down sample	0.8889	0.8805	0.8421	0.0824	0.0391
	Noise	0.8952	0.8552	0.7885	0.0636	0.0484
IMR	Down sample	0.8657	0.8955	0.8657	0.1057	0.0440
	Noise	0.8710	0.8787	0.8252	0.0848	0.0558
Twen	Down sample	0.8856	0.8407	0.7943	0.0583	0.0633
	Noise	0.8941	0.8457	0.7746	0.0476	0.0637
Sanmed_AI	Down sample	0.8709	0.8008	0.7482	0.0825	0.0687
	Noise	0.8815	0.7578	0.6662	0.0536	0.0719
Gexing	Down sample	0.8945	0.7548	0.6662	0.0413	0.0687
	Noise	0.8821	0.6949	0.5791	0.0373	0.0851

The results showed a general reduction in branch metrics and increased airway missing rates across all five approaches (p <0.05). Notably, the airway leakage ratio significantly decreased in the noisy dataset for all models, likely due to increased false negative predictions (more terminal branches missed). This prevalent reduction in performance underscores the urgent need for developing robust models capable of effectively handling low-quality data.

Beyond Segmentation: Clinical Value of Airway-Derived Biomarkers

To delve into the clinical relevance of airway-derived biomarkers, a survival analysis was conducted for patients with fibrotic lung diseases. Utilizing data from the Australian IPF Repository (AIPFR), which comprises 203 HRCT scans, we focused primarily on the Total Airway Volume (TAV), a biomarker representing the entire airway tree volume relative to total lung lobe volume. Additionally, probabilities from Task-II teams (top three: uAI, Tastefish, and Junqiangmler) were included in the analysis (denoted as Prob_II1, Prob_II2, Prob_II3).

The main steps of the survival analysis involved:

- Assessing the association of computational biomarkers (TAV, Prob_II1, Prob_II2, Prob_II3) with mortality.
- Analyzing these biomarkers alongside existing imaging biomarkers (SOFIA [41], average fibrosis score, FVC%), using the Cox Proportional-Hazards model in StataMP V17.0, to determine their independent predictive value.

Univariate analysis results (Table 5.10) indicate significant associations of all computational biomarkers with mortality ($p < 0.05$). Bivariate analysis with CPI, total fibrosis extent, FVC%, and SOFIA revealed TAV and Prob_II1 as independent predictors of mortality ($p < 0.05$). Multivariate analysis explored relationships between TAV, Prob_II1, Prob_II2, Prob_II3, and SOFIA. Among these, TAV, Prob_II1, and SOFIA emerged as independent predictors of mortality ($p < 0.05$), while Prob_II2 and Prob_II3 did not demonstrate independent predictive value.

Further multivariate analyses among TAV, ProbII1, SOFIA, AvgFib, and FVC% showed that TAV had a highly significant association with mortality ($p < 0.0001$, HR = 1.83). Prob_II1 indicated some risk with HR = 1.70 but lacked statistical significance ($p = 0.089$). SOFIA showed marginal effects on mortality, while the total fibrosis extent lacked significant independent predictive power. FVC%, however, was statistically significant ($p = 0.025$, HR = 0.35, 95% CI: 0.14-0.88), suggesting a considerable reduction in mortality risk with each unit increase.

Reproducibility of Radiomic Features

Radiomic features play a crucial role in facilitating explainable clinical diagnosis and prognosis tasks, offering a method for the quantitative description of medical images. These features are believed to capture previously unnoticed patterns and characteristics, which may go unnoticed by clinicians. However, existing studies have raised concerns regarding the low reproducibility of radiomic features, and their reliability in the context of lung fibrosis disease remains uncertain. While the overall accuracy (OvAcc) of the top-2 teams in Task I statistically demonstrates similarity ($p > 0.05$), it is unclear whether the radiomic features extracted by these teams are

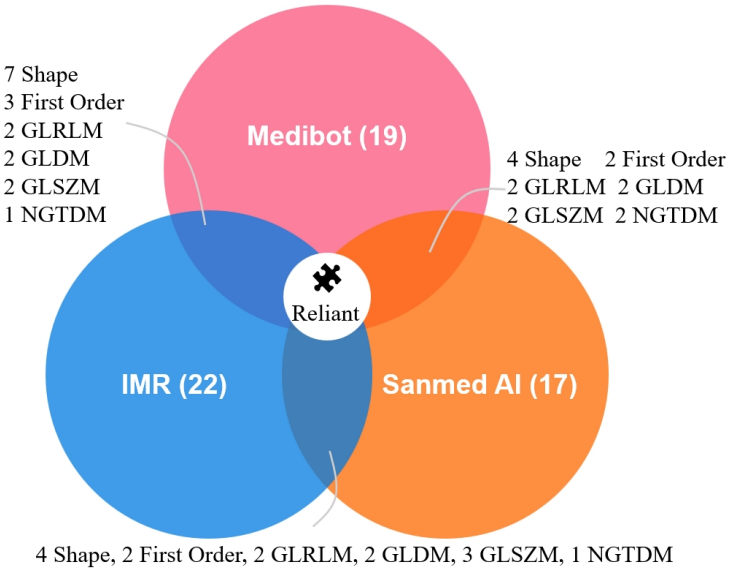
Table 5.10: Survival analysis for image-derived biomarkers

Methods	Variables	P Value	HR (95% CI)	Variables	P Value	HR (95% CI)
Univariate (N=203)	TAV	<0.0001	2.58 (2.05-3.24)	Prob_II2	0.003	1.80 (1.22-2.66)
	Prob_II1	<0.0001	3.06 (1.86-5.05)	Prob_II3	0.008	2.04 (1.21-3.43)
Bivariate with AvgFib (N=200)	Prob_II1	0.037	1.83 (1.04-3.24)	Prob_II3	0.025	1.84 (1.08-3.15)
	AvgFib	<0.0001	1.04 (1.02-1.05)	AvgFib	<0.0001	1.04 (1.03-1.06)
	Prob_II2	0.156	1.34 (0.89-2.00)	TAV	<0.0001	2.10 (1.57-2.82)
	AvgFib	<0.0001	1.04 (1.03-1.06)	AvgFib	0.001	1.03 (1.01-1.04)
Bivariate with FVC% (N=203)	Prob_II1	<0.0001	2.52 (1.50-4.24)	Prob_II3	0.195	1.43 (0.83-2.47)
	FVC%	<0.0001	0.19 (0.08-0.45)	FVC%	<0.0001	0.17 (0.07-0.43)
	Prob_II2	0.029	1.55 (1.05-2.31)	TAV	<0.0001	2.35 (1.80-3.07)
	FVC%	<0.0001	0.18 (0.08-0.42)	FVC%	0.004	0.27 (0.11-0.66)
Bivariate with SOFIA(N=203)	Prob_II1	<0.0001	2.59 (1.52-4.41)	Prob_II3	0.057	1.67 (0.98-2.85)
	SOFIA	<0.0001	1.36 (1.20-1.55)	SOFIA	<0.0001	1.38 (1.22-1.57)
	Prob_II2	0.104	1.40 (0.93-2.10)	TAV	<0.0001	2.19 (1.67-2.89)
	SOFIA	<0.0001	1.37 (1.20-1.56)	SOFIA	0.001	1.26 (1.10-1.44)
Methods	Variables		P Value		HR (95% CI)	
Multivariate (n=203)	TAV		<0.0001		2.10 (1.57-2.81)	
	Prob_II1		0.007		2.22 (1.24-3.99)	
	Prob_II2		0.735		0.924 (0.59-1.46)	
	Prob_II3		0.471		1.24 (0.69-2.26)	
	SOFIA		0.005		1.23 (1.07-1.42)	
Multivariate (n=200)	TAV		<0.0001		1.83 (1.35-2.49)	
	Prob_II1		0.089		1.70 (0.92-3.13)	
	SOFIA		0.052		1.17 (1.00-1.38)	
	AvgFib		0.163		1.01 (0.99-1.03)	
	FVC%		0.025		0.35 (0.14-0.88)	

consistent. This section begins by analyzing the variance of radiomic features across different AI models and subsequently reports on the relevant radiomic features for clinical prognostication.

We initiated our analysis by examining the significance of radiomic features extracted from MedibotTeam, IMR, and Sanmed AI (the 1st, 2nd, and 4th teams in Task I, respectively). Following this, we conducted a Cox Hazard analysis to identify variables associated with mortality. A total of 107 radiomic features were extracted for each case, encompassing 14 shape features, 18 first-order features, 24 Grey Level Co-occurrence Matrix features (GLCM), 14 Grey Level Dependence Matrix (GLDM), 16 Grey Level Run Length Matrix (GLRLM), 16 Grey Level Size Zone Matrix (GLSZM), and 5 Neighbouring Grey Tone Difference Matrix Features (NGTDM). Intriguingly, most of these features (>86) exhibited significant differences (Wilcoxon signed-rank test, $p < 0.05$) among different models, despite the absence of statistical differences ($p < 0.05$) in overall accuracy between MedibotTeam and IMR. Out of the 107 features, 19, 22, and 17 were found to be associated with mortality ($p < 0.05$ in the univariate Cox proportional

hazards Model), with 13 features common to all three models (Fig. 5.7).



	Reliant radiomic features
Shape	Voxel volume, Surface area, Elongation, Mesh volume
First Order	Energy, Total energy,
GLRLM	Run Length nonuniformity, Gray level nonuniformity
GLSZM	Gray level nonuniformity, Size zone nonuniformity
GLDM	Dependence nonuniformity, Gray level nonuniformity,
NGTDM	Coarseness

Figure 5.7: Reliant radiomic features in survival status prediction.

Limitations and Future Research Directions

There are several key areas where this research can be extended and improved. Firstly, the central line extraction algorithm used in airway modelling has limitations, particularly in dealing with the intricate structures of airway trees. These algorithms often result in errors, such as the overestimation of branch counts. Future research should focus on refining central line extraction methods to enhance accuracy. Secondly, the robustness of existing models needs attention. As demonstrated in Tables 6 and 7, many current approaches struggle when confronted with noisy and low-resolution CT scans. Developing models with improved generalization capabilities, capable of handling unharmonized and low-quality data effectively, is a pressing need. Lastly, the integration of clinical analysis into airway modelling remains underexplored. Bridging the gap between AI-based airway modelling and practical clinical applications can significantly enhance the utility of this technology in healthcare settings.

5.6 Conclusion

In conclusion, this research has highlighted the dominance of CNN-based methods in medical image segmentation, particularly in the context of airway modelling. These methods have

proven highly effective in accurately segmenting complex airway structures. Additionally, the study emphasized the importance of incorporating explainable image-derived features for mortality prediction, shedding light on valuable indicators for patient outcomes. High-performing teams achieved remarkable results in both airway modelling and mortality prediction tasks, showcasing the potential of AI in medical applications. Furthermore, there is a clear need to enhance model capacity, address robustness challenges, and explore the clinical implications of airway modelling for future research in this field.

Chapter 6

A Clinical Translation Study: Prognostication of patients with fibrotic lung disease

6.1 Abstract

Research indicates that the extent of traction bronchiectasis effectively forecasts the progression and outcomes of diseases. This study introduces a novel automated deep learning tool, SABRE (Smart Airway Biomarker Recognition Engine), designed to evaluate the severity of traction bronchiectasis in patients with idiopathic pulmonary fibrosis (IPF), using data from a national IPF registry. SABRE demonstrates its prowess in precisely measuring this severity, establishing itself as a reliable independent marker for mortality, surpassing traditional methods. The study highlights the significance of traction bronchiectasis's severity as a strong mortality predictor, underscoring its usefulness in monitoring disease progression and assisting in patient risk classification. These findings underscore the importance of SABRE as a prognostic tool, aiding clinicians in identifying patients at increased risk of rapid disease advancement and guiding treatment decisions, particularly concerning antifibrotic therapy.

6.2 Introduction

The approval of antifibrotic therapies for idiopathic pulmonary fibrosis (IPF) and their subsequent application in other progressive fibrotic lung diseases has highlighted the urgent need for reliable biomarkers. These biomarkers are crucial for accurately forecasting the course of the disease and the patient's response to treatment. The variable nature of fibrotic lung diseases, where some patients remain stable while others deteriorate rapidly, complicates treatment decisions, including the necessity for invasive procedures like surgical lung biopsies. The ability to predict disease trajectory using initial clinical data would enable healthcare providers to more effectively target antifibrotic therapy to those who would benefit most, and avoid unnecessary treatment in patients with naturally stable conditions.

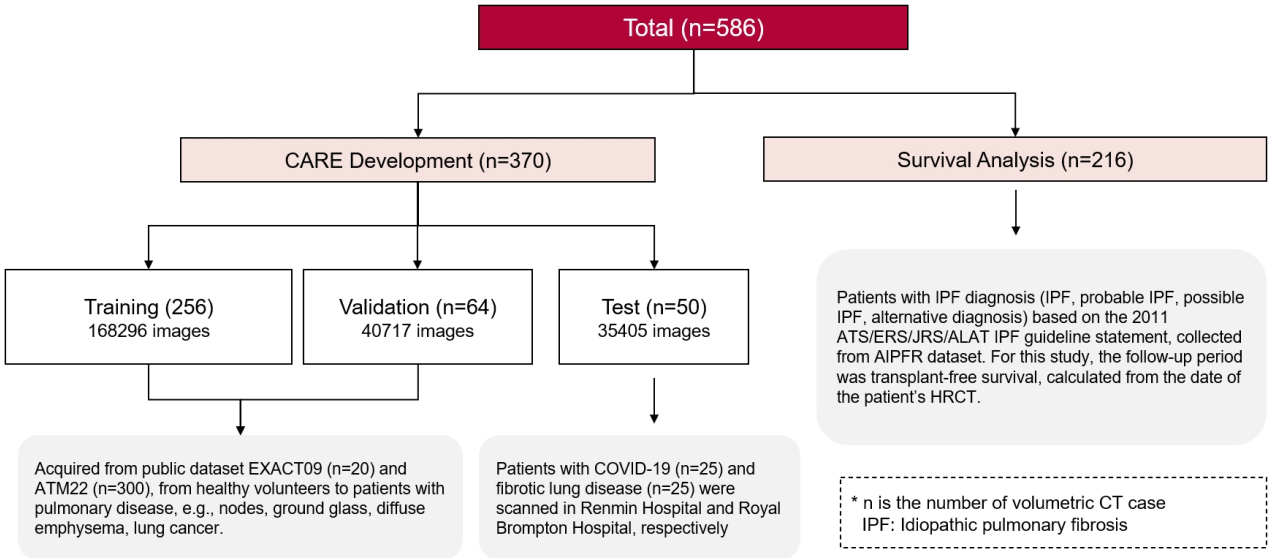


Figure 6.1: Data distribution of cohorts in this study.

In the last decade, there has been a growing interest in computer-based techniques for objectively measuring fibrotic lung disease severity through high-resolution chest CT (HRCT) scans. This interest stems from the limitations of traditional visual HRCT assessments, including their low reproducibility, high observer variability, and limited sensitivity. Visual quantification of traction bronchiectasis, characterized by the abnormal widening of airways in fibrotic areas, has been consistently recognized as a strong mortality indicator in fibrotic lung diseases. However, developing computer-based methods for accurately quantifying traction bronchiectasis severity has proven challenging, particularly in differentiating airways from honeycombing or

emphysema. This study assesses the prognostic value of a new automated deep-learning model designed to quantify the severity of traction bronchiectasis (the cohorts used in this study are shown in Fig 6.1). We apply this model to data from a national IPF registry, aiming to improve the precision of prognosis and treatment decision-making in fibrotic lung diseases.

6.3 Methods

6.3.1 Australian IPF Registry Data.

This study evaluates the effectiveness of SABRE as a tool for predicting outcomes in fibrotic lung diseases using data from the Australian IPF Registry (AIPFR), detailed in Table 6.1. The Sydney Local Health District has granted ethical approval for the AIPFR under protocol number X14-0264. Since its inception in 2012, the AIPFR's processes and goals have been previously documented [188]. In summary, patients diagnosed with IPF by their treating physicians are registered. For each participant in the initial recruitment phase, a panel comprising three expert radiologists, three histopathologists, and three specialist ILD physicians reviews the clinical, radiological, and, when available, histological data. They then assign a diagnosis (IPF, probable IPF, possible IPF, or an alternative diagnosis) following the criteria of the 2011 ATS/ERS/JRS/ALAT IPF guideline statement [180]. Any diagnostic disagreements are settled through panel discussions.

The registry collects both baseline and ongoing data for as long as participants are enrolled. In this specific study, the main focus was on transplant-free survival, starting from the date of the patient's high-resolution chest CT (HRCT). Disease progression within 12 months was identified based on one or more of the following criteria occurring within a year of the HRCT: (A) a sustained decline of 10% or more in the predicted percentage of Forced Vital Capacity (FVC) or 15% or more in the predicted percentage of Diffusing Capacity of the Lungs for Carbon Monoxide (DLCO) at 18 months, (B) death, or (C) lung transplantation.

Table 6.1: Baseline Demographics and Centralised MDM Review Characteristics. * indicates multidisciplinary diagnoses.

Variable	
Age (Median (IQR))	69.7 (64.9 – 76.1)
Gender (Female/Male)	168 (32.6%) / 347(67.4%)
Lung function	
FVC% (Median (IQR))	79.6 (66.1 – 93.5)
DLCO% (Median (IQR))	46.9 (37.9 – 58.7)
CPI (Median (IQR))	46.1 (36.8 – 54.8)
Anti-Fibrotic Therapy (No/Yes)	337 (65.4%) / 178 (34.6%)
Smoking History (Ever/Never)	333 (64.7%) / 182 (35.3%)
IPF Diagnosis Category (ATS 2011)*	
Definite IPF	249 (48.3%)
Probable IPF	26 (5.0%)
Possible IPF	110 (21.4%)
Inconsistent IPF	126 (24.5%)
No consensus reached	4 (0.8%)
UIP Diagnosis Category (ATS 2018)**	
(Radiologist consensus, n=515)	
Definite UIP	136 (26.4%)
Probable UIP	167 (32.4%)
Indeterminate UIP	84 (16.3%)
Alternative diagnosis	128 (24.9%)
Histology (n=86)	
Definite UIP	53 (61.6%)
Probable UIP	16 (18.6%)
Possible UIP	11 (12.8%)
Alternative to UIP	6 (7.0%)

6.3.2 ATM22 Data

SABRE, the Smart Airway Biomarker Recognition Engine, is an innovative AI-based system developed using data from the Airway Tree Modelling 2022 (ATM22) project [174]. The ATM22 dataset, first released on the Grand Challenge platform in conjunction with MICCAI 2022 (https://atm22.grand-challenge.org/), comprises 300 cases from various sources including the EXACT'09 public dataset, the LIDC-IDRI data repository, and Shanghai Chest Hospital. This collection includes CT scans from both healthy individuals and patients with diverse pulmonary conditions, such as pulmonary nodules, ground glass opacities, and lung cancer. These scans

were obtained from multiple centres using different scanning equipment, patient positions (either at full inspiration or expiration), and acquisition protocols.

The initial annotation of airway structures in these scans was performed by three junior radiologists using ITK-SNAP software [189], followed by a review and consensus resolution of any discrepancies by two senior radiologists.

6.3.3 SABRE- Smart Airway Biomarker Recognition Engine

SABRE integrates a fully connected fuzzy attention neural network and a smart branch decomposition module for its operation. For its development, 300 CT scans from ATM22 were utilized: 256 for training and 64 for validation. Additionally, to assess the model's effectiveness on external data, 25 HRCT scans from the Australian IPF Registry (patients with fibrotic lung disease) and 25 HRCT scans from Royal Brompton Hospital (patients with COVID-19) were used as a test set.

This system employs a 'split-merge' approach for training and inference with 3D HRCT scans, breaking down the scans into smaller patches and then reassembling the predictions into full volumetric masks. An optimization function incorporating the airway centerline was included to improve prediction accuracy, especially for terminal branches. SABRE was developed using Pytorch 1.9, CUDA 11.1.74, and cuDNN 8.0.5, operating on an NVIDIA RTX 3090 on an Ubuntu 20.04.2 LTS platform.

SABRE's methodology involves identifying the trachea and central point in 3D data, followed by projecting the 3D airway prediction onto two-dimensional images. The system first estimates the height of the airway trees, and then calculates normalized Euclidean distances for airway segmentation. Preliminary segments are identified using distance percentiles, which are then refined and computed in 3D, resulting in various segmented airway volumes normalized against the total lung volume. Further technical details, including image processing, training strategies, and evaluation metrics, are provided in the [121].

SABRE initially identifies airway segments by pinpointing the trachea and a central spatial

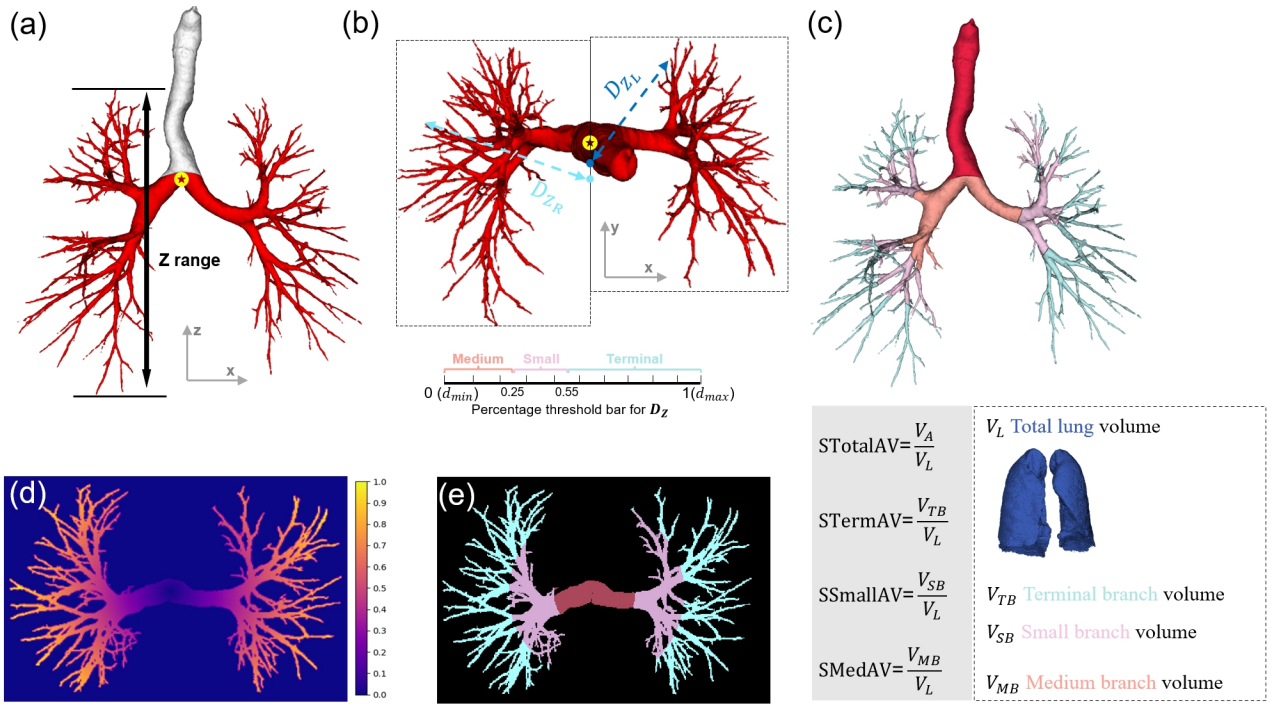


Figure 6.2: Approach used for obtaining airway segments.

point (marked as a yellow star in Figure 6.2) within 3D volumetric data. This is followed by a projection of the 3D airway model onto two-dimensional images along the y-axis (P_Y, as shown in 6.2(a)) and z-axis (P_Z, as shown in 6.2 (b)). The process begins by estimating the height of the airway trees, excluding the trachea. It then calculates normalized Euclidean distances, which vary from 0 to 1, between each pixel representing the airway in the z-axis projection and its associated central points (indicated by cyan dots for the right bronchus trees and blue dots for the left bronchus trees).

Using a distance map, initial segments are established based on thresholds determined by distance percentiles; specifically, the terminal, small, and medium branches are identified at 0.25 and 0.55 percentiles, respectively. These initial segments are re-projected into the 3D volumetric model and further refined based on the bronchus trees' distance constraints in the P_Y projection. Finally, specific airway segments are calculated and normalized against the total lung volume, resulting in four distinct measurements: SABRE-based terminal airway volume (STermAV), SABRE-based small airway volume (SSmallAV), SABRE-based medium airway volume (SMedAV), and the total airway volume (STotalAV).

Semiquantitative HRCT evaluation.

The semiquantitative assessment of High-Resolution Computed Tomography (HRCT) involves specific pattern definitions and scoring methods detailed in an online repository. In summary, two experienced thoracic radiologists, with 10 and 12 years of experience respectively, evaluate HRCTs for each patient. They assess the overall extent of Interstitial Lung Disease (ILD) and measure the extent and severity of four specific interstitial patterns: ground glass opacification, reticulation, honeycombing, and consolidation. They also evaluate the presence and severity of emphysema and traction bronchiectasis.

For each HRCT, a diagnostic probability is assigned to each of the radiologic diagnosis categories outlined by the 2018 Clinical Practice Guideline for Idiopathic Pulmonary Fibrosis (IPF). These probabilities, which total to 100% and are censored at a 5% minimum, include categories such as "Usual Interstitial Pneumonia (UIP): 75%", "Probable UIP: 25%", "Indeterminate for UIP: 0%", and "Alternative diagnosis: 0%". This structured approach ensures a thorough and standardized evaluation of HRCTs in the context of interstitial lung diseases.

Statistical Analysis.

Data presentation in this study adheres to standard statistical practices, using median or mean along with standard deviations to represent central tendencies and variability, depending on the distribution of data. Additionally, counts and percentages are used to describe categorical data. The Chi-squared test is employed for analyzing differences between categorical variables, while a two-sample t-test is used for parametric continuous variables. The threshold for statistical significance is established at a p-value of less than 0.05. The performance of the SABRE airway modelling technique is compared with other existing methods such as nnUNet [74], WingsNet [87], ERNet [103], Rose [105], FANN [121], etc., utilizing the Wilcoxon signed-rank test for analysis. Further, the interobserver agreement among radiologists on AIPFR airway tree annotations is measured using the Kendall correlation coefficient (KCC). This involves converting 3D binary annotation masks into 1D arrays and calculating KCC between pairs of

annotators.

In exploring the relationship between disease severity and airway volume indices, linear regression is applied. Cox proportional hazard modeling is then used to ascertain both crude and adjusted hazard ratios, focusing on transplant-free survival as the outcome. The survival period is calculated from the registry HRCT date to a specific endpoint. To ensure the assumptions of proportional hazards are met, various methods such as log-log plot inspection, comparison of Kaplan-Meier and Cox predicted curves, and Schoenfeld residuals analysis (not detailed here) are utilized. Finally, all results, including hazard ratios, 95% confidence intervals, and p-values, are meticulously reported. The statistical analysis is rigorously performed using Stata 17.0 MP Edition and Python SciPy 1.6.2, ensuring accuracy and reliability in the findings.

Table 6.2: Patient characteristics for those with an HRCT amenable to SABRE versus those whose HRCT was not.

Variable	SABRE		p-value
	possible (n=217)	not possible (n=298)	
Age	70.9	69.6	0.9663
Gender M:F (347:168)	140:77	207:91	0.1189
FVC (204:285)	80.5	80.3	0.5428
FEV1 (204:285)	83.1	84.2	0.2664
DLco (173:248)	48.9	49.2	0.4220
CPI (171:240)	45.1	45.4	0.4319
Radiologist-based TBX score (214:296)	1.75	1.73	0.5737
Fibrosis extent (214:296)	23.9	22.6	0.8663

Statistical Modelling strategy.

In the study's multivariable analyses, patient age and gender were consistently accounted for as controlling variables. For survival analyses, three distinct indices were employed to adjust for the varying severity of the disease: the semiquantitative total fibrosis extent score, which provides a detailed measure of fibrosis coverage; the DeepDTA fibrosis scores, a more nuanced metric derived from advanced imaging techniques; and the Composite Physiological Index (CPI), which amalgamates various lung function variables to gauge physiological impairment attributable to Interstitial Lung Disease (ILD). The CPI is particularly noted for its ability to reflect the impact of ILD while excluding the confounding effects of emphysema. Given that em-

physema can influence the apparent severity of traction bronchiectasis, the study also includes separate multivariate survival analyses specifically adjusted for the extent of emphysema. This ensures a more refined and accurate understanding of the disease's impact on patient survival, taking into account the complex interplay between ILD, fibrosis, and emphysema.

6.4 Results

6.4.1 SABRE analysis on the AIPFR.

Out of 868 patients, 515 patients had High-Resolution Computed Tomography (HRCT) scans suitable for analysis. The study observed high interobserver agreement among radiologists in annotating airway trees, as evidenced by a near-perfect Cohen's Kappa value of 0.937. Specifically, the agreement for terminal branches was substantial (Cohen's Kappa K= 0.742), while for small branches, it was nearly perfect (Cohen's Kappa K= 0.852). Of the initial patient group, 298 were excluded due to insufficient image slices for SABRE analysis, leaving 217 patients with HRCT scans suitable for such analysis. Comparative analysis showed no significant differences between the volumetric HRCT group and those excluded due to inadequate HRCTs. SABRE demonstrated enhanced performance relative to other state-of-the-art models, evidenced by an average detected branch ratio of 0.943±0.056 and an average detected length ratio of 0.959±0.037 in the AIPFR test set (n=90). These results in Table 6.3 highlight the tool's proficiency in closely matching AI predictions with clinicians' annotations and its

Table 6.3: Segmentation performance of SABRE, † indicates significant differences ($p<0.05$) evaluated by Wilcoxon signed-rank test.

Models	**IoU**	**DLR**	**DBR**	**ALR**	**AMR**
nnUNet	0.924±0.049†	0.748±0.179†	0.688±0.204†	0.005±0.006†	0.072±0.052†
WingsNet	0.786±0.085†	0.810±0.143†	0.768±0.172†	0.046±0.018†	0.178±0.089†
ERNet	0.845±0.096	0.704±0.189†	0.638±0.199†	0.046±0.036†	0.117±0.093†
ROSE	0.852±0.083†	0.755±0.156†	0.687±0.169†	0.077±0.077†	0.087±0.071†
CENet	0.824±0.104†	0.639±0.204†	0.563±0.204†	0.043±0.023†	0.140±0.113†
FANN	0.875±0.058†	0.861±0.099†	0.818±0.120†	0.087±0.045†	0.052±0.035†
SABRE	0.848±0.041	0.959±0.037	0.943±0.056	0.124±0.048	0.048±0.027

effective extraction of primary airway branches.

6.4.2 Univariate analysis of airway indicators.

Table 6.4: Univariable analysis Cox proportionate hazards models for airway segments

Variable	HR	P Value	CI 95%
STotalAV (n=217)	3.46	<0.0001	2.47-4.85
SMedAV (n=217)	51.59	<0.0001	14.09-188.89
SSmallAV (n=217)	32.96	<0.0001	10.94-99.30
STermAV (n=217)	9.74	<0.0001	5.16-18.37

In the study's univariate analysis (Table 6.4), different airway volumes measured by SABRE, namely total airway volume (STotalAV), medium airway volume (SMedAV), small airway volume (SSmallAV), and terminal airway volume (STermAV), were all predictors of mortality. These volumes continued to predict mortality even after adjusting for age and gender (Table 6.5). The relationships between medium, small, and terminal airway volumes were further

Table 6.5: Multivariate survival analysis of SABRE-based variables controlled for age and gender.

Variable	HR	p-value	95% CI
STotalAV (n=217)	3.24	<0.0001	2.29-4.59
SMedAV (n=217)	39.05	<0.0001	10.10-150.92
SSmallAV (n=217)	28.00	<0.0001	9.15-85.68
STermAV (n=217)	9.04	<0.0001	4.66-17.55

explored. It was found that STermAV had a significant correlation with SSmallAV (R^2=0.69, p<0.0001), and similarly, SSmallAV correlated with SMedAV (R^2=0.44, p<0.0001). However, when these variables were examined together in a multivariate analysis, only STermAV showed a trend towards statistical significance (shown in Table 6.6). To refine the analysis, SSmallAV

Table 6.6: Multivariate analysis including all airway segment volumes.

Variable	HR	p-value	95% CI
SMedAV (n=217)	3.821	0.172	0.558-26.174
SSmallAV (n=217)	3.418	0.221	0.477-24.480
STermAV (n=217)	3.337	0.051	0.993-11.216

and STermAV were combined into a single variable, referred to as the SABRE-based primary

airway volume (SPAV). This combined variable was then used in subsequent analyses to further investigate its predictive power on patient outcomes.

6.4.3 Multivariate analysis against disease severity.

The correlation analysis (in Table 6.7) explored the relationships between visual assessments of total fibrosis extent, the automated DeepDTA fibrosis score, and the Composite Physiological Index (CPI), revealing significant correlations. The CPI was found to correlate with the total fibrosis extent (R^2=0.21, p<0.0001) and had a stronger correlation with the DeepDTA fibrosis score (R^2=0.36, p<0.0001). Additionally, there was a robust correlation between the total fibrosis extent and the DeepDTA fibrosis score (R^2=0.49, p<0.0001). Further analysis examined

Table 6.7: Relationship between disease severity (using total fibrosis extent scores, DeepDTA score and the CPI) and SABRE-based airway segment volumes.

Variable	**Total fibrosis extent (n=214)**	**DeepDTA score (n=164)**	**CPI (n=171)**
STotalAV	$R^2 = 0.26, p < 0.0001$	$R^2 = 0.32, p < 0.0001$	$R^2 = 0.14, p < 0.0001$
SMedAV	$R^2 = 0.17, p < 0.0001$	$R^2 = 0.21, p < 0.0001$	$R^2 = 0.16, p < 0.0001$
SSmallAV	$R^2 = 0.21, p < 0.0001$	$R^2 = 0.22, p < 0.0001$	$R^2 = 0.09, p = 0.0001$
STermAV	$R^2 = 0.25, p < 0.0001$	$R^2 = 0.32, p < 0.0001$	$R^2 = 0.11, p < 0.0001$
SPAV	$R^2 = 0.25, p < 0.0001$	$R^2 = 0.31, p < 0.0001$	$R^2 = 0.11, p < 0.0001$

correlations between the SABRE-based primary airway volume (SPAV) and the total fibrosis extent, DeepDTA fibrosis score, and CPI. The SPAV showed a correlation with total fibrosis extent (R^2=0.25, p<0.0001), DeepDTA fibrosis score (R^2=0.31, p<0.0001), and CPI (R^2=0.10, p<0.0001). An increasing SPAV score was found to be associated with a higher risk of mortality, and this association remained significant independent of disease severity. Similar multivariate analyses of individual airway segments also indicated an association with increasing mortality risk (in Table 6.8). Moreover, when controlling for the effect of emphysema, both the SPAV and individual airway segment metrics demonstrated enhanced associations with mortality, underscoring the importance of considering emphysema's impact in these analyses.

Table 6.8: Multivariate survival analysis of each of SABRE-based airway segment volumes with disease severity

	Without Emphysema			With Emphysema		
Variable	HR	P-Value	CI 95%	HR	P-Value	CI 95%
STermAV (n=164)	4.18	0.001	1.86-9.36	5.77	<0.0001	2.51-13.27
CPI	1.05	<0.0001	1.04-1.07	1.05	<0.0001	1.03-1.07
STermAV (n=162)	2.51	0.051	1.00-6.33	3.41	0.01	1.35-8.62
DeepDTA fibrosis score	1.03	<0.0001	1.02-1.04	1.03	<0.0001	1.01-1.04
STermAV (n=214)	4.24	<0.0001	2.00-8.95	5.22	<0.0001	2.49-10.97
Total fibrosis extent	1.04	<0.0001	1.02-1.05	1.04	<0.0001	1.02-1.06
SSmallAV (n=164)	10.3	<0.0001	2.83-37.48	22.08	<0.0001	5.65-86.25
CPI	1.06	<0.0001	1.04-1.08	1.06	<0.0001	1.04-1.07
SSmallAV (n=162)	7.98	0.007	1.75-36.48	17.67	<0.0001	3.71-84.25
DeepDTA fibrosis score	1.03	<0.0001	1.02-1.04	1.03	<0.0001	1.02-1.04
SSmallAV (n=214)	10.47	<0.0001	3.00-36.57	19.10	<0.0001	5.32-68.57
Total fibrosis extent	1.4	<0.0001	1.02-1.06	1.04	<0.0001	1.02-1.06
SMedAV (n=164)	4.45	0.058	0.95-20.88	8.63	0.008	1.73-42.94
CPI	1.06	<0.0001	1.04-1.08	1.56	<0.0001	1.04-1.08
SMedAV (n=162)	2.11	0.382	0.40-11.26	4.22	0.106	0.74-24.11
DeepDTA fibrosis score	1.03	<0.0001	1.02-1.05	1.03	<0.0001	1.02-1.04
SMedAV (n=214)	12.07	0.001	2.94-49.46	23.79	<0.0001	5.59-101.24
Total fibrosis extent	1.04	<0.0001	1.03-1.06	1.04	<0.0001	1.03-1.06
SPAV (n=164)	2.70	<0.0001	1.59-4.58	3.56	<0.0001	2.05-6.20
CPI	1.05	<0.0001	1.04-1.07	1.05	<0.0001	1.03-1.07
SPAV (n=162)	2.06	0.019	1.13-3.77	2.66	0.002	1.44-4.91
DeepDTA fibrosis score	1.03	<0.0001	1.02-1.04	1.03	<0.0001	1.01-1.04
SPAV (n=164)	2.72	<0.0001	1.66-4.49	3.29	<0.0001	2.00-5.41
Total fibrosis extent	1.04	<0.0001	1.02-1.05	1.04	<0.0001	1.02-1.05

6.5 Discussion and Conclusion.

This study validates the prognostic value of a deep learning model, SABRE, in quantifying airway volume and, by extension, the severity of traction bronchiectasis in patients with suspected IPF. Its predictions were found to be significantly independent of traditional disease severity

measures, offering additional insights into patient prognosis.

In the realm of fibrotic lung diseases, the challenge of predicting and monitoring disease progression is paramount, particularly for conditions like progressive pulmonary fibrosis. Although anti-fibrotic therapies have shown benefits, the dilemma remains in the timely initiation of treatment. Traditional markers like FVC are helpful but have limitations, prompting a shift towards utilizing HRCT for more precise digital biomarkers. Despite the demonstrated prognostic value of semiquantitative HRCT assessments in various fibrotic lung diseases over two decades, their integration into clinical and trial practices remains limited due to subjectivity, variability among observers, and relative insensitivity to minor disease progressions. SABRE aims to overcome these limitations by offering an objective quantification of disease on HRCT. It outperformed existing models and provided enhanced prognostic utility compared to traditional radiologist-based scores.

The study further established the significance of traction bronchiectasis severity as a robust predictor of mortality, independent of overall disease severity. This severity potentially reflects the progressive nature of fibrosis, thereby presenting traction bronchiectasis as more than just a marker of disease extent. This insight could be crucial for early intervention strategies, particularly in identifying patients for antifibrotic treatments before significant deterioration.

The model's robustness is underscored by its effectiveness across diverse HRCT scans and protocols, as evidenced by its application in a national IPF registry. However, the study acknowledges limitations, including reliance on volumetric HRCTs and the need for standardized HRCT protocols for consistent and widespread application of advanced HRCT analysis like SABRE. Future directions might include automated evaluation of airway volume changes over time for even more precise patient stratification and monitoring in clinical trials and treatment regimens.

In conclusion, the study presents SABRE as a valuable tool for quantifying traction bronchiectasis severity, offering significant prognostic insights into IPF patient management and highlighting its potential for broader application in clinical and research settings.

Chapter 7

Data Harmonisation, Retrieval and Future Perspectives

7.1 Data harmonisation in digital healthcare

7.1.1 Introduction

Computational biomedical research plays a pivotal role in advancing digital healthcare and biomedical studies. Its primary objective is to develop computational models that enhance the accuracy of disease diagnosis, analyze gene expressions, and handle time series data like electroencephalograms and electrocardiograms. These models are instrumental in uncovering new risk biomarkers, predicting disease progression, optimizing treatment strategies, and identifying potential drug targets. They find applications across various domains, including cancer research, pulmonary disease management, and neurological disorder studies. While effective computational models should exhibit high performance, robustness, explainability, and reproducibility, they often encounter challenges arising from data distribution biases in large-scale studies. Data harmonization is a crucial aspect of working with medical data, with automated methods commonly employed for medical images and gene expression analysis. However, other modalities are often overlooked or harmonized manually. Machine learning approaches, partic-

ularly deep neural networks, have been found to be highly sensitive to training data distribution. Therefore, there is a pressing need to develop techniques that can effectively integrate device/site-invariant information from multiple datasets.

To address this issue, researchers have taken steps to establish standard acquisition protocols [191–193] and definitions [194,195] to facilitate the collection of standardized data. For example, recommendations for acquisition protocols have been made for various imaging techniques like F-FDG Positron emission tomography/computerized tomography (PET/CT) and MR imaging in the context of multiple sclerosis [192, 193]. Additionally, efforts have focused on integrating data from routine health information systems, involving manual harmonization and rule-based alignment of electronic data [194, 195]. While these acquisition protocols have been successful in reducing cohort bias in cross-scanner/site data, they are primarily suited for retrospective studies and may not accommodate prospective research or personalized digital healthcare needs.

Therefore, it is imperative to explore computational methods for harmonizing multicenter datasets, providing flexibility and effectiveness in handling diverse data sources and enabling the advancement of digital healthcare and biomedical research.

Table 7.1: Existing data harmonisation reviews.

Survey	[15]	[196]	[197]	[198]	Ours
Period	∼2020	∼2019	∼2020	∼2021	∼2021
# of reviewed studies	N/A	23	49	42	96
Domain	Radiomics	Radiomics	Radiomics	Radiomics	Radiomics, Gene, Pathology
Metric	x	x	x	x	✓
Checklist	x	x	x	x	✓
Guidance	x	✓	x	x	✓
Meta-analysis	x	✓	x	x	✓

The field of computational data harmonization in digital healthcare has seen some surveys [15, 196, 199] addressing specific aspects, such as MRI [200] or CT harmonization. However, these surveys have often been limited in scope, focusing on single modality or application areas, lacking comprehensive evaluation metrics, and offering limited guidance for researchers (Table 7.1). Furthermore, a standardized checklist for reporting and guiding methodology selection in computational data harmonization studies has been notably absent. This survey aims to fill these gaps by providing a comprehensive exploration of computational data harmonization

strategies in the context of multimodal data in digital healthcare.

This survey covers three primary areas: gene expression, radiomics, and pathology, encompassing over 96 qualified papers published within two decades. To facilitate better scientific practice in the data harmonization community, a comprehensive checklist is proposed. This checklist guides researchers in reporting their studies effectively, enhancing the clarity of explorations (the strategy) and advances (performance metrics) in their research. The main contributions of this survey include a three-fold taxonomy categorizing methodology, evaluation, and applications of computational data harmonization strategies, the introduction of a checklist for future data harmonization studies, and a critical examination of existing strategies and potential areas for improvement.

The remainder of the manuscript is organized as follows: Section 6.1.2 discusses the definition, motivation, utilization, and proposed solutions for computational data harmonization issues; Section 6.1.3 outlines the methodology used in conducting this survey; Sections 6.1.4, 6.1.5, and 6.1.6 present the three-fold taxonomy of harmonization strategies; Section 6.1.7 presents the results of the meta-analysis and introduces the checklist for data harmonization studies; Section 6.1.8 provides an in-depth examination of the checklist and summarizes critiques and limitations of current strategies; and finally, Section 6.1.9 concludes this survey, summarizing the key findings and contributions.

7.1.2 Computational data harmonisation: definition, origin, what for and how?

In this section, we delve into the intricacies of data harmonization, offering insights into its definition, origin, purpose, and potential solutions within the realm of computational data harmonization tasks. To facilitate a clearer understanding of these attributes, we provide a comprehensive terminology guide for computational data harmonization in Table 7.2.

Table 7.2: Terminology of computational data harmonisation.

Terminology	**Definitions**
Cohort	A group of data acquired by the same acquisition protocol and devices
Subjects	Patients (objects) involved in the study
Category	The classes that were involved in the study, e.g., cancer vs. normal
Cases	Samples (a subject can produce multiple samples with different acquisition protocols) involved in the study
Cohort bias	The non-biological related variances caused by acquisition protocols (also named as "batch effect" in gene expression studies)
Source cohorts	The cohort that needs to be harmonised from
Reference cohort	The cohort that needs to be harmonised to

What?

Computational data harmonization in digital healthcare aims to mitigate cohort bias, primarily originating from non-biological variations introduced by distinct data acquisition schemes. It employs computational strategies, such as machine learning and image/signal processing, to integrate multicenter datasets and reduce their non-biological heterogeneity. Unlike data cleansing, data normalization, and standardization, data harmonization possesses a broader scope and encompasses strategies for mitigating cohort biases caused by diverse acquisition protocols and devices. It encompasses actions like removing outliers (data cleansing), aligning the location-and-scale parameters of cohorts (data normalization), and converting multiple datasets into a common data format (data standardization/transformation, akin to manual harmonization). Importantly, data harmonization should not be confused with style transformation, which involves tasks like generating T1 images from T2 in MRI or generating CT images from X-ray images. Data harmonization exclusively focuses on intra-modality datasets.

Why?

Computational methods entail the automated analysis of digital healthcare data through the utilization of machine learning or mathematical modelling algorithms. Typically, these methods involve the extraction and fusion of data-derived features from raw data. For example, the grey level co-occurrence matrix (GLCM), a commonly used textual feature in radiomics, can

serve as an independent prognostic factor, representing metabolic intra-tumoral heterogeneity in surgically treated rectal cancer patients using F-FDG PET/CT images [201]. However, datasets obtained from different sites exhibit substantial variations (7.1), which can impede the efficacy of extracted features and result in unstable performance in both computational and manual diagnosis.

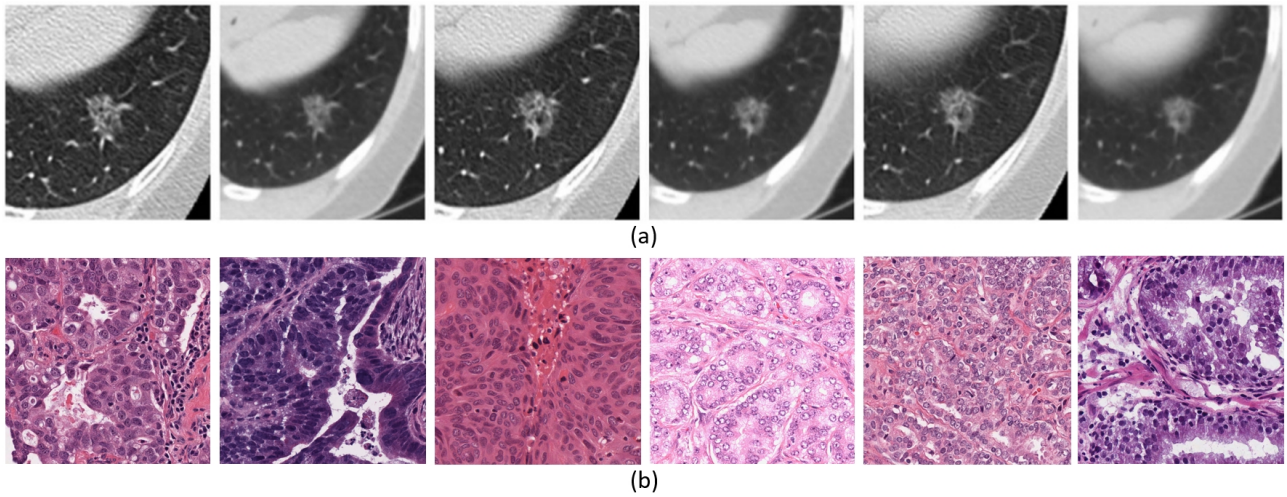


Figure 7.1: Visual Representation of Differences in (a) Radiomics and (b) Pathology Images. (a) Illustrates a lung tumour imaged using the same CT scanner but under six different acquisition protocols (Source: [1]). (b) Displays H&E stained tissue images collected from various sites (Source: [2]).

In particular, Zhao et al. [1] identified significant segmentation-based inconsistencies in lung tumours when multiple radiologists conducted repeated manual labelling. These inconsistencies could lead to a substantial reduction (from 0.76 to 0.28) in concordance correlation coefficients for certain radiomics features. Consequently, computational data harmonization is proposed to eliminate or mitigate these non-biological variances in multicenter datasets, with the aims of (1) enhancing the robustness and reproducibility of computational modules; (2) facilitating the fusion of knowledge acquired from prior tasks with knowledge obtained from new tasks; and (3) promoting comprehensive performance of computational modules.

The non-biological data variances primarily stem from hardware factors (e.g., scanners and platforms), acquisition protocols (e.g., signal/imaging acquisition parameters), and laboratory preparations (e.g., staining and slicing). These variances can undermine the reproducibility of quantitative biomarkers and constrain time-series studies based on multi-source datasets. Thus, there is an urgent demand for data harmonization strategies to generate reproducible

features [201, 202].

Heterogeneity of acquisition devices (inter-device variability)

The heterogeneity of acquisition devices, often referred to as inter-device variability, introduces variance in multicentre data and is particularly prominent in signals, CT, MRI, and pathological images. This heterogeneity arises from various factors, including differences in detector systems from different vendors, variations in the sensitivity of coils, positional and physiological differences during data acquisition, and magnetic field variations in MRI, among others (references [203–206]). Even when using a fixed acquisition protocol across different brands of scanners, studies have demonstrated that some radiomics features remain non-reproducible.

Table 7.3: Summary of the reproducibility/repeatability studies.

Reference	Intra-repro	Inter-repro	Repeatability	Condition	Variables	Object	Modality
Jha et al. [207], 2021	7.7% (332/1080)	1.4% (154/1080)	82.2% (888/1080)	ICC>0.90	Slice Sickness	Phantoms	CT
Emanimanijad et al. [208], 2021	8.0% (18/226)	/	/	CCC>0.90	Reconstruction	Patients	CT
Kim et al. [209], 2021	7.5% (112/1020)	/	/	CCC>0.90	Radiation Dose	Patients	CT
Yamashita et al. [210], 2020	/	5.6% (15/266)	/	CCC>0.90	Different Scanners	Patients	CECT
Fiset et al. [211], 2019	/	22.6% (398/1761)	/	ICC>0.90	Different Scanners	Patients	MRI
Saeedi et al. [212], 2019	20.5% (8/39)	/	/	CoV<5%	Tube Voltage	Phantoms	CT
	30% (13/39)	/	/	CoV<5%	Tube Current	Phantoms	CT
Meyer et al. [213], 2019	20.8% (22/106)	/	/	R^2>0.95	Radiation Dose	Patients	CT
	52.8% (56/106)	/	/	R^2>0.95	Reconstruction	Patients	CT
	39.6% (42/106)	/	/	R^2>0.95	Reconstruction	Patients	CT
	12.3% (13/106)	/	/	R^2>0.95	Slice Sickness	Patients	CT
Perrin et al. [214], 2018	24.8% (63/254)	/	/	CCC>0.90	Injection Rates	Patients	CECT
	13.4% (34/254)	/	/	CCC>0.90	Resolution	Patients	CECT
Midya et al. [215], 2018	11.7% (29/248)	/	/	CCC>0.90	Tube Current	Phantoms	CT
	19.8% (49/248)	/	/	CCC>0.90	Noise	Phantoms	CT
	63.3% (157/248)	/	/	CCC>0.90	Reconstruction	Patients	CT
Altazi et al. [216], 2017	21.5% (17/79)	/	/	Mean difference <25%	Reconstruction	Patients	PET
Zhao et al. [1], 2016	11.2% (10/89)	/	/	CCC>0.90	Reconstruction	Patients	CT
	/	/	69.7% (62/89)	CCC>0.90	/	Patients	CT
Hu et al. [217], 2016	/	/	64.0% (496/775)	ICC>0.80	/	Patients	CT
Choe et al. [218], 2019	15.2% (107/702)	/	/	CCC>0.85	Reconstruction	Patients	CT

Heterogeneity of acquisition protocols (intra-device variability)

The primary sources of cross-cohort variability in multicentre data are different acquisition protocols, which encompass scanning parameters (e.g., voltage, tube current, field of view, slice thickness, microns per pixel) and reconstruction approaches (e.g., various reconstruction kernels) [219]. To assess the reproducibility and repeatability of radiomics features, several studies have conducted test-reset experiments. In Table 7.3, good reproducibility/repeatability is defined as a high correlation coefficient (e.g., ICC, CCC, R^2) or low difference (e.g., mean difference, CoV) between two features. For example, a radiomics feature is considered reproducible/repeatable when the CCC between features extracted from two repeated scans is greater than 0.90. However, scanning parameters significantly impact radiomics features, leading to

challenges in statistical analysis. For instance, only 15.2% of radiomics features are reproducible when soft and sharp reconstruction kernels are used [218]. This limited reproducibility poses a significant obstacle to large-scale digital healthcare studies and the application of computational models. While implementing strict standard protocols can reduce non-biomedical variances, physicians often require non-standard acquisition protocols to address personalized image quality considerations. For instance, adjustments to thickness and pixel size are regularly made on a case-by-case basis to enhance data quality [220]. Consequently, the heterogeneity of acquisition protocols is inevitable, necessitating a general solution.

Heterogeneity of laboratory preparations (Preparation variability)

Laboratory variances, including sample preparation, assay, slicing, and staining, introduce heterogeneity in gene expression, radiomics, and pathological data. In single-cell RNA sequencing (scRNA-seq) and microarray data, various analysis platforms exhibit different biases, making it challenging to integrate and compare results from multi-centre or batch data [221, 222]. Radiomics data may also be affected by variances such as injection rate and radiation dose, which can impact data quality. In the case of pathology data, variances primarily arise from manual operations [223, 224], including biopsy sectioning, sample fixation, dehydration, and stain concentration. These factors collectively contribute to variations in pixel values and stain consistencies in pathological data.

What for?

The challenges of integrating and utilizing multicentre datasets highlight the importance of data harmonization in conducting large-scale and longitudinal studies [225]. Without harmonization, information fusion cannot yield reproducible results in such studies [226], and caution must be exercised when drawing conclusions, as some features may vary significantly due to minor non-biomedical changes [227]. Data harmonization is particularly crucial for patients monitored longitudinally and imaged on different scanners. For example, longitudinal PET data collected from multiple scanners may not provide useful information if not harmonized, as the relationship between standard uptake value (SUV) and outcomes may become obscured [202].

Another challenge is the transferability of computational models, as their performance tends to be unstable when applied to multicentre datasets [228]. To address this issue, transfer learning has been proposed to enhance the robustness of computational models by incorporating prior knowledge about potential data variability. Transfer learning requires additional training samples to reduce uncertainty regarding the data variability that models may encounter during inference, which can be impractical for prospective studies in the digital healthcare field. In contrast, computational data harmonization strategies can process data without requiring extra training or fine-tuning, providing a practical solution for multicentre studies. Moreover, there is growing evidence that combining data harmonization with machine learning algorithms enables robust and accurate predictions on multicentre datasets [229].

How?

The deployment of computational methods in data analysis involves several stages: preparation, pre-processing, modeling, and analysis. Preparation includes acquiring datasets through techniques like staining and scanning. The next stage, pre-processing, involves refining the data for further analysis. Data harmonization is a crucial step in this process, and it can be approached in two ways: sample-wise and feature-wise (shown in Fig. 7.2).

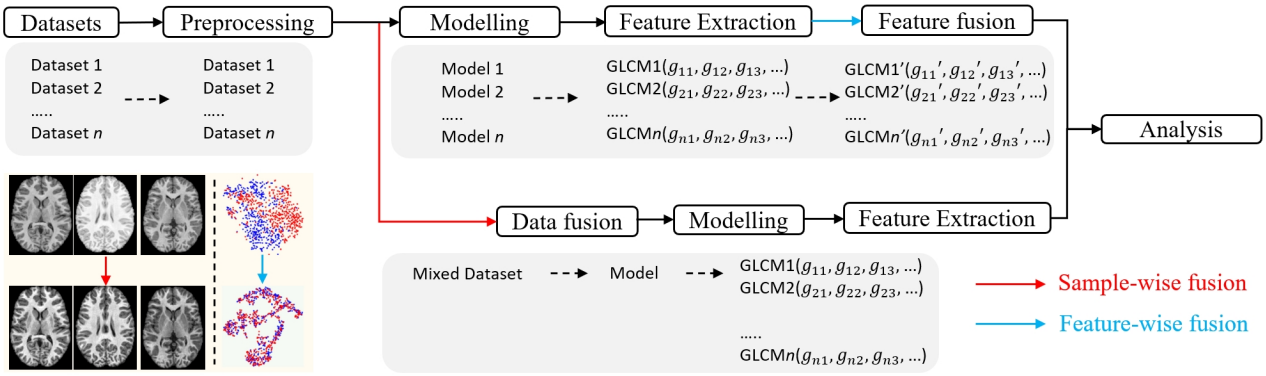


Figure 7.2: Scheme of computational data harmonisation.

Sample-wise harmonization, typically done before modelling, aims to minimize variance across training samples from different sources, effectively treating them as a single dataset. This process often involves image processing, synthesis, and learning invariant features, which helps in developing a single, robust model for clinical tasks. Feature-wise harmonization, on the other

hand, focuses on reducing bias in extracted features, such as the Gray Level Co-occurrence Matrix (GLCM) or the convex hull area in regions of interest. This approach is usually applied to extracted feature matrices, where variances are eliminated by integrating the features from different cohorts, as illustrated in the referenced figure with red and blue dots representing different cohorts.

Both methods are effective in reducing variances and enhancing analysis performance. However, feature-wise harmonization is more complex as it requires multiple models to extract relevant features. This complexity increases when dealing with small sample sizes in each cohort, making it challenging to develop appropriate models.

7.1.3 Methods

Literature search and review

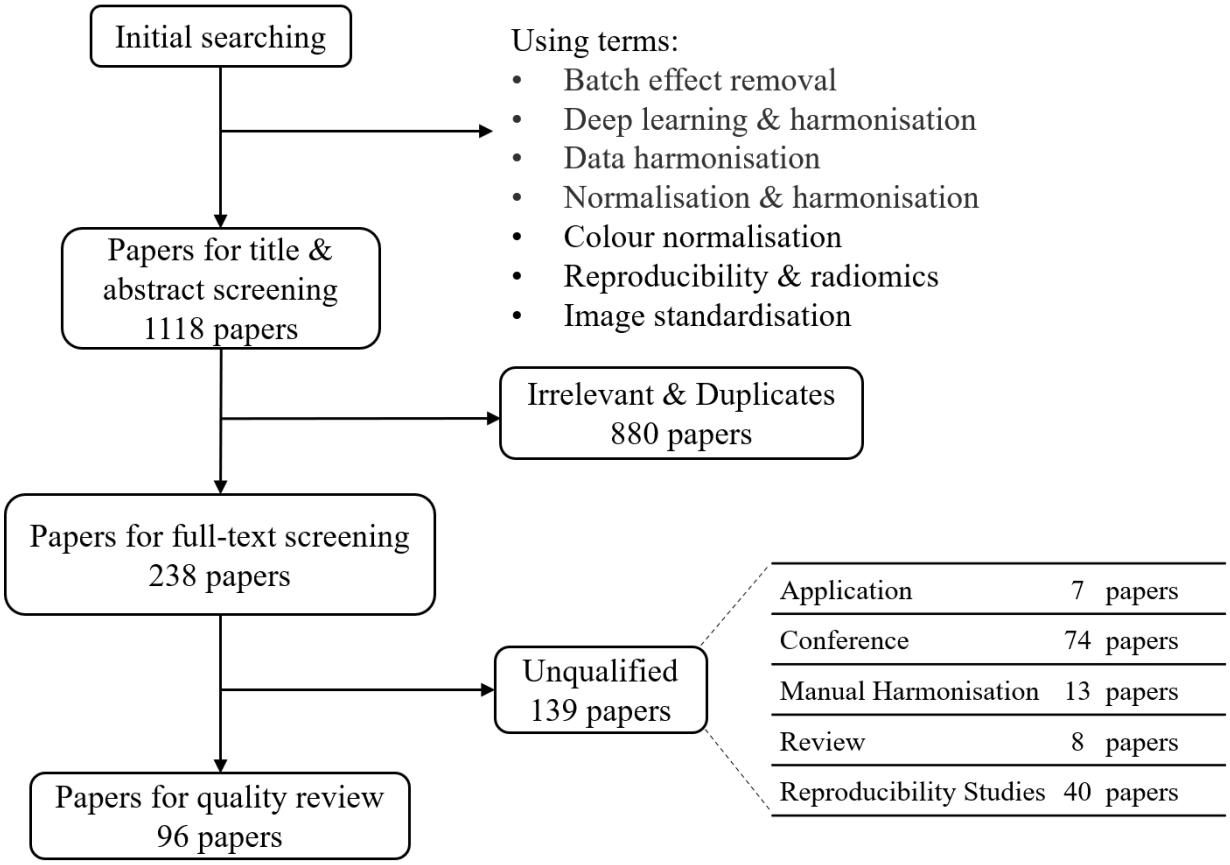


Figure 7.3: Procedure of literature selection.

The literature search for our systematic review was a meticulously structured process, carried out by two researchers with backgrounds in computer science and biomedical research. To ensure objectivity and thoroughness, disagreements were resolved by consulting a third reviewer specializing in biomedical data analysis. This comprehensive search was conducted using the Scopus Preview (Elsevier) database, encompassing publications up to July 10, 2021. Our investigation focused on harmonization strategies in information fusion, guided by a set of keywords: 'batch effect removal', 'deep learning', 'harmonisation', 'data harmonisation', 'normalisation' and 'harmonisation', 'colour normalisation', 'reproducibility' and 'radiomics', 'image standardisation'. These terms were searched both individually and in combination to maximize the breadth of literature covered. It's important to note that 'normalisation' and 'standardisation' are both considered methods of harmonisation.

The initial stage of screening involved reviewing titles and abstracts to exclude irrelevant articles. This was followed by a more detailed evaluation against our specified criteria, to determine the eligibility of articles for full-text review. A flowchart in Fig. 7.3 visualizes this literature selection process. After discarding irrelevant and duplicate articles based on title and abstract screening, 238 articles were identified for full-text evaluation. Applying our eligibility criteria, 139 of these were deemed unqualified, leaving 96 papers to be included in the systematic review.

7.1.4 Data harmonisation strategies for information fusion

In our systematic review, we categorized data harmonization approaches into four distinct groups: distribution-based methods, image-processing techniques, synthesis strategies, and invariant feature learning. To clarify the underlying concepts and interrelationships among these computational approaches, we created a taxonomy. This taxonomy is visually represented in Fig 4, providing an intuitive guide to understanding the different techniques. This illustration is followed by an in-depth description of each harmonization method, further elucidating the nuances and applications of these techniques in the field of data harmonization.

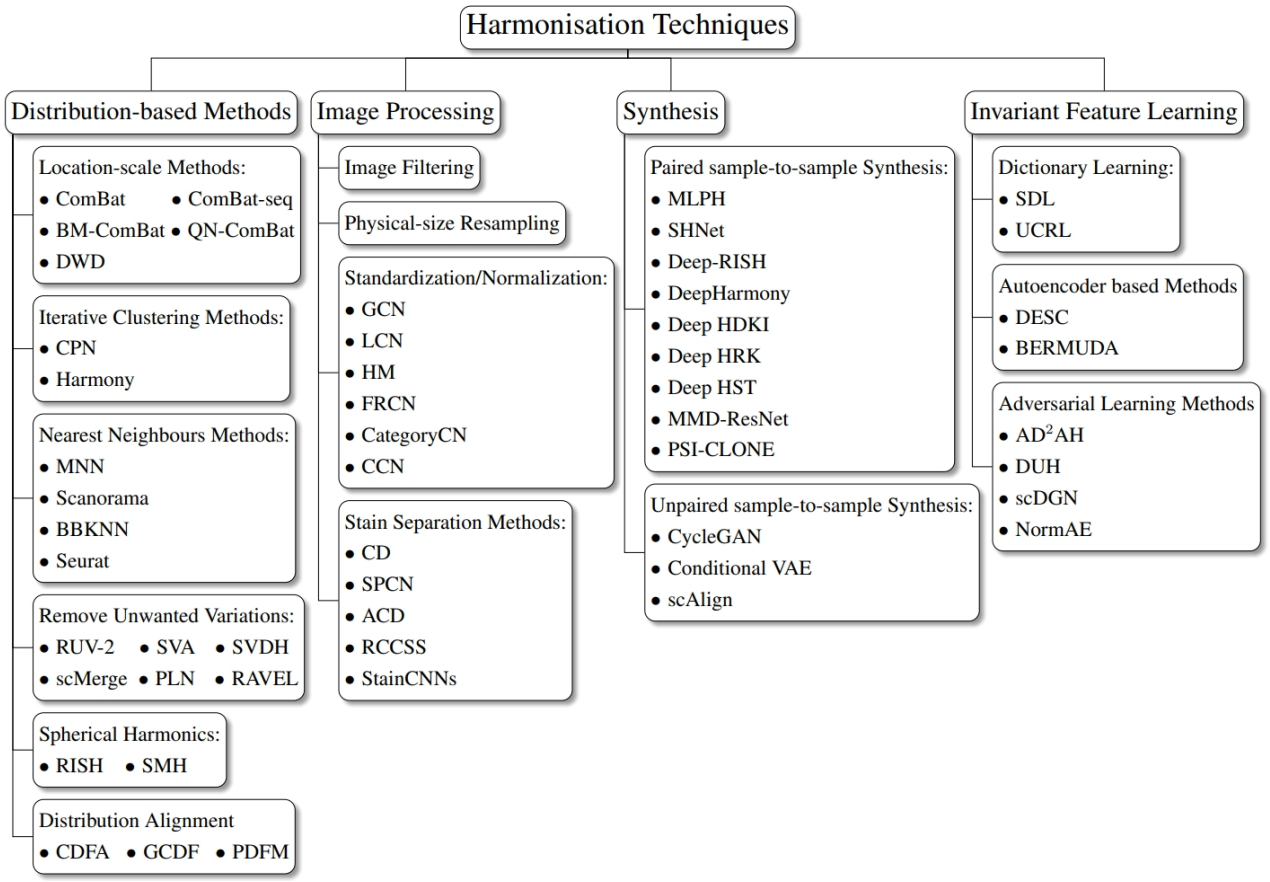


Figure 7.4: Taxonomy of computational data harmonisation strategies.

Distribution based methods

Distribution-based methods in data harmonization are designed to estimate or calculate biases between cohorts from latent spaces and then align or map source data to target data using bias correction vectors or alignment functions.

Location-Scale Methods (LS). Location-scale methods involve adjusting the mean and variance of each dataset to a common scale. A notable example is the ComBat algorithm [230], which uses empirical Bayes methods to robustly estimate these parameters and then standardizes the data to reduce biases. Variations of ComBat, such as ComBat-seq [231] and QN-ComBat [232], offer flexibility by accommodating different data distributions and incorporating quantile normalization for improved performance in specific contexts like gene expression data.

Iterative clustering methods (IC). These methods employ iterative clustering to refine the harmonization process through repeated bias corrections. Techniques like Cross-platform nor-

malization (XPN [233]) and Harmony [234] utilize clustering to identify and harmonize groups of similar data points or genes, thus aligning the overall data distribution. These approaches are especially effective in contexts where data needs to be segmented into distinct clusters before harmonization.

Nearest neighbours methods (NNM). NNM focuses on identifying and aligning the nearest neighbour pairs across different datasets. Methods like Mutual Nearest Neighbours [235] and Scanorama [236] calculate bias correction vectors based on these pairs and apply adjustments to achieve harmonization. These techniques [223, 237, 238] are crucial for datasets where relationships between data points can be represented as geometric distances.

Remove unwanted variations (RUV). RUV methods [239–241] operate on the premise that cohort bias is distinct from biological variance, treating the latter as "unwanted variations." For instance, biases observed in negative control genes (genes known to be unaffected by the biological changes of interest) are considered indicative of cohort bias [242, 243]. Under this assumption, raw data is harmonized by subtracting these unwanted variations.

Spherical harmonics (SH). Spherical harmonics approaches are specialized techniques used to harmonize MRI data. Their primary objective is to align data from different cohorts into the same spherical harmonic domain by adjusting spherical variables. Two notable methods in this category are Rotation Invariant Spherical Harmonics (RISH) [203, 244–246] and Spherical Moment Harmonics (SMM) [247, 248].

Distribution alignment (DA). DA methods, such as Cumulative Distribution Functions Alignment [249] and Probability Density Function Matching [250], transform the data distributions of the source cohort to match those of the reference cohort. These techniques are adept at handling complex data transformations and are essential when datasets require alignment to a highly specific reference distribution.

Image processing

Image processing techniques are utilized in data harmonization across multiple cohorts, incorporating a range of digital image processing algorithms. These include image filtering (also known as image convolution), registration, resampling, and normalization. **Image Filtering**

and Resampling. Image filtering techniques [251, 252], such as convolution with arrays like the 2D second-order Butterworth low-pass filter, help eliminate cohort biases by adjusting the data's frequency components. This is particularly effective for CT images where voxel size discrepancies can skew data interpretation. Additionally, physical size resampling, such as bi-linear resampling [251], standardizes pixel or voxel sizes across images, which is critical for maintaining the integrity of radiomic and pathological features.

Standardization/Normalization (SN). Standardization and Normalization: These methods are employed to reduce variability and inter-cohort differences by applying linear transformations across image datasets. Techniques like global colour normalization (GCN) [253, 254] and histogram matching [255, 256] adjust the colour statistics and contrast of images, respectively, aligning them with a reference standard. Local colour normalization [257] further refines this by focusing adjustments on specific image regions, preserving essential information while aligning background luminance.

Stain separation methods (SS). Stain separation approaches [258–261] decompose input images into distinct channels, such as the haematoxylin, eosin, and background channels in H&E-stained images, to analyze and align stain feature matrices from source to target cohort data. The foundational principle of stain separation is based on Lambert Beer's law [262], which in RGB space suggests that stain concentrations are nonlinearly dependent on light intensity:

$$I_C = I_0 e^{-OD_c} \tag{7.1}$$

where I_0 is the incident light value, and OD_c represents the optical density (OD) of images. The objective of most stain separation methods is to factorize OD values into two matrices:

$$OD_c = \log\left(\frac{I_0}{I_C}\right) = S \times D \tag{7.2}$$

Here, S denotes the stain depth matrix, and D represents the stain colour appearance (SCA) matrix.

Synthesis

Synthesis aims to accurately recreate a sample from a missing modality or domain, facilitating the harmonization of multi-cohort datasets. It views harmonization as a style transfer task, treating each cohort as a unique 'style' and converting all samples to a uniform 'style'. Synthesis methods are categorized based on training sample characteristics into paired synthesis and unpaired synthesis.

Paired Sample-to-Sample Synthesis (P-s2s). P-s2s methods utilize paired samples from the same subject acquired through different protocols to train the system. They focus on learning data transfer between source and reference cohorts, necessitating repeated subject acquisition under varying protocols. This method is mostly applicable to radiomic data due to the impracticality of repeated acquisition for gene expression and pathology data. Techniques such as the Spherical Harmonic Network [263], Deep Rotation Invariant Spherical Harmonics [264] and DeepHarmony [218, 265, 266] utilize deep learning frameworks to adjust the data to a consistent style across different cohorts. For example, DeepHarmony uses a U-Net architecture to ensure consistent contrast across various data samples.

Unpaired sample-to-sample synthesis (Up-s2s). Unpaired Sample-to-Sample Synthesis (Up-s2s) methods create harmonized data using cycle-consistent generative adversarial networks [267, 268] or conditional variational autoencoder-decoders [269–271], requiring ample samples and cohort labels from different cohorts for effective network training.

Invariant feature learning

Invariant feature learning techniques aim to identify cohort-invariant features from diverse cohorts, subsequently utilizing these features for primary tasks such as segmentation, classification, or regression. The essence of representation learning for harmonization lies in constructing a sparse dictionary or mapping from multi-cohort data, ensuring learned representations exclude inter/intra-cohort variability.

Dictionary Learning (DictL). Sparse dictionary learning [272, 273] reduces data complexity

by identifying key invariant features that can be used to reconstruct data accurately. This method is pivotal in scenarios where the essential data features need to be distilled from large datasets with substantial variability.

Autoencoder-Based Methods. Techniques like DESC [274] and BERMUDA [275] use autoencoders to learn a representation of data that minimizes within-cohort variability while maximizing the distinction between different cohorts. These learned features are then used for tasks such as classification or regression, ensuring that the models perform well across all data sources.

Adversarial Learning Methods (AdvL). Adversarial learning methods develop systems that concentrate on scanner/protocol invariant features while maintaining performance on the main task, thus mitigating cohort bias in predictions. These methods [276–279] typically consist of an adversarial module for cohort identification, a backbone for feature extraction, and a main task module for classification, regression, and/or segmentation.

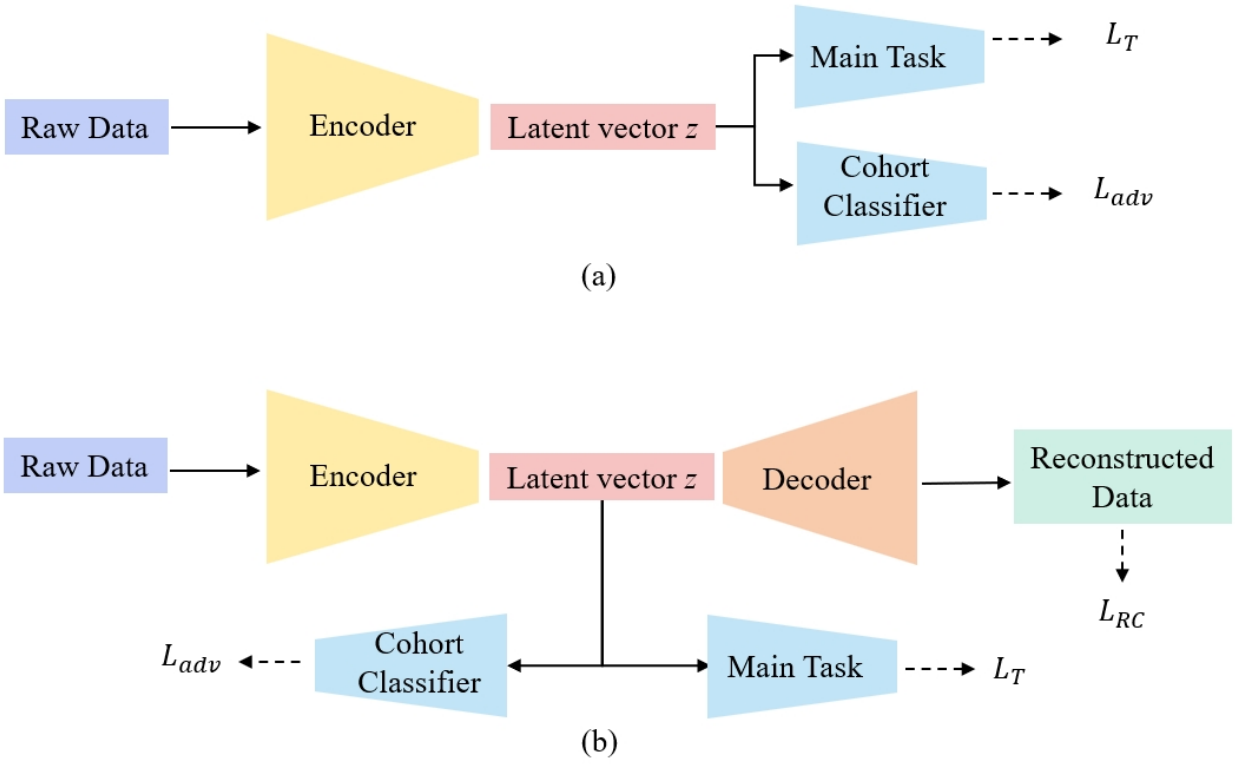


Figure 7.5: Description of adversarial learning methods.

Adversarial learning methods applied in data harmonization predominantly adopt two struc-

tural frameworks, as depicted in Fig. 7.5 (a) and (b). Techniques such as AD2AH (Attention-guided Deep Domain Adaptation Harmonics) [276], DUH (Deep Unlearning Harmonics) [277], and scDGN (Single-Cell Domain Generalisation Network) [278] utilize an adversarial module (domain classifier). This module aids the encoder in learning cohort-invariant features by maximizing the adversarial loss L_{adv} while minimizing the main task loss L_T, as shown in Fig. 7.5(a). For precise feature representation z, methods like NormAE [279] incorporate a decoder to reconstruct the input raw data, aiming to minimize the reconstruction loss L_{RC}, illustrated in Fig. 7.5(b). Through these optimization functions, the main task is enabled to maintain stable performance across multi-cohort data.

7.1.5 Evaluation approaches of the data harmonisation strategies

This section delves into evaluation methodologies for harmonization performance, categorizing them as distribution-based, correlation-based, value-based, and task-based metrics. Distribution-based metrics assess harmonization through cluster analysis or location-scale parameters across cohorts. Correlation and value-based metrics examine the variability of data-derived features from different cohorts to evaluate reproducibility. Cohort classification is also employed as an evaluation method, showcasing the effects of harmonization through classification results. Visualization is a commonly utilized approach for direct comparison of datasets pre and post-harmonization.

Distribution-Based Evaluation. Distribution-based metrics evaluate harmonization by analyzing clustering or local-scale parameters. Clustering-related metrics include the Adjusted Rand Index (ARI) [274], k-Nearest Neighbour Batch-Effect Test (kBET) [280], and Local Inverse Simpson's Index (LISI) [234, 281]. Location-scale metrics encompass Structural Similarity Families [46, 282], Normalised Median Intensity [223, 261, 283–287], Coefficient of Variation [52, 203, 212, 243, 288–292], and KL Divergence [274, 293].

Correlation-Based Evaluation. Correlation-based evaluation measures reproducibility through statistical analysis, calculating correlations between features before and after harmonization. This includes the Pearson correlation coefficient [222,223,261,294–297], concordance correlation

coefficient [218, 251, 252, 266, 298, 299], intra-class correlation coefficient [249, 250, 252, 288, 300], and others.

Value-based evaluation. Value-based evaluation primarily focuses on assessing intensity differences between data or data-derived features before and after harmonization. A low value of intensity differences indicates effective harmonization, often relying on a "ground truth" as a benchmark for results. These evaluation metrics include Mean Absolute Error (MAE) [244, 265, 277, 296, 301–307], Root-Mean Square Error (RMSE) [239, 248, 249, 263, 264, 266, 290, 294, 303, 307–310, 310], Peak Signal to Noise Ratio (PSNR) [309, 311].

Main Task-Based Performance Evaluation. The effectiveness of harmonization methods is often demonstrated by comparing the performance of primary tasks (e.g., regression, segmentation, classification) before and after harmonization [242,249,303,312]. While this result-oriented approach may be influenced by the random initialization of models, it provides substantial evidence of harmonization effectiveness. Regression tasks are evaluated using metrics like the Dice coefficient or Intersection over Union (IoU) [256, 259, 261, 267, 277, 305, 313]. Classification tasks employ various metrics, including Area Under the Receiver Operating Characteristics Curve (AUC) [243, 259, 263, 276, 279, 284, 286, 314–318], accuracy [222, 237, 247, 255, 259, 269–271, 274, 276, 278, 289, 305, 306, 314, 318–321], True Positive Rate [231, 281], sensitivity [276, 319], specificity [319], and Matthews Correlation Coefficient (MCC) [289].

Cohort Classification. Some studies assess harmonization strategies using cohort classification [251, 294, 306]. This metric is based on the premise that effective harmonization makes cohort identification more challenging. For example, Wachinger et al. [306] evaluated cohort identification accuracy across various datasets, finding that ComBat yielded the least accurate results, indicating superior harmonization.

Visualization. Visualization techniques depict data distributions in low-dimensional feature spaces or sample intensities. Approaches like Principal Component Analysis (PCA) [232, 239, 241, 279, 295, 303, 312, 320], Uniform Manifold Approximation and Projection (UMAP) [222, 237, 271, 275, 281, 314, 322], and t-Distributed Stochastic Neighbour Embedding (t-SNE) [235, 238, 241, 270, 271, 276, 278, 315, 323, 324] are employed to visualize data distribution in latent spaces.

Effective harmonization is indicated by mixed data distributions rather than distinct clusters. Researchers also plot intensities or location-scale parameters (mean and variance) of samples pre and post-harmonization to evaluate performance [230, 241, 243, 254, 256, 257, 259, 260, 290, 309].

7.1.6 Meta-analysis

This review categorizes methodologies and metrics based on diverse theories and conducts meta-analyses in three key areas: gene expression, radiomics, and pathology. The differentiation in modalities is essential due to distinct data properties: gene analysis primarily involves expression matrices, radiomics encompasses high-dimensional volume arrays (grayscale images per slice), and pathology relies on large-sized colour images.

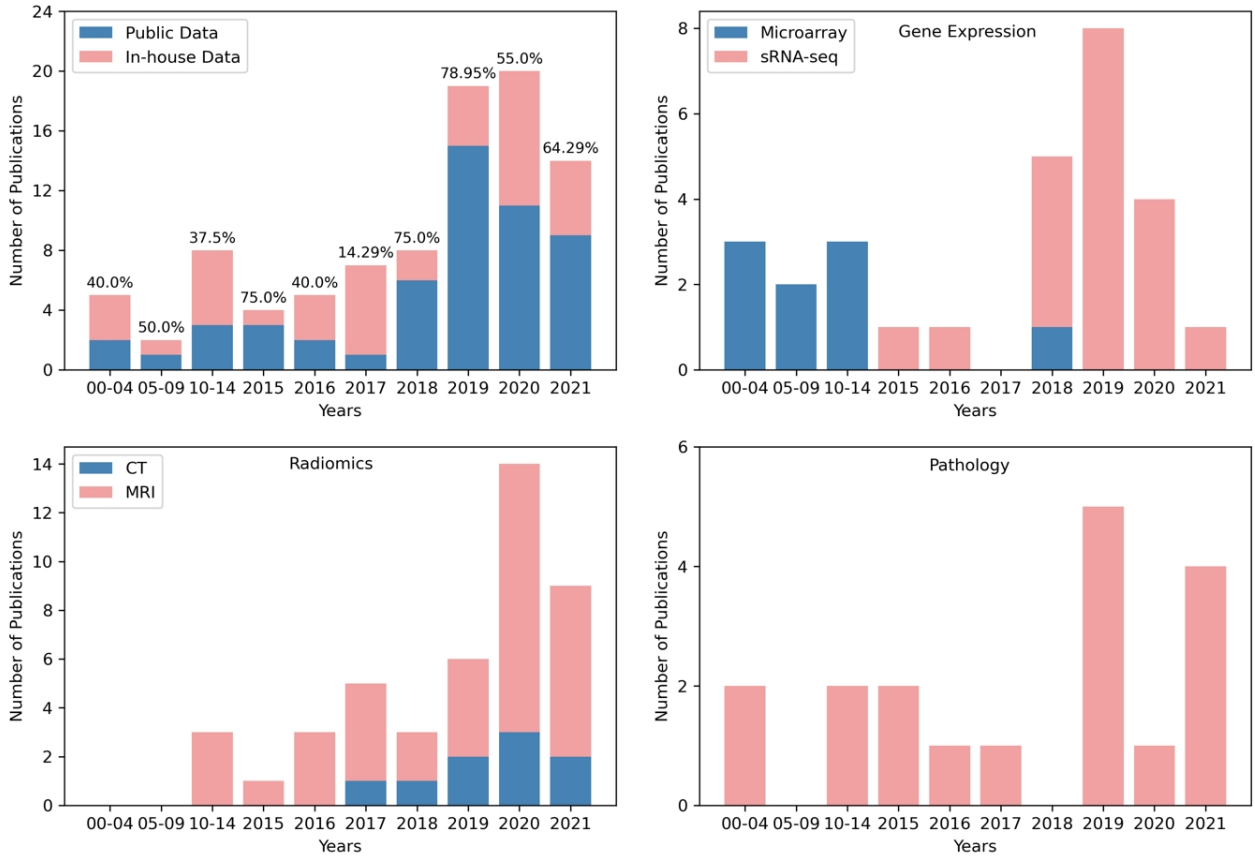


Figure 7.6: Trend in Publications Over Years Based on Data Properties and Modalities. 'Public data' refers to open-source data accessible online, while 'in-house data' denotes datasets not available on the internet. The percentage shown in the top left subfigure represents the proportion of studies conducted using public datasets.

Data Properties and Study Trends. Fig. 7.6(top left) illustrates the evolution of har-

monization studies from 2000 to 2021, emphasizing the proportion of studies using public datasets, which are accessible via open-source platforms, unlike in-house data. A notable surge in harmonization research since 2019 highlights the growing need for large-scale studies and sophisticated harmonization strategies. We also explore harmonization research trends across various sub-modalities, including Microarray and sRNA-seq in gene expression, CT and MRI in radiomics, and digital pathology. In gene expression, microarray harmonization peaked before 2015, whereas sRNA-seq has gained prominence recently. Radiomics studies, particularly in MRI, are increasingly focusing on enhancing feature reproducibility. Pathology data harmonization has been a longstanding focus, receiving heightened attention in recent years.

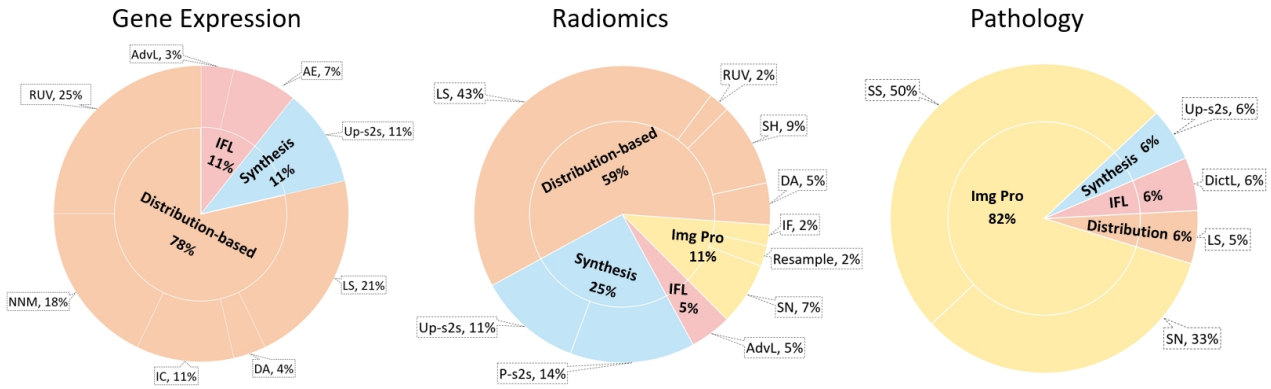


Figure 7.7: Harmonization Strategies Across Different Modalities. 'IFL' denotes Invariant Feature Learning approaches, and 'Img Pro' stands for Image Processing approaches. Each pie chart is annotated with the percentage of various sub-methods, identified by their abbreviations.

Strategies and Modalities. The intersection of diverse biomedical data modalities with harmonization strategies is presented in Fig. 7.7. Distribution-based methods are prevalent in gene expression and radiomics (79% and 59%, respectively), but less so in digital pathology (5%). Empirical Bayes methods are favored for their generalizability and robustness, while RUV and SH are more modality-specific. Image processing approaches are predominantly utilized in digital pathology, focusing on standardization/normalization and stain separation. Recently, invariant feature learning and synthesis, driven by deep learning algorithms, have been applied across all three modalities.

Evaluation Metric. Evaluation metrics, essential in harmonization method development,

assess performance by analyzing distributions, correlations, and values between source and target cohorts. Visualization is the most widely used evaluation method, followed by main task performance analysis (Fig. 7.8). Some studies use cohort classification for evaluation, but this approach has limitations as the inability to distinguish cohorts doesn't necessarily imply effective harmonization. Despite various evaluation options, challenges persist in clinical implementation, including the need for paired datasets (often impractical in clinical settings), machine learning-based modules for performance comparison, clinician involvement for subjective assessment, and predefined ROIs. Moreover, harmonization's role is pivotal for model federation, raising questions about the extent to which local datasets should be harmonized without compromising locally specific data nuances beneficial for local model generalization.

Along with the utilization of various metrics, Fig. 7.8 details the evaluation metrics specific to different modalities. Visualization is notably preferred in gene expression and digital pathology, encompassing data distribution visualization (e.g., using UMAP, t-SNE) and comparison of samples pre and post-harmonization. Many pathology studies leverage distribution-based metrics, such as structural similarity and normalized median intensity. Task-based evaluation is also deemed reliable, comprising 19%, 28%, and 13% of all evaluation methods across modalities.

Table 7.4: Data scale (image size) in previous studies. The small, middle, large image sizes are defined as $\epsilon \in [0, 256)$, $\epsilon \in [256, 512)$, $\epsilon > 512$, respectively, N/A indicates there is no report of image size.

	Small	Middle	Large	N/A
Radiomics	9.1% (6)	18.2% (12)	3.0% (2)	69.7% (46)
Pathology	5.0% (1)	15.0% (3)	60.0% (12)	20% (4)

Data Scale - Images. The scale of samples in radiomics and pathology is intricately linked to image resolution and quality. We analyzed the variable $\epsilon = \sqrt{w \times h}$ for studies involving images of width w and height h. In radiomics, most studies utilized 256-pixel images, while some used 512 and 128 pixels, often constrained by GPU or RAM limitations, especially for 3D or multi-slice datasets. Notably, 69.7% of radiomics studies did not report image size, compared to 20% in pathology. Pathological images are predominantly large-scale (ϵ>500), as image processing algorithms yield better results with broader fields of view.

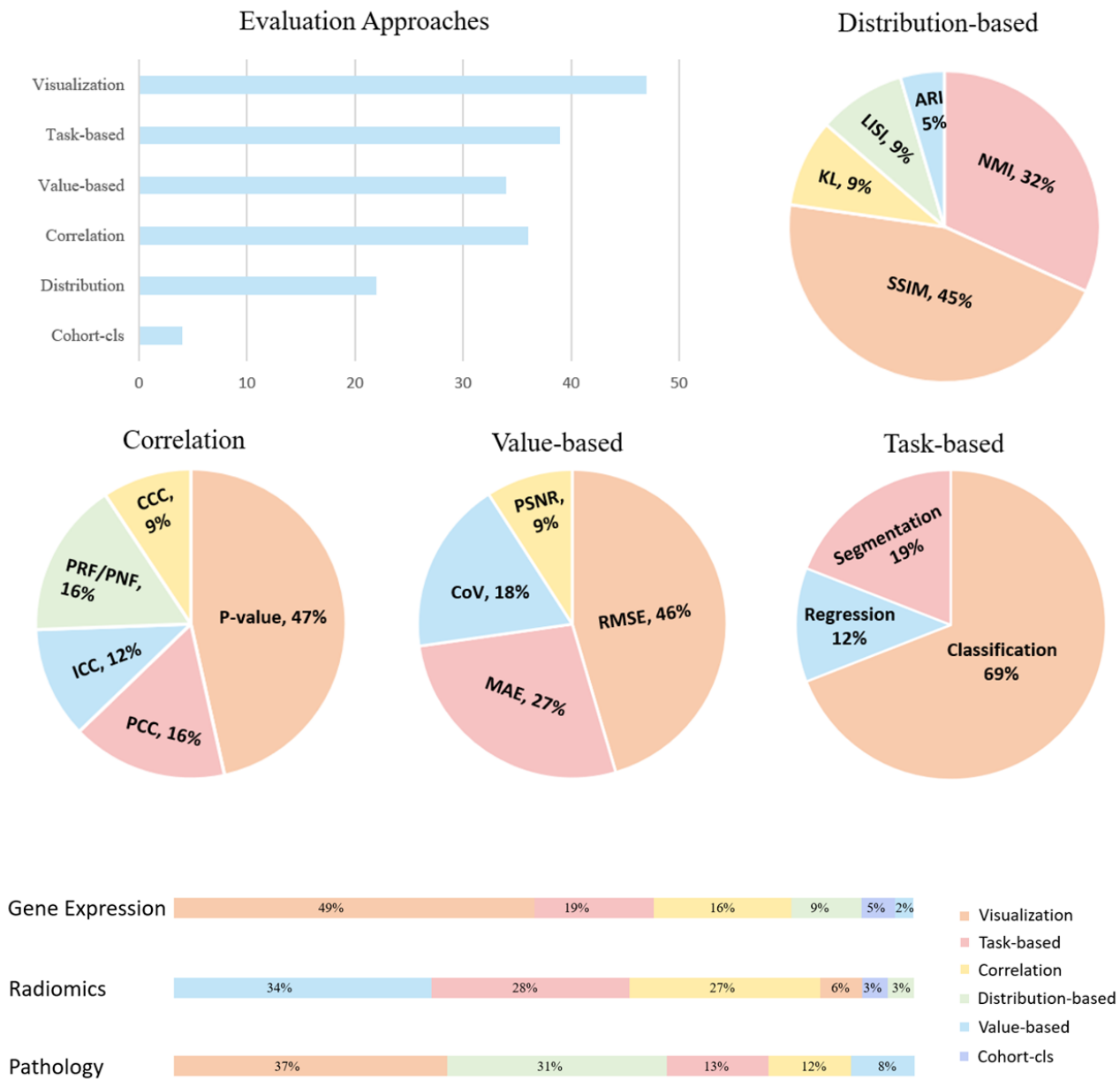


Figure 7.8: Evaluation metrics regarding different modalities.

7.1.7 Research directions

This section addresses research directions for various data harmonization approaches, highlighting their limitations and common restrictions in prior studies.

Directions for Distribution-Based Methods. Distribution-based methods map source data to targets by estimating cohort variances, but face several challenges:

1. They often rely on refined feature vectors requiring prior knowledge of the region of

interest, conflicting with their initial purpose for robust multicentre dataset processing.

2. Although effective in phantom studies (e.g., ComBat), they struggle with images or high-dimensional signals due to computational complexity.

3. Harmonization must be reapplied to entire datasets when new data are added.

4. Pairwise approaches (e.g., XPN, DWD, CCA, MNN, Seurat) complicate training for multicentre datasets [274].

To address these, future research should focus on harmonizing raw datasets, developing efficient and robust strategies, and avoiding pairwise algorithm development.

Directions for Image Processing-Based Methods. Image processing-based methods offer simpler harmonization but have their limitations, such as field-specific applications (e.g., stain separation) and reliance on empirical settings (e.g., filtering kernels). Future work should develop general approaches that minimize reliance on empirical settings.

Directions for Synthesis. Deep learning-based synthesis has advanced rapidly but faces issues with reproducibility and generalizability:

1. Many methods lack evaluations on new datasets.

2. GAN-based models risk instability and unrealistic changes.

3. Training GANs requires extensive data, often impractical in clinical studies.

Researchers should focus on reporting performance on new datasets, enhancing stability, and building strategies requiring less training data.

Directions for Invariant Feature Learning. Invariant feature learning focuses on extracting cohort-invariant features, while it cannot produce harmonized data. Future studies should explore generating harmonized data using extracted features.

Explainable AI for Further Harmonisation Studies. Explainable AI (XAI) methods [325] remain unexplored in harmonization literature. XAI can identify data representation inconsistencies, aiding in detecting biases and out-of-distribution examples. Improved harmonization

can benefit XAI by standardizing data and reducing biases [326, 327]. The intersection of harmonization and XAI presents exciting research opportunities.

Limitations in Methodology Design. Many harmonization studies lack a stepwise design methodology, impeding reproducibility. For instance, over half of radiomic studies do not report image scale. Diverse definitions of 'reproducible' and varied evaluation metrics also hinder method comparison for future research.

7.1.8 Guidance of data harmonisation strategies and metrics

Research indicates that improper data harmonization strategies can introduce significant bias, leading to less accurate predictions [328]. To aid in method selection, a flowchart detailing potential harmonization approaches is depicted in Fig. 7.9 (left). According to the flowchart, distribution-based methods are effective for refined features or gene matrices. For high-dimensional images, image processing methods are preferable in the absence of high-performance GPUs. Deep learning-based methods, including invariant feature learning and synthesis, are suitable for various modalities but require ample training samples. Invariant feature learning is particularly recommended when the main task aligns with the training process, as synthesis might introduce unrealistic artefacts into the data. Regarding evaluation, the

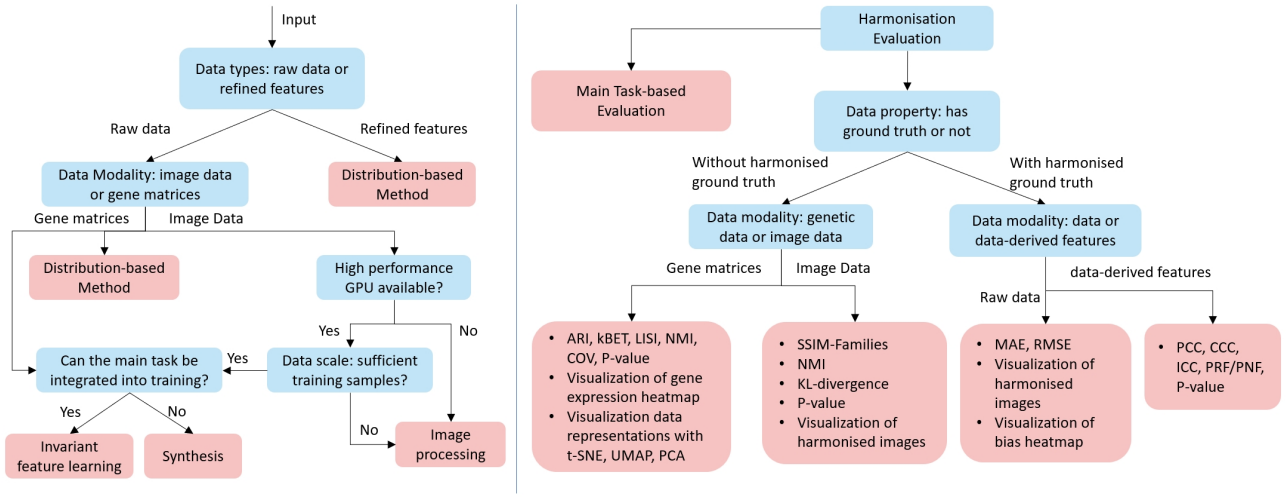


Figure 7.9: Flowchart of how to choose harmonisation approaches (left) and evaluation metrics (right).

choice of metrics crucially determines the reliability of results. Fig. 7.9 (right) summarizes and

recommends data harmonization metrics for different scenarios. Visualization offers an intuitive method for analyzing harmonization outcomes, achievable through techniques like t-SNE, UMAP, or PCA on both raw and harmonized data. Main task-based evaluation can effectively demonstrate the impact of harmonization strategies by comparing task performance on data pre- and post-harmonization. In the absence of harmonized ground truth, distribution-based metrics can be employed to gauge sample mixing, though cohort labels might be necessary. When harmonized ground truth is available, value-based or correlation-based metrics can precisely indicate the effectiveness of data harmonization.

7.1.9 Conclusion

Computational data harmonization has played a crucial role in digital healthcare research for years. However, integrating basic science research models and data fusion into multicentre, multimodal, and multi-scanner medical practice and clinical trials remains challenging without effective harmonization. Transfer learning, federated learning, multitask learning, and other knowledge-exchange methodologies require minimal distribution shifts to maintain coherence across models and centres. In cases of significant distribution variation, data harmonization becomes essential. However, the appropriate approaches and metrics for handling multimodal datasets are not well-established, and the absence of standardized methodologies hinders reproducibility.

To address these challenges, this section provides a comprehensive summary and categorization of existing data harmonization strategies and metrics, informed by diverse theoretical underpinnings. The introduction of the CHECDHA criteria aims to standardize the harmonization process, enhancing academic reproducibility and development. Furthermore, we summarize data harmonization approaches and evaluation metrics specific to three modalities (Fig. 7.7 and Fig. 7.8), offering guidance on selecting suitable strategies and metrics (Fig. 7.9) based on various conditions. Additionally, limitations and future directions for different harmonization methods were discussed.

Data harmonization is increasingly recognized as a pivotal process in large multicentre studies

within computational biomedical research. It is particularly adaptable to federated learning systems, significantly contributing to the advancement of computational modules in areas like radiomics, genetics, and pathology. Given the critical need for standardization in reporting harmonization research, we advocate for researchers to adhere to and build upon the checklist presented in this survey.

7.2 Deep Vectorised Hashing for Image Retrieval

Image retrieval in large-scale clinical databases is effectively addressed by leveraging CNN and hashing techniques, surpassing traditional retrieval methods. However, some current methods simplistically treat retrieval as classification, resulting in over-centralization by assigning identical hash codes to similar class samples. Other approaches attempt to control the Hamming distance between similar and dissimilar samples using probability functions, but their sharpness often leads to training instabilities like vanishing gradients. Additionally, most methods overlook the out-of-distribution (OOD) challenge, particularly under adversarial attacks. To overcome these limitations, we introduce a novel hybrid model combining deep hashing with OOD detection for medical image retrieval.

7.2.1 Introduction

The rising importance of medical imaging for enhancing diagnoses and patient outcomes has led to an accumulation of substantial imaging data, necessitating effective management and retrieval mechanisms. Image retrieval, pivotal for case analysis and clinical education, enables swift access to pertinent images, thereby aiding efficient diagnosis and prognostication. However, the sheer volume of data in large-scale repositories poses a challenge, which automated image retrieval can address. Among various techniques, hashing, notable for transforming high-dimensional data into binary codes (hash codes), facilitates the identification of similar samples via Hamming distance computation. Nonetheless, traditional hashing methods often yield approximate results, limited by their feature extraction capabilities.

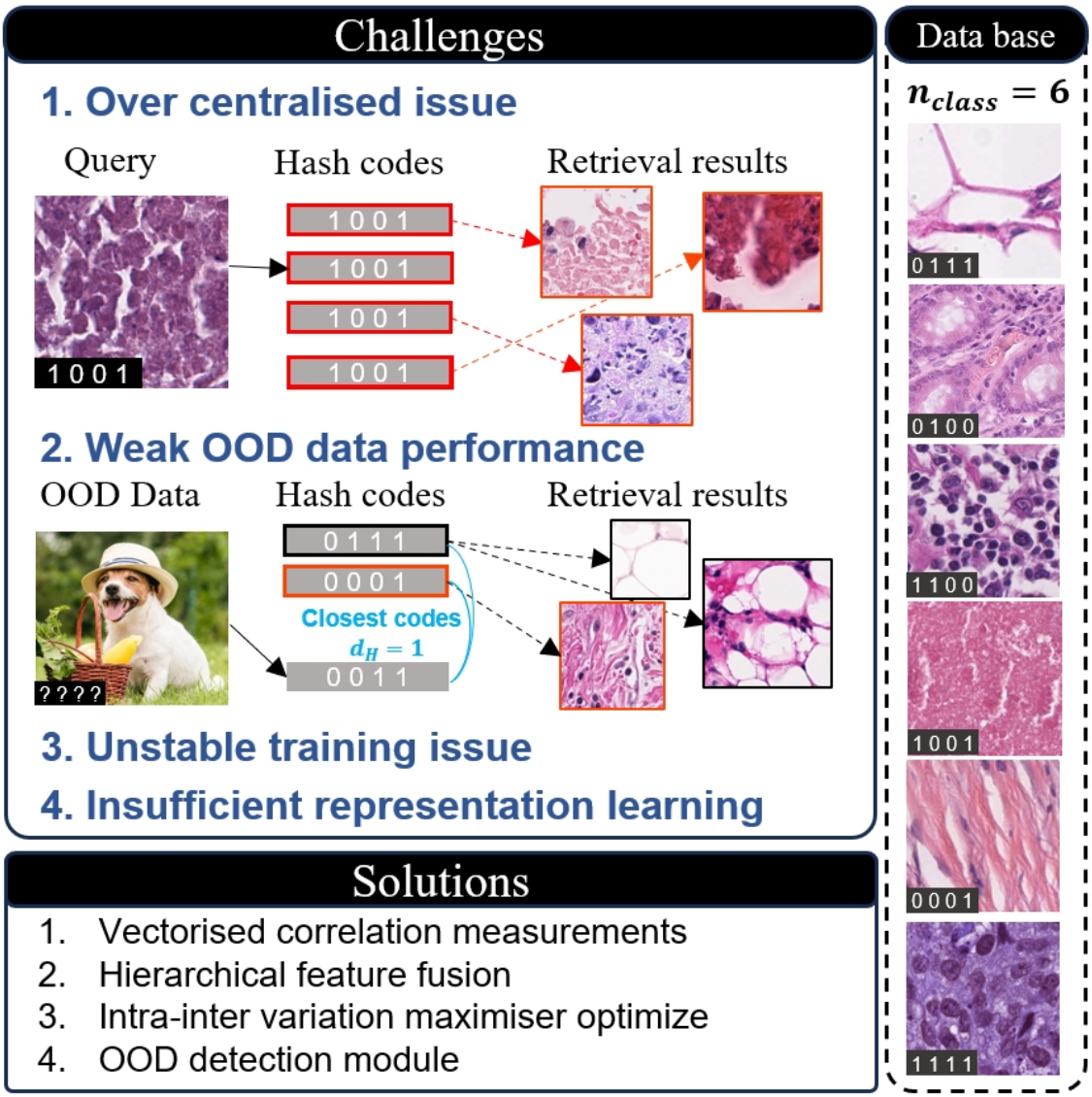


Figure 7.10: Challenges in Retrieval Tasks for Medical Images: This includes issues like over-centralization and poor performance on Out-of-Distribution (OOD) data. Over-centralization refers to assigning identical hash codes to samples within the same category without accounting for variations within the class. Weak OOD performance implies the model's inability to identify OOD samples, forcing it to incorrectly associate them with known classes.

In the realm of AI systems, accuracy, reliability, and safety are paramount, indicating the necessity for robust retrieval systems capable of recognizing out-of-distribution (OOD) samples. Recent advancements in deep hashing [329–331] have shown promise. Deep hashing is typically divided into two categories: paired (e.g., doublet, triplet) and pointwise approaches. Paired hashing, initially integrated with deep neural networks [332, 333], employs optimization functions to cluster similar pairs and disperse dissimilar ones. Pointwise hash learning, exemplified by CSQ [334], simplifies retrieval tasks to classification challenges by training networks to learn

predefined target hash codes.

Despite progress, image retrieval faces several persistent challenges:

1. Training instability and inefficiency, often caused by probability functions that transform Hamming distance into probabilities [335, 336].
2. Over-centralization, where pointwise hashing leads to minimal intra-class variation [334, 337].
3. Limited representation learning, with some methods underperforming using compact hash codes and inadequate utilization of intermediate feature representations.
4. Poor performance with OOD data, reflecting weak generalization and robustness capabilities (Fig. 7.10).

To address these issues, this section introduces an innovative approach for medical image retrieval, integrating deep hashing with OOD detection. This approach utilizes the Pearson correlation coefficient for hash code distance evaluation, replacing traditional probability functions. It also incorporates multi-head self-attention layers into standard CNN architectures to enhance feature representation [338]. Moreover, it leverages intermediate feature representations for hash code generation, providing a more nuanced feature perspective. An OOD detection module is included to identify anomalous samples. The efficacy of this method is demonstrated through rigorous testing on diverse datasets, including CIFAR-10, BCSS, and RadIN-CTL, showing significant improvements in retrieval performance. The study's contributions are as follows:

1. A novel architecture blending convolutional and multi-head self-attention layers for hash encoding.
2. The use of the Pearson coefficient for distance measurement, addressing over-centralization and gradient issues.
3. An OOD detection module based on image reconstruction disparity.

4. Establishing a comprehensive benchmark for evaluating medical image retrieval across various databases.

7.2.2 Related works

In the context of a given query image, the objective of a deep hashing model, denoted as $\phi(\Theta)$, is to identify related images. This is achieved by transforming a set of raw images $X = \{x_1, x_2, \ldots, x_N\}$ into K-bit binary hash vectors $H = \{h_1, h_2, \ldots, h_N\}$, where the mapping function ϕ is defined as $\phi : X \to H$ and each h_N is within the range $[-1, 1]^K$. Subsequently, the vectors H are converted into binary hash codes $B \in \{0, 1\}^{N \times K}$ through a sign function:

$$B = \text{sgn}(f(\phi(X; \Theta))), \tag{7.3}$$

Here, Θ represents the network's trainable parameters, f is an activation function which could be either linear or nonlinear, and sgn is the signum function that outputs the sign of its input elements.

To facilitate the training of the model, the optimization functions in deep hashing generally comprise a distance loss L_D and a quantization loss L_Q which is scaled by a weighting parameter α:

$$L = L_D + \alpha L_Q, \tag{7.4}$$

Distance Loss L_D. In the scenario of pointwise hashing as described in references [334, 337], the network endeavours to learn specific hash codes t. In this setup, L_D is defined using the binary cross-entropy loss between the generated hash vectors h and the target t:

$$L_D = \frac{1}{K} \sum_{i=1}^{C} [t_{i,k} \log(h_{i,k}) + (1 - t_{i,k}) \log(1 - h_{i,k})] \tag{7.5}$$

where C is the number of classes and K is the bit-length of the hash codes. Nonetheless, it is argued that pointwise hashing may lead to an excessive concentration of samples, assigning identical hash codes to samples from the same class, which diminishes the granularity needed

for effective similarity assessment. In contrast, paired hashing evaluates the 'distance' between paired samples primarily through the conventional Hamming distance d_H or a softened version of it d_RH, as discussed in reference [339]. The conventional Hamming distance $d_H = \|h_i - h_j\|_2^2$ calculates the count of differing elements between the hash vectors h_i and h_j. For example, Liu et al. [329] proposed a marginal distance loss:

$$L_D = (1 - y)d_H + y \max(m - d_H, 0), \tag{7.6}$$

where m is a predefined margin value, and y signifies the similarity between paired images, with 0 for similar pairs. The relaxed Hamming distances $d_RH = P(d_H)$ predominantly apply probability functions P to scale the Hamming distance into the interval [0, 1]:

$$L_D = (1 - y) \log(d_RH) + y \log\left(P(\max(m - d_H, 0))\right), \tag{7.7}$$

The commonly used probability functions include the sigmoid σ [335, 340, 341] (adjusted as $P(d_H) = \sigma(d_H - \frac{K}{2})$), Cauchy [336], Gaussian [342], and tanh [343]. However, these functions tend to exhibit steep or flat shapes for small hamming distances, leading to potential training instabilities such as gradient vanishing or exploding.

Quantization Loss L_D. For quantization loss L_Q, the goal is to quantize hash vectors into binary hash codes. This is typically represented as

$$L_Q = \|(1 - |h|)\|_1, \tag{7.8}$$

where 1 is a vector of ones with length K [331]. Additionally, some researchers have utilized sigmoid or tanh activation functions to limit the network output to the range $[-1, 1]$.

7.2.3 Methods

Problem Definition and Deep Vectorised Hashing

Consider a set of images X with their corresponding labels $L = \{l_1, l_2, \ldots, l_N\}$. The similarity $s_{i,j}$ between two image pairs x_i and x_j is determined based on their class labels as follows:

$$s_{i,j} = \begin{cases} 1, & \text{if } l_i = l_j \\ 0, & \text{if } l_i \neq l_j \end{cases} \tag{7.9}$$

Given the similarity matrix $S = \{s_{0,0}, s_{0,1}, \ldots, s_{N,N}\}$ for N paired samples, the model $\phi(\Theta)$ is trained by optimizing the a posteriori estimation of hash vectors $H = \{h_1, h_2, \ldots, h_N\}$ as:

$$\begin{aligned} \hat{\Theta}_{\text{MAP}} =& \arg\max_{\theta} P(H \mid S) \\ =& \arg\min_{\theta} -\log P(H \mid S) \\ =& -\arg\min_{\theta} (\log P(S \mid H) + \log P(H)), \end{aligned} \tag{7.10}$$

where $\log P(H)$ represents the prior probability of H, and $\log P(S \mid H)$ is the log-likelihood estimation. The conditional probability of the similarity $s_{i,j}$ given two hash vectors h_i, h_j is calculated by:

$$P(s_{i,j} \mid h_i, h_j) = \begin{cases} \mathcal{P}(D(h_i, h_j)), & \text{if } s_{i,j} = 1 \\ 1 - \mathcal{P}(D(h_i, h_j)), & \text{if } s_{i,j} = 0 \end{cases} \tag{7.11}$$

where $\mathcal{D}$ denotes the distance measurement assessing the proximity between h_i and h_j, and $\mathcal{P}$ is the probability function that converts distances, such as the hamming distance, into probabilities.

Numerous investigations have quantified D by computing the relaxed hamming distance $d_{RH} = \mathcal{P}(d_H)$, employing a variety of probability functions $\mathcal{P}$ such as sigmoid, tanh, Gaussian, or Cauchy functions. However, these functions (see Fig. 7.11(a)) typically exhibit a steep (or flat) profile for small (or large) hamming distances. Furthermore, when considering a hamming

sphere with centre h_Q and radius r, these relaxed hamming distances predominantly evaluate the magnitude between samples. For instance, samples a and c in Fig. 7.11(b) are considered equivalent as their hamming distances to the centre h_Q are identical. Consequently, cosine similarity has been utilized to determine the orientation of hamming vectors [344], which only accounts for the directionality and disregards magnitude; for example, samples a and b exhibit identical cosine similarity (direction) in relation to the query image. In this work, we assert

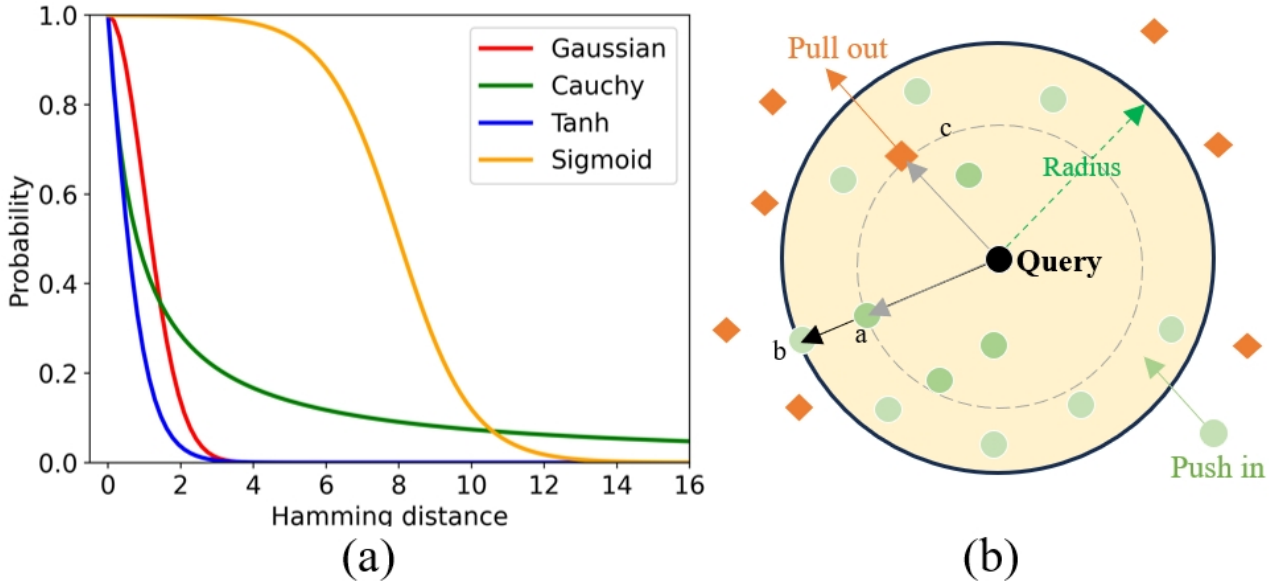


Figure 7.11: Limitations of Measurement Techniques: (a) Relaxed Hamming Distance shows nearly constant patterns for small Hamming distances and sharp patterns for large distances. (b) Cosine Similarity evaluates only the directional aspect of hash vectors, overlooking other characteristics.

that the hash space is multidimensional and the distinction between various hash vectors should incorporate both magnitude and direction. The normalized cosine similarity is introduced to quantify the directional discrepancies between two vectorized hash codes as:

$$S_{PCC}(x, y) = \frac{\sum(x_{p,q} - \bar{x}_q)(y_{p,q} - \bar{y}_q)}{\|x_{p,q} - \bar{x}_q\|_2 \|y_{p,q} - \bar{y}_q\|_2}, \tag{7.12}$$

where $x, y \in \mathbb{R}^{N \times K}$ are K-bit hash vectors. This coefficient, ranging from $[-1, 1]$, aligns well with Bayesian learning and can be effectively scaled to $[0, 1]$. Utilizing this similarity metric, the conditional probability in Eq. (6.27) is expressed as:

$$P(s_{i,j} \mid h_i, h_j) = S_{PCC}(h_i, h_j)^{s_{i,j}} (1 - S_{PCC}(h_i, h_j))^{1 - s_{i,j}}, \tag{7.13}$$

and the $\log P(S \mid H)$ is computed by:

$$\log P(S \mid H) = s_{i,j} \log S_{PCC}(h_i, h_j) + (1 - s_{i,j}) \log(1 - S_{PCC}(h_i, h_j)), \tag{7.14}$$

The prior probability for quantization is formulated as:

$$\log P(H) = \log(1 - S(|h_i|, 1)), \tag{7.15}$$

where 1 is a vector of ones of the same length as the hash code h_i. To harmonize the quantization loss with distance loss within the same domain, we revise the equation by introducing a constant 1 to ensure the outcome surpasses 0:

$$\log P(H) = \log(2 - S(|h_i|, 1)). \tag{7.16}$$

Besides the innate limitations of these probability functions, the challenge of class imbalance also persists. This common issue tends to skew the retrieval towards more populous classes. While current methodologies to mitigate data imbalance are discussed in [335,336], they are primarily concentrated on balancing similar and dissimilar pairs, neglecting the variations within classes. In response, we propose a method to fine-tune the log-likelihood estimate, thus providing a counterbalance to the bias induced by class imbalance in retrieval tasks. Particularly, for dissimilar pairs, this method accounts for the prevalent data imbalance across various classes. With N samples distributed among C categories, the global class weights $w = \{w_1, w_2, \ldots, w_C\}$ can be determined by:

$$w_i = \frac{N}{N_i \cdot C} \tag{7.17}$$

where N_i represents the number of samples in class i. To integrate these global class weights into pairwise hashing, the class weight map for each batch is derived from:

$$W_M = w^T \cdot w \tag{7.18}$$

Consequently, the MAP estimation synergizes with gradient descent through the vectorised

correlation loss (VCL), defined as:

$$VCL = -\arg\min_{\theta} \left(W_M \cdot \left(s_{i,j} \log S(h_i, h_j) + (1 - s_{i,j}) \log(1 - S(h_i, h_j))\right) + \beta \log\left(2 - S(|h_i|, 1)\right)\right), \tag{7.19}$$

where β denotes the weight assigned to the quantization loss.

Details of Deep Vectorised Hashing

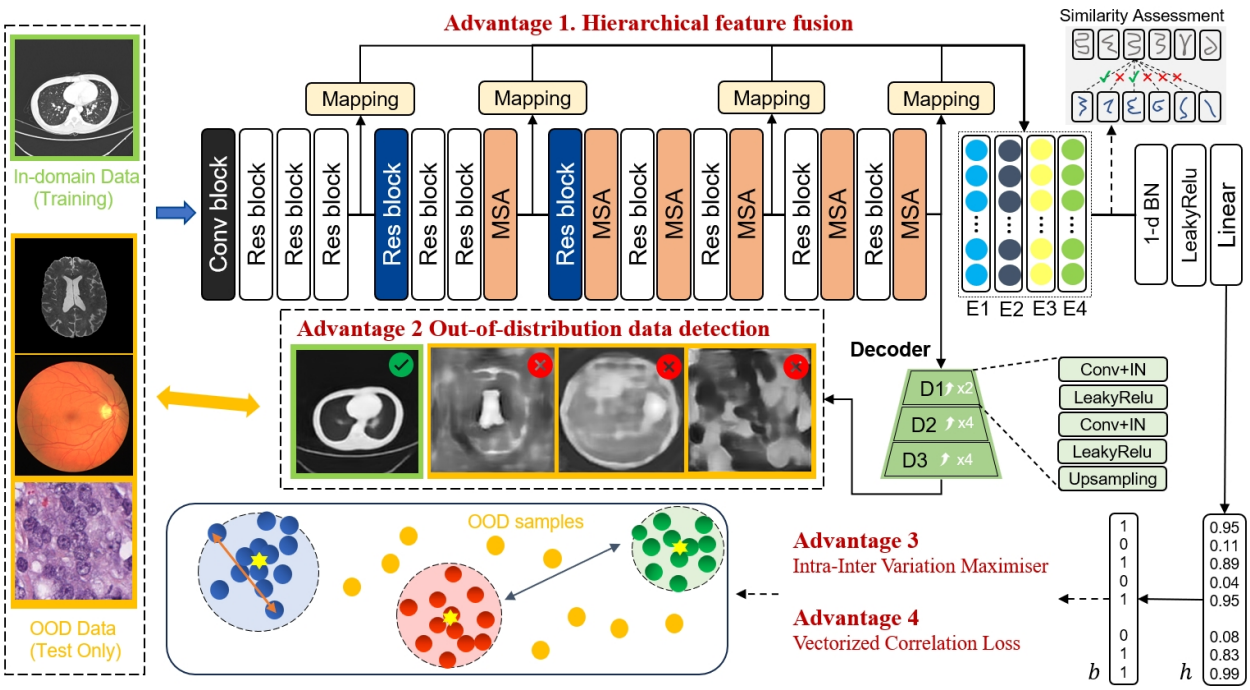


Figure 7.12: DVH integrates three key components: (1) A feature extraction stem with multi-head self-attention (MSA) and hierarchical feature fusion; (2) Vectorised correlation loss (VCL) applied to vectorised hash codes, enhanced by the IntraInterVariationMaximizer (IIVM) mechanism; (3) An Out-of-distribution (OOD) data detection module. Notably, during training, DVH's output h ranges from [-1, 1], with OOD samples incorporated only in the testing phase.

The DVH framework integrates multi-head self-attention layers, multi-level fusion layers, and an out-of-distribution (OOD) detector, as depicted in Fig. 7.12. Given a query image X_Q, DVH produces the corresponding hash vector h_Q, the reconstructed image R_Q, and hierarchical feature embeddings E_Q through the function ϕ parameterized by Θ:

$$h_Q, R_Q, E_Q = \phi(X_Q; \Theta) \tag{7.20}$$

DVH is architectured on the principles that:

Enhanced feature representations are pivotal for retrieval efficiency. Contrasting with classification-centric tasks, image retrieval resembles clustering, where the goal is to project images into compressed hash vectors while promoting category-based aggregation. It has been proved that retrieval can be significantly improved with superior feature representation learning capabilities. To achieve such representations, multi-head self-attentions (MSAs) are amalgamated into the baseline ResNet 50 [52]. Research indicates MSAs can centralize feature embeddings, in contrast to convolutions that tend to distribute features [338]. MSAs are utilized for spatial smoothing of the feature map, hence, they are inserted into a customized ResNet50 architecture at the culmination of each stage in Fig. 3. The term "Res block" signifies residual blocks [52] comprising convolution layers, batch normalizations, leaky ReLU activations, and skip connections.

Hash codes are binary representations encompassing comprehensive image features. For current stuides, images were first encoded into 512-dimensional vectors by the backbone network, then reduced the dimension to hash code length. Nevertheless, it is contended that a single 512-dimensional vector inadequately represents the intricate details of images, thus constraining retrieval capabilities.

To remedy this, a feature pyramid is constructed using 'Mapping' layers atop the ResNet50. Positioned at each stage's end, these 'Mapping' blocks consist of adaptive average pooling and a linear layer to transform diverse stage feature maps into 512-dimensional vectors. These vectors are then concatenated and processed through a 1-d batch normalization, a LeakyReLU, and a final linear layer to generate the hash vectors. The 1-d batch normalization ensures the embeddings stay within a normalized range, aiding in the stabilization of the training.

Self-supervised learning can be implemented for OOD detection. In real applications, the reconstruction model exhibits higher reconstruction errors for OOD samples due to their divergence from in-distribution data. Based on this, we modified the decoder, which aims to rebuild the original input from condensed latent space representations, by excluding skip connections to prioritize the last encoder layer's feature representations. This enhances the

quality of the learned features. The decoder's structure includes convolution blocks with layers, instance normalization, leaky ReLU activations, and upsampling, optimized using an l1 loss function. To optimize computation, the upsampling scales for D1, D2, and D3 are strategically set to 2, 4, and 4, respectively.

To enhance the assessment of reconstruction errors in grayscale images, the multi-scale structural similarity index (MSSSIM) [345] is utilized. Incorporating exponential weights α_M, β_j, and γ_j, MSSSIM computes the resemblance between two images x and y by the formula:

$$\text{MSSSIM}(x, y) = [l_M(x, y)]^{\alpha_M} \prod_{j=1}^{M} [c_j(x, y)]^{\beta_j} [s_j(x, y)]^{\gamma_j}, \tag{7.21}$$

where M denotes the number of scales, and l_M, c_j, and s_j are the luminance, contrast, and structure comparison metrics, respectively. Hence, OOD instances are identifiable by their MSSSIM scores, contrasting the source images with their reconstructed counterparts. For RGB images, OOD samples typically manifest a 'domain shift', evident through changes in mean values. Consequently, the OOD threshold is determined based on the mean disparity between the original and reconstructed images.

Specifically, for the i-th original image X_i and its reconstructed counterpart R_i, the OOD detection threshold is established as:

$$\tau = \frac{1}{N} \sum_{i=1}^{N} D(X_i, R_i) \pm 1.5 \cdot \delta(D(X, R)), \tag{7.22}$$

where i enumerates the images in the dataset, D signifies the distance measure such as MSSSIM, $l1$ error, and δ is the standard deviation of these distances. For the distance function D, we apply MSSSIM (as in Eq. (6.37))) for RGB images and the $L1$ norm for grayscale images.

Intra-Inter Variation Maximiser

The performance of DVH is also augmented by the Intra-Inter Variation Maximiser (IIVM), an effective strategy that aids the model in finding an improved optimum. Traditional deep-

learning techniques often modulate the learning rate dynamically by observing a selected metric and its progression. The metric's behavior significantly impacts the training steps in terms of size and direction. While most hashing methods have relied on mean average precision (mAP) or validation loss for learning rate adjustment and model evaluation, these may not sufficiently reflect true performance. Particularly, mAP assessments within hamming balls do not account for intra-class variability, expecting hash values of query samples to strictly match those at the class centre.

To overcome this, we introduce a soft mean average precision (smAP) to approximate mAP at varying hamming ball radii. The monitored metric during training is thus the maximal mAP calculated within hamming balls with a radius of $K/4+1$. For example, for an 8-bit hash code, the smAP is defined as:

$$\text{smAP} = \max(\text{mAP}_0, \text{mAP}_1, \text{mAP}_2), \tag{7.23}$$

where mAP_i denotes the mAP at a hamming ball of radius i. The optimal radius determined during training dictates the hamming ball radius for inference.

Ranked Retrieval

Given a query image X_Q, DVH filters potential candidates X_S from the database X_D by calculating the hamming distance within hamming balls of radius $r = K/4 + 1$. The msssim score is computed between X_Q and its reconstructed counterpart R_Q. If the msssim score is below the OOD threshold τ, X_Q is classified as OOD. The average S_PCC between multi-level feature embeddings E_1, E_2, E_3, E_4 is then calculated to evaluate content similarity. Finally, this composite similarity score is used to order the retrieved samples.

7.2.4 Experiments and Results

Datasets and training details

To validate the efficacy and stability of DVH, we conducted trials on three publicly accessible datasets: a natural image dataset and two medical image datasets, with details regarding the

medical datasets provided in Table 7.5.

Table 7.5: Details of the BPS and RadIN-CTL Datasets.

BPS Dataset									
Subset	Image size	Magnification	Tumour	Stroma	Inflammatory	Necrosis	Fat	Gland	Total
Train & Valid	256x256	20x,40x	6459	3852	2598	1676	1474	1419	17478
Query set	256x256	20x,40x	200	200	200	200	200	200	1200
RadIN-CTL Dataset									
Subset	Image size	Planes	Normal	Nodule	Bronchiectasis	PD*	ILD*	AO*	Total
Train & Valid	224x224	Axial, Coronal	4273	2963	1519	1153	2921	3664	16493
Query set	224x224	Axial, Coronal	1425	988	506	385	974	1221	5499

* PD indicates parenchyma destruction, ILD refers to interstitial lung disease, and AO indicates airspace opacification.

CIFAR-10. The CIFAR-10 dataset comprises 60,000 32×32 colour images distributed across 10 classes. The designated training set containing 50,000 images was utilized for model training and validation, while the official test set of 10,000 images served as the query set.

Basic Pathological Structures (BPS) Dataset: The BPS dataset includes images derived from references [346] and [347]. Following semantic annotations, 256×256 image patches were extracted and categorized into six classes. The selection was contingent upon the mask ratio for a given class exceeding 0.9.

RadIN-CTL: The RadIN-CTL dataset, constructed from RadioImageNet [348], with 1.35 million annotated images from 131,872 patients. It covers six clinical patterns: normal, nodule, bronchiectasis, parenchyma destruction, interstitial lung disease, and airspace opacification. Predominantly, the images are axial slices, with a minor subset ($\sim$1%) comprising coronal slices.

Training Strategy: We employed He-normal initialisation for all trainable parameters. Data augmentation during training on BPS and RadIN-CTL datasets included randomized flips (up, down, left, right) and randomized resized cropping (scale $\in [0.8, 1.0]$). Specifically for BPS, adjustments to brightness, contrast, saturation, and hue were randomly applied. CIFAR-10 was exempt from augmentation. All models underwent training on an NVIDIA RTX 3090 Ti GPU over 200 epochs, with an initial learning rate of $1e^{-4}$ and a decay factor of 0.9 after a tolerance of five epochs. The coefficient β in Eq. (6.35) was maintained at 0.5 throughout training.

Table 7.6: Ablation studies

Model	8 bits (mAP)	16 bits (mAP)	32 bits (mAP)
Baseline	0.5333	0.5528	0.5486
Baseline + VCL	0.6835	0.6884	0.6866
Baseline + VCL + MSAs	0.7001	0.7047	0.7041
Baseline + VCL + MSAs + HFF	0.7122	0.7149	0.7167
Baseline + VCL + MSAs + HFF + OOD	0.7308	0.7345	0.7359
Baseline + VCL + MSAs + HFF + OOD + IIVM	0.7457	0.7571	0.7532

VCL: Vectorised correlation loss, MSAs: multi-head self-attention, HFF: hierarchical feature fusion, OOD: out-of-distribution detection module, IIVM: inter-intra variation maximiser protocol.

Experimental Settings and Evaluation Metrics

Comparison Studies. We conducted comprehensive comparison studies on three datasets, assessing various models: central similarity quantisation (CSQ) [334], deep polarised network (DPN) [337], deep Cauchy hashing (DCH) [336], deep supervised discrete hashing (DSDH) [349], deep balanced discrete hashing (DBDH) [340], deep fisher hashing (DFH) [350], pairwise correlation-based deep hashing (PCDH) [351], quadratic spherical mutual information hashing (QSMIH) [352], HashFormer [353], TransHash [354], and Attention-based Triplet Hashing (ATH) [355]. All models were trained from scratch, following their specific training strategies.

Ablation Studies. To elucidate the impact of different modules - VCL, MSAs, HFF, OOD, and IIVM - ablation studies were performed using the RadIN-CTL dataset.

Out-of-Distribution (OOD) Assessment. An additional class, "OOD", was integrated into the RadIN-CTL dataset to test model robustness. This included 200 query images randomly chosen from retina [356], brain MRI [357], BPS, and VOC2012 [358] datasets. Samples unrelated to lung CT modality were classified as OOD. Hash centres were determined by averaging their hash codes, and samples were identified as OOD if their hash codes fell outside the Hamming balls (r=K/4+1) relative to their hash centre. The resilience of the top-6 methods against OOD data was then evaluated.

Ranked Retrieval. For the RadIN-CTL dataset, the top-4 similar and dissimilar samples were selected from the retrieval results, which were sorted by Hamming distance. The performance of three methods - DVH, CSQ, and HashFormer - was evaluated to explore different hashing approaches, including point-based, pairwise, and Transformer-based methods.

Evaluation Metrics. Performance was assessed using mean Average Precision (mAP). True

positive predictions were counted when the retrieved samples matched the query image's class and hash code, and when samples were within the specified Hamming ball radius.

Experimental Results

Table 7.7: Comparison studies on CIFAR-10, BPS, and RadIN-CTL

Model	CIFAR-10			BPS			RadIN-CTL		
	8 bits	16 bits	32 bits	8 bits	16 bits	32 bits	8 bits	16 bits	32 bits
CSQ	0.7188	0.7251	**0.7361***	0.9706	0.9726	0.9678	0.6897	0.6950*	0.6829
DPN	0.7344	**0.7449***	0.7414	**0.9701***	0.9690	0.9657	0.6631	0.6748*	0.6640
DCH	0.7284	**0.7411**	0.7357	0.9692*	0.9650	0.9647	0.5925	0.5982*	0.5794
DSDH	0.7593	0.8244	**0.8240***	0.9668*	0.9538	0.9048	0.5339	**0.5484***	0.4759
DBDH	0.7057	0.8386	**0.8446***	0.8548	0.8797*	0.8225	0.4881*	0.4823	0.4448
DFH	0.3873	0.4871	**0.5516***	0.8111	0.9690	0.9718	0.4903	0.4917	**0.5025***
PCDH	0.1000	0.3062	**0.7123***	0.6062	0.9121	0.8013	0.4931	**0.5024***	0.4559
QSMIH	0.6456	**0.6704***	0.6407	0.9690*	0.9557	0.9083	0.6144*	0.6018	0.4663
HashFormer	0.7857	**0.7892***	0.7704	0.9684*	0.9661	0.9580	0.6866	0.6895*	0.5268
TransHash	0.7443	0.7601	**0.7596***	0.9618	0.9621*	0.9588	0.6592	0.6684*	0.4508
ATH	**0.3345***	0.2641	0.2327	**0.5881***	0.5337	0.5475	0.5982*	0.5770	0.5216
DVH (our proposed)	0.8512	**0.8547**	0.8543	0.9831	**0.9849**	0.9833	0.7457	**0.7571**	0.7532

* indicates significant differences (with Wilcoxon signed-rank test p-value¡0.05) compared with the best baseline (among 8, 16, and 32 bits) for each comparison method was bolded, the best and the least good results were green and red bolded, respectively. It is of note that all the retrieved samples were included assessment instead of top-k results.

This section presents the experimental results all comparisons. **Ablation Studies.** Table 7.6 shows the ablation study results. The baseline DVH model achieved mAPs of 0.5333 and 0.5528 for 8 and 16 bits, respectively. Significant improvements (0.11 mAP for both 8-bit and 16-bit) were observed when integrating vectorised correlation losses. Incorporating MSAs blocks led to a 1.7% mAP increase. Hierarchical feature fusion provided a further 1.2% increase in mAP. The implementation of a multi-task learning scheme (image reconstruction) resulted in nearly a 2% mAP improvement, and the introduction of IIVM further stabilized the model, contributing to nearly a 2% mAP promotion.

Comparison Experiments on CIFAR-10. In the CIFAR-10 dataset (Table 7.7), DVH achieved the best performance, surpassing 0.85 mAP, followed closely by DSDH (0.8244 mAP) and DBDH (0.8446 mAP). HashFormer and TransHash showed comparable results, with 0.7892 and 0.7601 mAP, respectively. DFH had a weaker performance at 0.5516 mAP, possibly due to inappropriate hyperparameters. PCDH failed in 8 and 16 hash-bit models (0.1 and 0.3062 mAP, respectively) but performed well in the 32 bits model (0.7123 mAP).

Comparison Experiments on the BPS Dataset. On the BPS dataset (Table 7.7), DVH outperformed others with 0.9831, 0.9849, and 0.9833 mAPs for 8-bit, 16-bit, and 32-bit models, respectively. DFH was comparable in the 32-bit model but underperformed in smaller hash sizes (e.g., 0.8111 mAP for 8 bits). CSQ-16bit also showed similar results to DVH-16bit with a mAP of 0.9726. Transformer-based methods, HashFormer-8bits and TransHash-16bits, demonstrated impressive performance with 0.9684 and 0.9621 mAP, respectively.

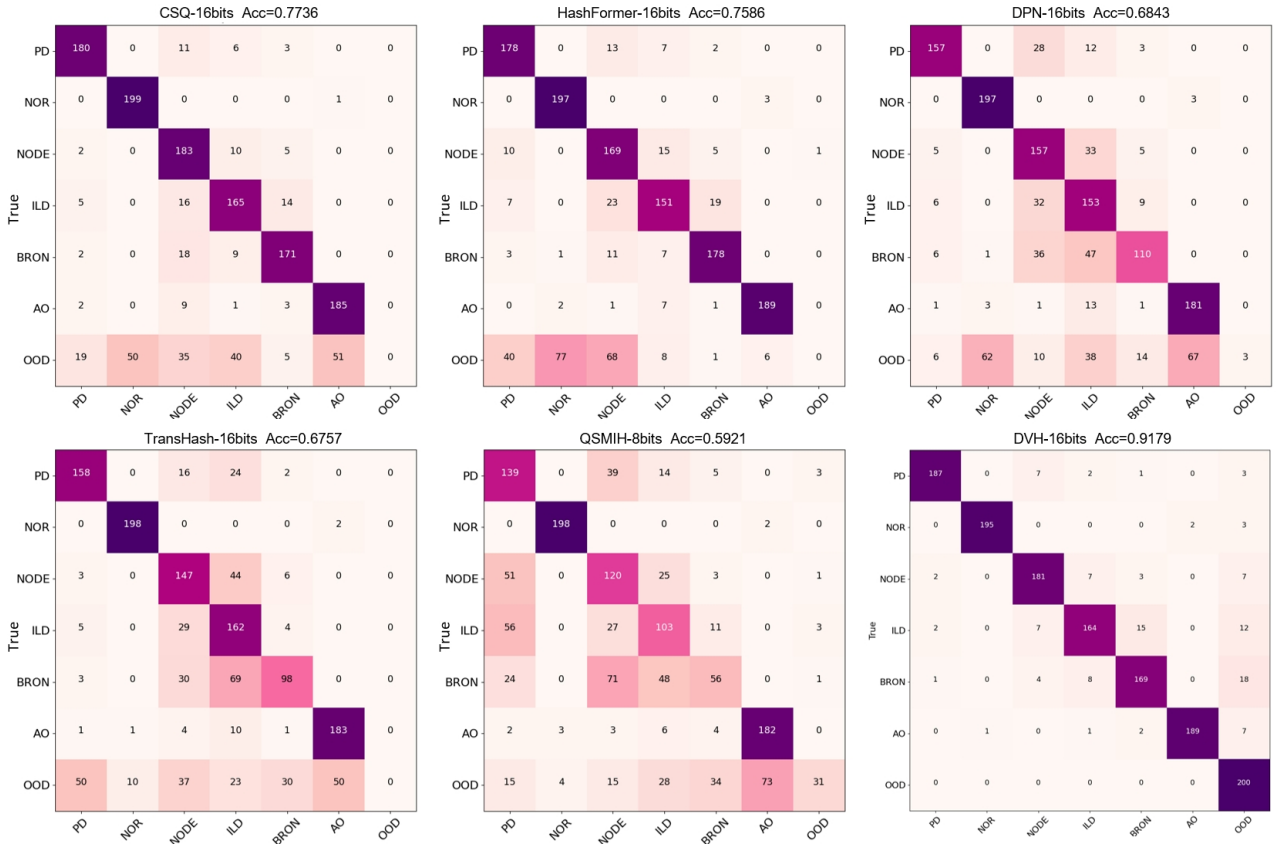


Figure 7.13: Confusion Matrix for HashFormer, CSQ, DPN, and DVH on the RadioIN-CTL Test Set (n=1200) with an Additional 200 OOD Samples from Various Datasets. The classification was determined by calculating the Hamming distance to the hash center of each class from the training set, considering samples with $h_D < \frac{K}{4} + 1$ as true positives. Labels: 'PD' for Parenchyma Destruction, 'NOR' for Normal, 'NODE' for Nodule, 'ILD' for Interstitial Lung Disease, 'BRON' for Bronchiectasis, and 'AO' for Airspace Opacification.

Comparison Experiments on RadIN-CTL. For the RadIN-CTL dataset (Table 7.7), DVH achieved the highest mAP (0.7571), significantly outperforming other methods. CSQ and HashFormer-16bits followed with mAPs of 0.6950 and 0.6895, respectively. Surprisingly, DBDH, DFH, PCDH, and DSDH showed lower-than-expected mAPs (0.4881, 0.4903, 0.4931,

and 0.5339, respectively). The attention-based triplet hashing (ATH) method underperformed, achieving a mAP of 0.5982.

Out-of-Distribution Assessment. The OOD evaluation results (Fig. 7.13, $\tau = 0.8$) indicated strong performance in predicting normal patients and those with airspace opacification. Some models struggled to distinguish bronchiectasis (DPN, TransHash, QSMIH), nodules (TransHash, QSMIH), and interstitial lung disease (QSMIH). CSQ, HashFormer, DPN, and TransHash failed to predict OOD samples accurately (TP_OOD=0), except for DPN (TP_OOD=3). QSMIH recognized 31 OOD samples but had a low overall accuracy (0.5921). DVH excelled in OOD detection (accuracy=0.9179), correctly identifying all OOD samples but misclassifying 12 in-distribution ILD samples as OOD.

7.2.5 Discussion

In this study, we introduce a novel methodology that integrates Vectorised Correlation Losses, Multi-Head Self-Attention, Hierarchical Feature Fusion, and Out-Of-Distribution (OOD) detection for medical image retrieval. This method has been rigorously evaluated on a spectrum of images, encompassing both pathological and radiological types, as well as natural images. Our approach has demonstrated superior performance in overcoming existing challenges in various image databases when compared to prior work.

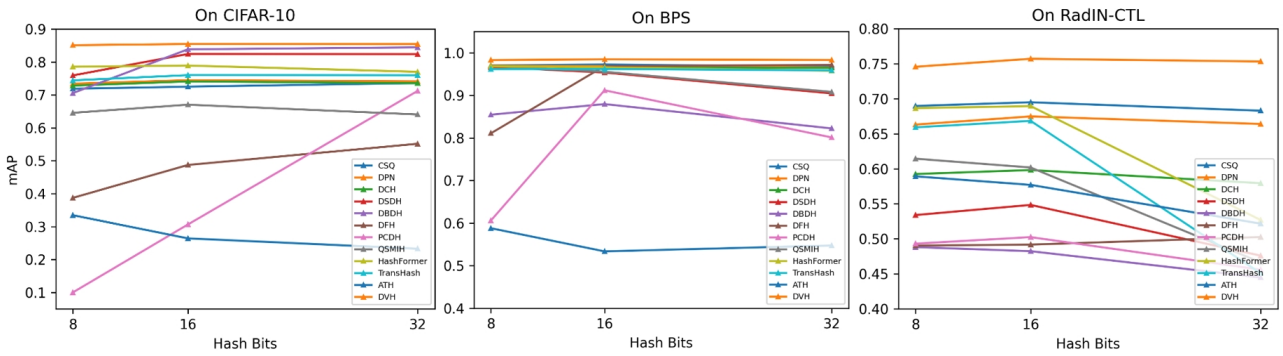


Figure 7.14: mAP on CIFAR-10, BPS, and RadIN-CTL datasets.

Overall Performance. The experimental analysis on the BPS and CIFAR-10 datasets, which display substantial inter-class variability relative to the RadIN-CTL dataset, shed light on the general retrieval efficiency. Meanwhile, the studies on RadIN-CTL, targeting lung lesions, focus

on the precision of similar lesion retrieval within a particular category, thus highlighting the models' granular retrieval prowess. Notably, DVH exhibited robust and consistent performance across natural, pathological, and radiological modalities (refer to Fig. 7.14), outperforming other existing methodologies. Point-wise hashing approaches like CSQ and DPN displayed capabilities in both general and detailed retrieval. However, these methods centralized samples excessively towards the hash centre, failing to present organized ranked retrieval outcomes (see Fig. 7.13). Some methods demonstrated unstable results across diverse datasets. For instance, DBDH had a lower performance on RadIN-CTL dataset (mAP of 0.4881) but fared comparatively on the CIFAR-10 dataset (mAP of 0.8405 with 16 bits). This variance indicates differing retrieval competencies for general versus fine-grained samples. The inconsistent performance of DFH across datasets (ranking 11th, 3rd, and 12th on CIFAR-10, BPS, and RadIN-CTL, respectively) points to its limited generalization capacity and robustness. Moreover, the requirement of manual tuning of five main hyperparameters for DFH suggests a dependency on hyperparameter optimization. Interestingly, methods based on Transformers exhibited greater robustness than CNN-based methods, securing top-5 positions on all datasets. Overall, DVH outclassed existing retrieval systems, setting a new benchmark in performance.

Ablation Studies. As depicted in Table 7.6, the Vectorised Correlation Loss contributes significantly to the retrieval accuracy, with nearly a 15% improvement in mAP over the baseline. This underscores the efficacy of integrating correlation-based distance metrics as opposed to traditional relaxed hamming distances. By factoring in both the magnitude and direction of hash vectors, the model attains an all-encompassing grasp of the inherent similarities between features, enabling it to discern subtle associations and variances, thereby enhancing retrieval precision. The marked enhancement with the combination of "Baseline+VCL+MSAs+HFF" implies that augmenting feature representation learning can further elevate retrieval performance. The addition of self-supervised learning also bolsters retrieval efficacy, as the network is driven to learn more generalizable features through the task of reconstructing the raw data from reduced-dimension features. This extraction of meaningful features fortifies the model's ability to discern underlying similarities and patterns, leading to more accurate image retrieval. The gains accrued via the IIVM approach underscore the importance of monitoring metrics

during training. By utilizing soft mAP scores, the model adeptly negotiates the inter-class and intra-class variations within the hamming space, allowing for a more refined learning rate adjustment and training optimization.

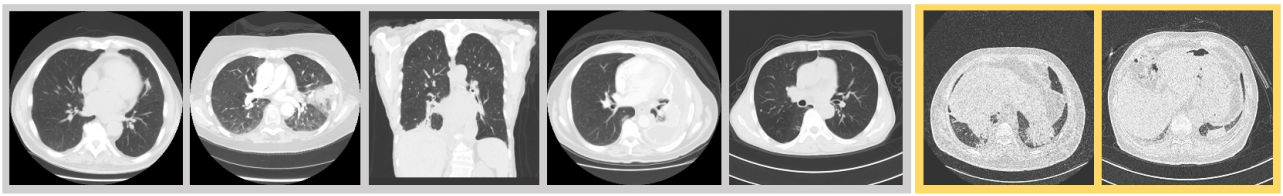


Figure 7.15: OOD samples that misrecognized by first (a-e) and second (f, e) steps.

Out-of-Distribution Assessment of DVH. DVH proves competent in identifying OOD samples, while it may occasionally misclassify known samples as OOD. There are two conditions under which DVH determines a sample as OOD. Firstly, if the hamming distance from the known hash centres exceeds the pre-set radius $r = K/4 + 1$, and secondly, if the reconstruction metrics, such as MSSSIM for grayscale or L1 distance for RGB images, are below the threshold τ (as per Eq. (6.38)). Further analysis into false positives revealed that most incorrectly identified OOD samples were 'hard' samples with minimal class-related features (Fig. 7.15). Misclassifications in the second stage were typically noisy images, explaining their OOD classification. These insights suggest that the OOD detection module has the potential to curate training samples for downstream tasks like classification and segmentation.

Ranked Retrieval. The outcomes of ranked retrieval are showcased in Fig. 7.16. The top-4 related and unrelated samples to the query images are depicted with green and orange backdrops, respectively. It is observed that DVH's retrieval results are orderly sorted by similarity, unlike CSQ and HashFormer. For instance, DVH deftly identifies images with corresponding patterns of airspace opacity as closely related samples (indicated by small green squares in Fig. 7.16), whereas CSQ and HashFormer struggled to order the retrieved items based on hamming distance, with their top-ranked samples showing significant visual disparities from the query.

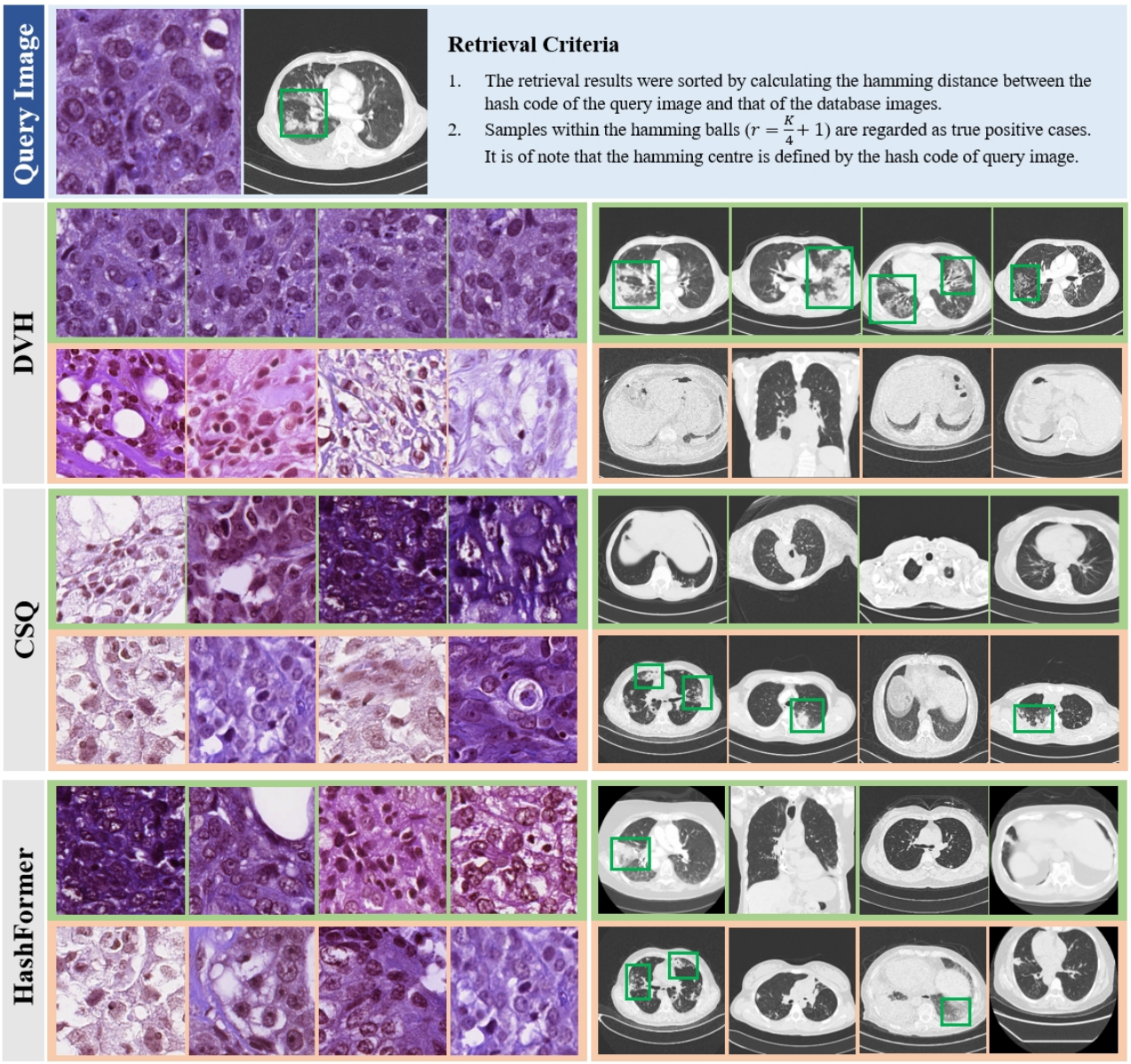


Figure 7.16: Ranked Retrieval Outcomes by Our method, CSQ, and HashFormer. Displays the top-4 similar (in green background) and dissimilar (in orange background) samples for query images. Small green boxes signify that DVH arranges retrieved samples according to patterns observed in the query images.

7.2.6 Conclusion

This manuscript delineates a robust retrieval system, DVH, adept at generating ordered retrieval outcomes through content similarity assessments. DVH surpasses traditional approaches that utilize relaxed Hamming distance-based probability functions by employing the Pearson correlation coefficient to ascertain sample similarity. The integration of a deep hierarchical feature fusion module along with an OOD detection mechanism substantially augments the

model's retrieval efficacy, particularly for out-of-distribution data sets. Rigorous experimental evaluations have established DVH as the frontrunner in context-based retrieval and OOD detection proficiency, surpassing other extant methodologies across both natural and biomedical image data sets. The empirical evidence advocates for DVH's potential as a versatile and scalable solution for image retrieval tasks in a myriad of imaging environments.

Chapter 8

Summary and future work

8.1 Summary of Thesis Achievements

In this thesis, we presented how deep learning-based approaches have addressed complex multi-centre clinical issues by introducing supervised/unsupervised learning and data harmonisation approaches. Additionally, further studies including a benchmark and a clinical translation study were performed to contribute to the medical AI community.

In the realm of supervised learning (Chapters 2 and 3), this research has elucidated the importance of introducing novel techniques to common deep learning models, such as uncertainty estimation, fuzzy logic, adversarial learning, etc. The investigation into Fine-grained recognition in patients with IgAN revealed that the uncertainty of predictions can be used in improving the capacity of classification models, as well as purifying the training datasets for better representation learning. The study also highlighted the positive effects when adopting instance-based structure similarity as the segmentation optimization function to densely distributed targets. The development of the fuzzy attention neural network provided insights into integrating fuzzy logic into deep learning models. The comprehensive experiments demonstrated that by combining fuzzy logic with a well-designed data sampling strategy and optimization functions, the deep learning model can achieve accurate predictions of tiny objects, e.g., small and terminal branches for airways. This research also highlights the importance of addressing

data-imbalanced issues for medical AI applications. Additionally, the Adversarial Transformer for Airway Purification emphasized the achievements when introducing adversarial learning into segmentation models. This study has shown that patch-scale discriminators can significantly enhance airway modelling results in an adversarial learning format, especially for improving the connectivity of terminal airway branches.

In Chapter 4, this thesis presented a novelty work (the DCGN) in the unsupervised learning domain. DCGN incorporated the Gaussian mixture model with a deep neural network and a centralized constraint for fully unsupervised segmentation. The study first investigated the degenerative issues, including missing class, redundant class, dominant class and poor reproducibility. The missing class issue often arises when modules fail to correctly handle outliers, such as classifying outliers as a unique class while merging different categories into a single class. The redundant class issue, known as an additional trick to enhance performance in unsupervised segmentation, arises from the discrete variation within a single class. The instability or low reproducibility stems from random initialization. By evaluating the model on two pathological datasets, the results demonstrated that DCGN outperformed existing unsupervised segmentation methods, addressing or alleviating degenerative issues effectively by centralising the components of deep mixture models during the optimization function calculation. In addition, DCGN achieved comparable results with some semi-supervised and fully supervised vanilla U-Net.

Chapter 5 presented a benchmark study for the challenging task, the airway segmentation. It mainly focused on the data collected from patients with fibrotic pulmonary disease, corresponding to more complex clinical patterns such as honeycombing. To investigate how current approaches perform on this severe disease, the "Airway-Informed Quantitative CT Imaging Biomarker for Fibrotic Lung Disease 2023" (AIIB23) challenge was held in conjunction with the 2023 International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI2023), which provided comprehensive assessments of AI models on airway segmentation and mortality prediction. Based on the obtained results, this study showed that the segmentation performance can be effectively improved by introducing a weighted general union optimization function and continuity losses. Meanwhile, existing approaches were found

to be unstable and weak robust when implemented on low-quality as well as unharmonised data. The radiomic features were also explored, with only certain known radiomic features (13) discovered to be associated with the mortality status. In addition, the normalized total airway volume was found to be a strong imaging biomarker for the prognosis of FLD patients.

Based on the previous study in fibrotic lung disease, Chapter 6 demonstrated the clinical value of imaging-derived biomarkers in mortality prediction. The proposed method SABRE, which has been proven as a valuable tool for quantifying traction bronchiectasis severity, can independently predict survival status against clinical measurements such as FVC, CPI, clinicians' estimations (e.g., average fibrosis score), and existing AI biomarkers. This offered significant prognostic insights into IPF patient management and highlighted its potential for broader application in clinical and research settings.

Chapter 7 first demonstrated the importance of data harmonisation through a comprehensive review, followed by presenting a novel image retrieval model as the first step of data harmonisation. This section delved into different harmonisation approaches in computational healthcare applications. The results showed that the distribution-based harmonisation approaches were popular in structured tabular data while the synthesis and invariant feature learning were commonly used in the image domain. Additionally, it provided many insights into methodology and evaluation metric selection, and a checklist for enhancing the reproducibility of harmonisation studies. Interestingly, this study also highlights the adaptability of computational data harmonisation when combined with federated learning systems. The anomaly-aware deep vectorised hashing was then presented as the preliminary study of data harmonisation. It presented a novel image retrieval system for locating context-based similar targets given a certain reference sample, with a vectorised correlation loss, multi-head self-attention layers, a hierarchical feature fusion and an OOD detection module. The integration of a deep hierarchical feature fusion module along with an OOD detection mechanism substantially augments the model's retrieval efficacy, particularly for out-of-distribution data sets. In addition, DVH deftly identifies images with corresponding patterns as closely related samples, which ranks the retrieved samples by content similarity as the given reference samples.

8.2 Future Work

Although efforts have been made in this thesis to show the feasibility of models in multicentre clinical data analysis, potential research topics remain to be addressed.

High-performance Large foundation models. Both studies in Chapters 2 and 3 revealed the research gaps for accurate segmentation models. For instance, models were struggling to segment adjacent objects (connected airway and lung tissues, glomeruli, cells, etc.) and tiny objects. This indicates the need for high-performance segmentation models in digital healthcare. The foundation models such as the Segment Anything Model (SAM) have dominated the supervised segmentation task recently [359]. These models have shown better performance and generalization ability on common segmentation tasks. However, for complex tasks in biomedical research, large foundation models are still facing some challenges, e.g., the lack of large amounts of panoramic annotated datasets, and unsatisfied segmentation capabilities. As a result, integrating SAM with novel techniques or tricks for complex segmentation tasks becomes an urgent and imperative need. Based on my existing studies, I intend to 1) develop a large foundation model for pulmonary disease in the future; 2) increase the capacity of large foundation models to achieve better performance. The research will be focused on fibrotic lung diseases as they lack prognostication biomarkers to assist clinicians with treatment plans and prognoses. This model would be able to segment different patterns (e.g., nodes, veins, arteries, fibrosis, and airways) with high robustness and precision, providing more insights into disease prognosis.

Vision-language models. Image retrieval is pivotal in data harmonization, and in clinical contexts, the ability to search cases through textual descriptions is highly valued. Chapter 7 has shown the capacity of image retrieval while ignoring the multi-modality retrieval approaches. Leveraging advancements in vision-language models such as CLIP [360], text-image retrieval systems could be substantially enhanced by developing robust correlations between image and text pairs. Future work could focus on establishing bidirectional and multimodal correlations that are fine-tuned for clinical specificity and accuracy. This could extend to various clinical applications, including generating detailed clinical reports from medical imagery, enhancing diagnostic accuracy by retrieving similar cases or anomalies based on descriptive clinical inputs,

and even creating educational tools for medical training. Further, integrating clinical metadata with visual signs could provide a more holistic view of patient cases, aiding in personalized treatment strategies.

Content-based Data harmonisation. Effective data harmonization is crucial in realizing the full potential of AI models, particularly in multicentre studies where data inconsistency can significantly hinder model performance. Building on computational data harmonization, future efforts should aim to develop innovative strategies specifically tailored for cell images in biomedical research, mitigating batch effects and other forms of data variance that can obscure true biological signals [221]. Techniques could be involved using machine learning algorithms to identify and adjust for these variances, ensuring that cell counting, detection, and other downstream tasks are more accurate and reliable. Moreover, there is a potential to extend these harmonization techniques beyond cell imaging to other types of medical and biological data, enhancing the robustness and applicability of computational models across the board. By improving how models handle diverse datasets, we can facilitate broader collaboration in the medical community and accelerate the pace of discovery and innovation. Continued focus on federated learning and privacy-preserving methods will also be essential, ensuring that data harmonization contributes to the development of AI solutions that are both powerful and ethically responsible.

Few-Shot Learning in Medical Imaging. While the development of high-performance models and vision-language approaches is critical, the reliance on large annotated datasets remains a significant bottleneck in medical imaging research. In addition to unsupervised learning which can only handle preliminary tasks, few-shot learning presents a promising avenue to address this limitation by enabling models to generalize from a limited number of examples. Future work will focus on leveraging few-shot learning techniques to enhance the performance of segmentation and classification tasks in medical imaging, particularly for rare diseases and conditions where data scarcity is prevalent. Key areas of exploration will include developing few-shot segmentation models to improve accuracy for rare anatomical structures and disease markers, integrating few-shot learning with large foundation models like SAM to boost their adaptability and precision in complex biomedical tasks, and applying few-shot learning to vision-language models

to improve the generation of clinical reports from medical images. Additionally, extending few-shot learning to multimodal data integration can provide a comprehensive approach to clinical decision-making, correlating sparse clinical data with imaging findings to enhance diagnostic accuracy and personalized treatment planning. By advancing few-shot learning techniques, we aim to reduce the dependency on extensively annotated datasets, thereby accelerating the development and deployment of AI-driven solutions in medical imaging and healthcare.

Bibliography

[1] B. Zhao, Y. Tan, W.-Y. Tsai, J. Qi, C. Xie, L. Lu, and L. H. Schwartz, "Reproducibility of radiomics for deciphering tumor phenotype with imaging," *Scientific reports*, vol. 6, no. 1, p. 23428, 2016.

[2] N. Kumar, R. Verma, S. Sharma, S. Bhargava, A. Vahadane, and A. Sethi, "A dataset and a technique for generalized nuclear segmentation for computational pathology," *IEEE transactions on medical imaging*, vol. 36, no. 7, pp. 1550–1560, 2017.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

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

[5] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature medicine*, vol. 25, no. 1, pp. 44–56, 2019.

[6] J. Jiménez-Luna, F. Grisoni, and G. Schneider, "Drug discovery with explainable artificial intelligence," *Nature Machine Intelligence*, vol. 2, no. 10, pp. 573–584, 2020.

[7] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[8] L. Xie, Q. Yu, Y. Zhou, Y. Wang, E. K. Fishman, and A. L. Yuille, "Recurrent saliency transformation network for tiny target segmentation in abdominal ct scans," *IEEE transactions on medical imaging*, vol. 39, no. 2, pp. 514–525, 2019.

[9] S. L. Walsh, N. Sverzellati, A. Devaraj, G. J. Keir, A. U. Wells, and D. M. Hansell, "Connective tissue disease related fibrotic lung disease: high resolution computed tomographic and pulmonary function indices as prognostic determinants," *Thorax*, vol. 69, no. 3, pp. 216–222, 2014.

[10] S. L. Walsh, N. Sverzellati, A. Devaraj, A. U. Wells, and D. M. Hansell, "Chronic hypersensitivity pneumonitis: high resolution computed tomography patterns and pulmonary function indices as prognostic determinants," *European Radiology*, vol. 22, no. 8, pp. 1672–1679, 2012.

[11] A. W. G. of the International, I. N. Network, I. S. Roberts, H. T. Cook, S. Troyanov, C. E. Alpers, A. Amore, J. Barratt, F. Berthoux, S. Bonsib, *et al.*, "The oxford classification of iga nephropathy: pathology definitions, correlations, and reproducibility," *Kidney international*, vol. 76, no. 5, pp. 546–556, 2009.

[12] Y. Nan, F. Li, P. Tang, G. Zhang, C. Zeng, G. Xie, Z. Liu, and G. Yang, "Automatic fine-grained glomerular lesion recognition in kidney pathology," *Pattern Recognition*, vol. 127, p. 108648, 2022.

[13] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, and A. J. Aljaaf, "A systematic review on supervised and unsupervised machine learning algorithms for data science," *Supervised and unsupervised learning for data science*, pp. 3–21, 2020.

[14] Y. Nan, J. Del Ser, S. Walsh, C. Schönlieb, M. Roberts, I. Selby, K. Howard, J. Owen, J. Neville, J. Guiot, *et al.*, "Data harmonisation for information fusion in digital healthcare: A state-of-the-art systematic review, meta-analysis and future research directions," *Information Fusion*, vol. 82, pp. 99–122, 2022.

[15] R. Da-Ano, D. Visvikis, and M. Hatt, "Harmonization strategies for multicenter radiomics investigations," *Physics in Medicine & Biology*, vol. 65, no. 24, p. 24TR02, 2020.

[16] J. Enguehard, P. O'Halloran, and A. Gholipour, "Semi-supervised learning with deep embedded clustering for image classification and segmentation," *Ieee Access*, vol. 7, pp. 11093–11104, 2019.

[17] M. Jabi, M. Pedersoli, A. Mitiche, and I. B. Ayed, "Deep clustering: On the link between discriminative models and k-means," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 6, pp. 1887–1896, 2019.

[18] W. Le, S. Liang, Y. Hu, K. Deng, H. Bao, C. Zeng, and Z. Liu, "Long-term renal survival and related risk factors in patients with iga nephropathy: results from a cohort of 1155 cases in a chinese adult population," *Nephrology Dialysis Transplantation*, vol. 27, no. 4, pp. 1479–1485, 2012.

[19] M. Haas, J. C. Verhave, Z.-H. Liu, C. E. Alpers, J. Barratt, J. U. Becker, D. Cattran, H. T. Cook, R. Coppo, J. Feehally, *et al.*, "A multicenter study of the predictive value of crescents in iga nephropathy," *Journal of the American Society of Nephrology: JASN*, vol. 28, no. 2, p. 691, 2017.

[20] J. D. Bukowy, A. Dayton, D. Cloutier, A. D. Manis, A. Staruschenko, J. H. Lombard, L. C. S. Woods, D. A. Beard, and A. W. Cowley Jr, "Region-based convolutional neural nets for localization of glomeruli in trichrome-stained whole kidney sections," *Journal of the American Society of Nephrology: JASN*, vol. 29, no. 8, p. 2081, 2018.

[21] B. Ginley, B. Lutnick, K.-Y. Jen, A. B. Fogo, S. Jain, A. Rosenberg, V. Walavalkar, G. Wilding, J. E. Tomaszewski, R. Yacoub, *et al.*, "Computational segmentation and classification of diabetic glomerulosclerosis," *Journal of the American Society of Nephrology: JASN*, vol. 30, no. 10, p. 1953, 2019.

[22] O. Simon, R. Yacoub, S. Jain, J. E. Tomaszewski, and P. Sarder, "Multi-radial lbp features as a tool for rapid glomerular detection and assessment in whole slide histopathology images," *Scientific reports*, vol. 8, no. 1, p. 2032, 2018.

[23] M. Hermsen, T. de Bel, M. Den Boer, E. J. Steenbergen, J. Kers, S. Florquin, J. J. Roelofs, M. D. Stegall, M. P. Alexander, B. H. Smith, *et al.*, "Deep learning–based histopathologic

assessment of kidney tissue," *Journal of the American Society of Nephrology: JASN*, vol. 30, no. 10, p. 1968, 2019.

[24] T. de Bel, M. Hermsen, B. Smeets, L. Hilbrands, J. van der Laak, and G. Litjens, "Automatic segmentation of histopathological slides of renal tissue using deep learning," in *Medical Imaging 2018: Digital Pathology*, vol. 10581, pp. 285–290, SPIE, 2018.

[25] M. Gadermayr, A.-K. Dombrowski, B. M. Klinkhammer, P. Boor, and D. Merhof, "Cnn cascades for segmenting sparse objects in gigapixel whole slide images," *Computerized Medical Imaging and Graphics*, vol. 71, pp. 40–48, 2019.

[26] J. Gallego, A. Pedraza, S. Lopez, G. Steiner, L. Gonzalez, A. Laurinavicius, and G. Bueno, "Glomerulus classification and detection based on convolutional neural networks," *Journal of Imaging*, vol. 4, no. 1, p. 20, 2018.

[27] S. Sheehan, S. Mawe, R. E. Cianciolo, R. Korstanje, and J. M. Mahoney, "Detection and classification of novel renal histologic phenotypes using deep neural networks," *The American Journal of Pathology*, vol. 189, no. 9, pp. 1786–1796, 2019.

[28] J. N. Marsh, M. K. Matlock, S. Kudose, T.-C. Liu, T. S. Stappenbeck, J. P. Gaut, and S. J. Swamidass, "Deep learning global glomerulosclerosis in transplant kidney frozen sections," *IEEE transactions on medical imaging*, vol. 37, no. 12, pp. 2718–2728, 2018.

[29] G. O. Barros, B. Navarro, A. Duarte, and W. L. Dos-Santos, "Pathospotter-k: A computational tool for the automatic identification of glomerular lesions in histological images of kidneys," *Scientific reports*, vol. 7, no. 1, p. 46769, 2017.

[30] G. Bueno, M. M. Fernandez-Carrobles, L. Gonzalez-Lopez, and O. Deniz, "Glomerulosclerosis identification in whole slide images using semantic segmentation," *Computer methods and programs in biomedicine*, vol. 184, p. 105273, 2020.

[31] G.-S. Xie, X.-Y. Zhang, W. Yang, M. Xu, S. Yan, and C.-L. Liu, "Lg-cnn: From local parts to global discrimination for fine-grained recognition," *Pattern Recognition*, vol. 71, pp. 118–131, 2017.

[32] L. Qi, X. Lu, and X. Li, "Exploiting spatial relation for fine-grained image classification," *Pattern Recognition*, vol. 91, pp. 47–55, 2019.

[33] X.-S. Wei, C.-W. Xie, J. Wu, and C. Shen, "Mask-cnn: Localizing parts and selecting descriptors for fine-grained bird species categorization," *Pattern Recognition*, vol. 76, pp. 704–714, 2018.

[34] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 11, pp. 3212–3232, 2019.

[35] L. Chen, S. Wang, K.-M. Lam, H. Zhou, M. Jian, and J. Dong, "Cascaded one-vs-rest detection network for fine-grained recognition without part annotations," *Multimedia Tools and Applications*, vol. 78, pp. 4381–4395, 2019.

[36] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

[37] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9627–9636, 2019.

[38] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

[39] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[40] K. Sirinukunwattana, D. R. Snead, and N. M. Rajpoot, "A stochastic polygons model for glandular structures in colon histology images," *IEEE transactions on medical imaging*, vol. 34, no. 11, pp. 2366–2378, 2015.

[41] Y. Wu and K. He, "Group normalization," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

[42] S. Asgari Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh, "Deep semantic segmentation of natural and medical images: a review," *Artificial Intelligence Review*, vol. 54, pp. 137–178, 2021.

[43] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pp. 240–248, Springer, 2017.

[44] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

[45] W. Zhu, Y. Huang, L. Zeng, X. Chen, Y. Liu, Z. Qian, N. Du, W. Fan, and X. Xie, "Anatomynet: deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy," *Medical physics*, vol. 46, no. 2, pp. 576–589, 2019.

[46] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[47] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, 2017.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

[49] M. A. Nielsen, *Neural networks and deep learning*, vol. 25. Determination press San Francisco, CA, USA, 2015.

[50] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, "Mask scoring r-cnn," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6409–6418, 2019.

[51] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6154–6162, 2018.

[52] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[53] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

[54] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[55] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, pp. 6105–6114, PMLR, 2019.

[56] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[57] Y. Chen, Y. Bai, W. Zhang, and T. Mei, "Destruction and construction learning for fine-grained image recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5157–5166, 2019.

[58] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

[59] C. Zeng, Y. Nan, F. Xu, Q. Lei, F. Li, T. Chen, S. Liang, X. Hou, B. Lv, D. Liang, *et al.*, "Identification of glomerular lesions and intrinsic glomerular cell types in kidney diseases via deep learning," *The Journal of pathology*, vol. 252, no. 1, pp. 53–64, 2020.

[60] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021.

[61] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural computation*, vol. 29, no. 9, pp. 2352–2449, 2017.

[62] A. J. Edey, A. A. Devaraj, R. P. Barker, A. G. Nicholson, A. U. Wells, and D. M. Hansell, "Fibrotic idiopathic interstitial pneumonias: Hrct findings that predict mortality," *European radiology*, vol. 21, no. 8, pp. 1586–1593, 2011.

[63] S. L. Walsh, L. Calandriello, N. Sverzellati, A. U. Wells, and D. M. Hansell, "Interobserver agreement for the ats/ers/jrs/alat criteria for a uip pattern on ct," *Thorax*, vol. 71, no. 1, pp. 45–51, 2016.

[64] J. Jacob, B. J. Bartholmai, S. Rajagopalan, M. Kokosi, A. Nair, R. Karwoski, S. L. Walsh, A. U. Wells, and D. M. Hansell, "Mortality prediction in idiopathic pulmonary fibrosis: evaluation of computer-based ct analysis with conventional severity measures," *European Respiratory Journal*, vol. 49, no. 1, 2017.

[65] S. M. Humphries, K. Yagihashi, J. Huckleberry, B.-H. Rho, J. D. Schroeder, M. Strand, M. I. Schwarz, K. R. Flaherty, E. A. Kazerooni, E. J. van Beek, *et al.*, "Idiopathic pulmonary fibrosis: data-driven textural analysis of extent of fibrosis at baseline and 15-month follow-up," *Radiology*, vol. 285, no. 1, pp. 270–278, 2017.

[66] D. M. Hansell, J. G. Goldin, T. E. King Jr, D. A. Lynch, L. Richeldi, and A. U. Wells, "Ct staging and monitoring of fibrotic interstitial lung diseases in clinical practice and treatment trials: a position paper from the fleischner society," *The Lancet Respiratory Medicine*, vol. 3, no. 6, pp. 483–496, 2015.

[67] J. Tschirren, E. A. Hoffman, G. McLennan, and M. Sonka, "Intrathoracic airway trees: segmentation and airway morphology analysis from low-dose ct scans," *IEEE Transactions on Medical Imaging*, vol. 24, no. 12, pp. 1529–1539, 2005.

[68] A. P. Kiraly, W. E. Higgins, G. McLennan, E. A. Hoffman, and J. M. Reinhardt, "Three-dimensional human airway segmentation methods for clinical virtual bronchoscopy," *Academic Radiology*, vol. 9, no. 10, pp. 1153–1168, 2002.

[69] M. W. Graham, J. D. Gibbs, D. C. Cornish, and W. E. Higgins, "Robust 3-d airway tree segmentation for image-guided peripheral bronchoscopy," *IEEE Transactions on Medical Imaging*, vol. 29, no. 4, pp. 982–997, 2010.

[70] Z. Xu, U. Bagci, B. Foster, A. Mansoor, J. K. Udupa, and D. J. Mollura, "A hybrid method for airway segmentation and automated measurement of bronchial wall thickness on ct," *Medical Image Analysis*, vol. 24, no. 1, pp. 1–17, 2015.

[71] D. Aykac, E. A. Hoffman, G. McLennan, and J. M. Reinhardt, "Segmentation and analysis of the human airway tree from three-dimensional x-ray ct images," *IEEE Transactions on Medical Imaging*, vol. 22, no. 8, pp. 940–950, 2003.

[72] P. Lo, B. Van Ginneken, J. M. Reinhardt, T. Yavarna, P. A. De Jong, B. Irving, C. Fetita, M. Ortner, R. Pinho, J. Sijbers, *et al.*, "Extraction of airways from ct (exact'09)," *IEEE Transactions on Medical Imaging*, vol. 31, no. 11, pp. 2093–2107, 2012.

[73] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *International Conference on 3D Vision (3DV)*, pp. 565–571, 2016.

[74] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021.

[75] J.-P. Charbonnier, E. M. Van Rikxoort, A. A. Setio, C. M. Schaefer-Prokop, B. van Ginneken, and F. Ciompi, "Improving airway segmentation in computed tomography using

leak detection with convolutional networks," *Medical Image Analysis*, vol. 36, pp. 52–60, 2017.

[76] J. Yun, J. Park, D. Yu, J. Yi, M. Lee, H. J. Park, J.-G. Lee, J. B. Seo, and N. Kim, "Improvement of fully automated airway segmentation on volumetric computed tomographic images using a 2.5 dimensional convolutional neural net," *Medical Image Analysis*, vol. 51, pp. 13–20, 2019.

[77] A. Garcia-Uceda Juarez, H. A. Tiddens, and M. d. Bruijne, "Automatic airway segmentation in chest ct using convolutional neural networks," in *Image Analysis for Moving Organ, Breast, and Thoracic Images*, pp. 238–250, Springer, 2018.

[78] Y. Qin, M. Chen, H. Zheng, Y. Gu, M. Shen, J. Yang, X. Huang, Y.-M. Zhu, and G.-Z. Yang, "Airwaynet: a voxel-connectivity aware approach for accurate airway segmentation using convolutional neural networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 212–220, Springer, 2019.

[79] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 424–432, Springer, 2016.

[80] H. Li, Z. Tang, Y. Nan, and G. Yang, "Human treelike tubular structure segmentation: A comprehensive review and future perspectives," *Computers in Biology and Medicine*, p. 106241, 2022.

[81] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, *et al.*, "Attention u-net: Learning where to look for the pancreas," in *International Conference on Medical Imaging with Deep Learning*, p. 15, 2018.

[82] Y. Deng, Z. Ren, Y. Kong, F. Bao, and Q. Dai, "A hierarchical fused fuzzy deep neural network for data classification," *IEEE Transactions on Fuzzy Systems*, vol. 25, no. 4, pp. 1006–1012, 2016.

[83] C. Guan, S. Wang, and A. W.-C. Liew, "Lip image segmentation based on a fuzzy convolutional neural network," *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 7, pp. 1242–1251, 2019.

[84] T. Shen, J. Wang, C. Gou, and F.-Y. Wang, "Hierarchical fused model with deep learning and type-2 fuzzy learning for breast cancer diagnosis," *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 12, pp. 3204–3218, 2020.

[85] A. Garcia-Uceda, R. Selvan, Z. Saghir, H. A. Tiddens, and M. de Bruijne, "Automatic airway segmentation from computed tomography using robust and efficient 3-d convolutional neural networks," *Scientific Reports*, vol. 11, no. 1, pp. 1–15, 2021.

[86] Y. Qin, Y. Gu, H. Zheng, M. Chen, J. Yang, and Y.-M. Zhu, "Airwaynet-se: A simple-yet-effective approach to improve airway segmentation using context scale fusion," in *International Symposium on Biomedical Imaging (ISBI)*, pp. 809–813, IEEE, 2020.

[87] H. Zheng, Y. Qin, Y. Gu, F. Xie, J. Yang, J. Sun, and G.-Z. Yang, "Alleviating class-wise gradient imbalance for pulmonary airway segmentation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 9, pp. 2452–2462, 2021.

[88] P. J. Reynisson, M. Scali, E. Smistad, E. F. Hofstad, H. O. Leira, F. Lindseth, T. A. Nagelhus Hernes, T. Amundsen, H. Sorger, and T. Langø, "Airway segmentation and centerline extraction from thoracic ct–comparison of a new method to state of the art commercialized methods," *PloS One*, vol. 10, no. 12, p. e0144282, 2015.

[89] Q. Meng, H. R. Roth, T. Kitasaka, M. Oda, J. Ueno, and K. Mori, "Tracking and segmentation of the airways in chest ct using a fully convolutional network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 198–207, Springer, 2017.

[90] S. A. Nadeem, E. A. Hoffman, J. C. Sieren, A. P. Comellas, S. P. Bhatt, I. Z. Barjaktarevic, F. Abtin, and P. K. Saha, "A ct-based automated algorithm for airway segmentation using freeze-and-grow propagation and deep learning," *IEEE Transactions on Medical Imaging*, vol. 40, no. 1, pp. 405–418, 2020.

[91] C. Wang, Y. Hayashi, M. Oda, H. Itoh, T. Kitasaka, A. F. Frangi, and K. Mori, "Tubular structure segmentation using spatial fully connected network with radial distance loss for 3d medical images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 348–356, Springer, 2019.

[92] T. Lei, P. Liu, X. Jia, X. Zhang, H. Meng, and A. K. Nandi, "Automatic fuzzy clustering framework for image segmentation," *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 9, pp. 2078–2092, 2019.

[93] T. Lei, X. Jia, Y. Zhang, S. Liu, H. Meng, and A. K. Nandi, "Superpixel-based fast fuzzy c-means clustering for color image segmentation," *IEEE Transactions on Fuzzy Systems*, vol. 27, no. 9, pp. 1753–1766, 2018.

[94] S. Zhou, Q. Chen, and X. Wang, "Fuzzy deep belief networks for semi-supervised sentiment classification," *Neurocomputing*, vol. 131, pp. 312–322, 2014.

[95] S. R. Price, S. R. Price, and D. T. Anderson, "Introducing fuzzy layers for deep learning," in *International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–6, IEEE, 2019.

[96] K. Huang, Y. Zhang, H. Cheng, P. Xing, and B. Zhang, "Semantic segmentation of breast ultrasound image with fuzzy deep learning network and breast anatomy constraints," *Neurocomputing*, vol. 450, pp. 319–335, 2021.

[97] W. Ding, M. Abdel-Basset, H. Hawash, and W. Pedrycz, "Multimodal infant brain segmentation by fuzzy-informed deep learning," *IEEE Transactions on Fuzzy Systems*, vol. 30, no. 4, pp. 1088–1101, 2022.

[98] W. Xie, C. Jacobs, J.-P. Charbonnier, and B. Van Ginneken, "Relational modeling for robust and efficient pulmonary lobe segmentation in ct scans," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2664–2675, 2020.

[99] T.-C. Lee, R. L. Kashyap, and C.-N. Chu, "Building skeleton models via 3-d medial surface axis thinning algorithms," *CVGIP: Graphical Models and Image Processing*, vol. 56, no. 6, pp. 462–478, 1994.

[100] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, *et al.*, "The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans," *Medical Physics*, vol. 38, no. 2, pp. 915–931, 2011.

[101] A. Wang, T. C. C. Tam, H. M. Poon, K.-C. Yu, and W.-N. Lee, "Naviairway: a bronchiole-sensitive deep learning-based airway segmentation pipeline for planning of navigation bronchoscopy," *arXiv preprint arXiv:2203.04294*, 2022.

[102] A. Garcia-Uceda, R. Selvan, Z. Saghir, H. A. Tiddens, and M. de Bruijne, "Automatic airway segmentation from computed tomography using robust and efficient 3-d convolutional neural networks," *Scientific Reports*, vol. 11, no. 1, pp. 1–15, 2021.

[103] L. Xia, H. Zhang, Y. Wu, R. Song, Y. Ma, L. Mou, J. Liu, Y. Xie, M. Ma, and Y. Zhao, "3d vessel-like structure segmentation in medical images by an edge-reinforced network," *Medical Image Analysis*, vol. 82, p. 102581, 2022.

[104] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, and J. Liu, "Ce-net: Context encoder network for 2d medical image segmentation," *IEEE transactions on medical imaging*, vol. 38, no. 10, pp. 2281–2292, 2019.

[105] Y. Ma, H. Hao, J. Xie, H. Fu, J. Zhang, J. Yang, Z. Wang, J. Liu, Y. Zheng, and Y. Zhao, "Rose: a retinal oct-angiography vessel segmentation dataset and new model," *IEEE transactions on medical imaging*, vol. 40, no. 3, pp. 928–939, 2020.

[106] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, eds.), (Cham), pp. 234–241, Springer International Publishing, 2015.

[107] J.-P. Charbonnier, E. M. van Rikxoort, A. A. Setio, C. M. Schaefer-Prokop, B. van Ginneken, and F. Ciompi, "Improving airway segmentation in computed tomography using leak detection with convolutional networks," *Medical Image Analysis*, vol. 36, pp. 52–60, 2017.

[108] D. Jin, Z. Xu, A. P. Harrison, K. George, and D. J. Mollura, "3d convolutional neural networks with graph refinement for airway segmentation using incomplete data labels," in *Machine Learning in Medical Imaging* (Q. Wang, Y. Shi, H.-I. Suk, and K. Suzuki, eds.), (Cham), pp. 141–149, Springer International Publishing, 2017.

[109] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[110] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li, "Transbts: Multimodal brain tumor segmentation using transformer," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I*, (Berlin, Heidelberg), p. 109–119, Springer-Verlag, 2021.

[111] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1748–1758, IEEE, 1 2022.

[112] S. Li, X. Sui, X. Luo, X. Xu, Y. Liu, and R. Goh, "Medical image segmentation using squeeze-and-expansion transformers," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 807–815, International Joint Conferences on Artificial Intelligence Organization, 8 2021.

[113] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10002, IEEE, 10 2021.

[114] H. Zhang, M. Shen, P. L. Shah, and G.-Z. Yang, "Pathological airway segmentation with cascaded neural networks for bronchoscopic navigation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9974–9980, IEEE, 5 2020.

[115] H. Zhao, X. Qiu, W. Lu, H. Huang, and X. Jin, "High-quality retinal vessel segmentation using generative adversarial network with a large receptive field," *International Journal of Imaging Systems and Technology*, vol. 30, pp. 828–842, 9 2020.

[116] X. Guo, C. Chen, Y. Lu, K. Meng, H. Chen, K. Zhou, Z. Wang, and R. Xiao, "Retinal vessel segmentation combined with generative adversarial networks and dense u-net," *IEEE Access*, vol. 8, pp. 194551–194560, 2020.

[117] K.-B. Park, S. H. Choi, and J. Y. Lee, "M-gan: Retinal blood vessel segmentation by balancing losses through stacked deep fully convolutional networks," *IEEE Access*, vol. 8, pp. 146308–146322, 2020.

[118] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-Supervised Nets," in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics* (G. Lebanon and S. V. N. Vishwanathan, eds.), vol. 38 of *Proceedings of Machine Learning Research*, (San Diego, California, USA), pp. 562–570, PMLR, 09–12 May 2015.

[119] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967–5976, 2017.

[120] S. Shit, J. C. Paetzold, A. Sekuboyina, I. Ezhov, A. Unger, A. Zhylka, J. P. W. Pluim, U. Bauer, and B. H. Menze, "cldice - a novel topology-preserving loss function for tubular structure segmentation," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16555–16564, 2021.

[121] Y. Nan, J. Del Ser, Z. Tang, P. Tang, X. Xing, Y. Fang, F. Herrera, W. Pedrycz, S. Walsh, and G. Yang, "Fuzzy attention neural network to tackle discontinuity in airway segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[122] W. Kuo, A. Perez-Rovira, H. Tiddens, and M. de Bruijne, "Airway tapering: an objective image biomarker for bronchiectasis," *European Radiology*, vol. 30, pp. 2703–2711, 5 2020.

[123] P. U. Adiga and B. Chaudhuri, "An efficient method based on watershed and rule-based merging for segmentation of 3-d histo-pathological images," *Pattern recognition*, vol. 34, no. 7, pp. 1449–1458, 2001.

[124] S. Wienert, D. Heim, K. Saeger, A. Stenzinger, M. Beil, P. Hufnagl, M. Dietel, C. Denkert, and F. Klauschen, "Detection and segmentation of cell nuclei in virtual microscopy images: a minimum-model approach," *Scientific reports*, vol. 2, no. 1, p. 503, 2012.

[125] I. Nogues, L. Lu, X. Wang, H. Roth, G. Bertasius, N. Lay, J. Shi, Y. Tsehay, and R. M. Summers, "Automatic lymph node cluster segmentation using holistically-nested neural networks and structured optimization in ct images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 388–397, Springer, 2016.

[126] S. Ragothaman, S. Narasimhan, M. G. Basavaraj, and R. Dewar, "Unsupervised segmentation of cervical cell images using gaussian mixture model," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 70–75, 2016.

[127] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm," *IEEE transactions on medical imaging*, vol. 20, no. 1, pp. 45–57, 2001.

[128] Z. Guo, H. Liu, H. Ni, X. Wang, M. Su, W. Guo, K. Wang, T. Jiang, and Y. Qian, "A fast and refined cancer regions segmentation framework in whole-slide breast pathological images," *Scientific reports*, vol. 9, no. 1, p. 882, 2019.

[129] S. Graham, Q. D. Vu, S. E. A. Raza, A. Azam, Y. W. Tsang, J. T. Kwak, and N. Rajpoot, "Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images," *Medical image analysis*, vol. 58, p. 101563, 2019.

[130] S. Graham, H. Chen, J. Gamper, Q. Dou, P.-A. Heng, D. Snead, Y. W. Tsang, and N. Rajpoot, "Mild-net: Minimal information loss dilated network for gland instance segmentation in colon histology images," *Medical image analysis*, vol. 52, pp. 199–211, 2019.

[131] A. Mahbod, G. Schaefer, B. Bancher, C. Löw, G. Dorffner, R. Ecker, and I. Ellinger, "Cryonuseg: A dataset for nuclei instance segmentation of cryosectioned h&e-stained histological images," *Computers in biology and medicine*, vol. 132, p. 104349, 2021.

[132] D. Liu, D. Zhang, Y. Song, F. Zhang, L. O'Donnell, H. Huang, M. Chen, and W. Cai, "Unsupervised instance segmentation in microscopy images via panoptic domain adaptation and task re-weighting," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4243–4252, 2020.

[133] Q. Liang, Y. Nan, G. Coppola, K. Zou, W. Sun, D. Zhang, Y. Wang, and G. Yu, "Weakly supervised biomedical image segmentation by reiterative learning," *IEEE Journal of biomedical and health informatics*, vol. 23, no. 3, pp. 1205–1214, 2018.

[134] S. Hu, Y. Gao, Z. Niu, Y. Jiang, L. Li, X. Xiao, M. Wang, E. F. Fang, W. Menpes-Smith, J. Xia, *et al.*, "Weakly supervised deep learning for covid-19 infection detection and classification from ct images," *IEEE Access*, vol. 8, pp. 118869–118883, 2020.

[135] H. E. Atlason, A. Love, S. Sigurdsson, V. Gudnason, and L. M. Ellingsen, "Unsupervised brain lesion segmentation from mri using a convolutional autoencoder," in *Medical Imaging 2019: Image Processing*, vol. 10949, pp. 372–378, SPIE, 2019.

[136] A. Kanezaki, "Unsupervised image segmentation by backpropagation," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 1543–1547, IEEE, 2018.

[137] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[138] Y. Shen and H. Zhou, "Double dip: Re-evaluating security of logic encryption algorithms," in *Proceedings of the on Great Lakes Symposium on VLSI 2017*, pp. 179–184, 2017.

[139] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[140] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 11, pp. 1222–1239, 2001.

[141] D. A. Clausi, "K-means iterative fisher (kif) unsupervised clustering algorithm applied to image texture segmentation," *Pattern Recognition*, vol. 35, no. 9, pp. 1959–1972, 2002.

[142] Z. Ji, Y. Xia, Q. Sun, Q. Chen, D. Xia, and D. D. Feng, "Fuzzy local gaussian mixture model for brain mr image segmentation," *IEEE transactions on Information Technology in Biomedicine*, vol. 16, no. 3, pp. 339–347, 2012.

[143] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[144] Z. F. Knops, J. A. Maintz, M. A. Viergever, and J. P. Pluim, "Normalized mutual information based registration using k-means clustering and shading correction," *Medical image analysis*, vol. 10, no. 3, pp. 432–439, 2006.

[145] L.-H. Juang and M.-N. Wu, "Mri brain lesion image detection based on color-converted k-means clustering segmentation," *Measurement*, vol. 43, no. 7, pp. 941–949, 2010.

[146] Z. Fan, J. Lu, C. Wei, H. Huang, X. Cai, and X. Chen, "A hierarchical image matting model for blood vessel segmentation in fundus images," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2367–2377, 2018.

[147] A. B. Tosun, M. Kandemir, C. Sokmensuer, and C. Gunduz-Demir, "Object-oriented texture analysis for the unsupervised segmentation of biopsy images for cancer detection," *Pattern Recognition*, vol. 42, no. 6, pp. 1104–1112, 2009.

[148] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 132–149, 2018.

[149] S. Huang, Z. Kang, Z. Xu, and Q. Liu, "Robust deep k-means: An effective and simple method for data clustering," *Pattern Recognition*, vol. 117, p. 107996, 2021.

[150] L. Manduchi, K. Chin-Cheong, H. Michel, S. Wellmann, and J. Vogt, "Deep conditional gaussian mixture model for constrained clustering," *Advances in Neural Information Processing Systems*, vol. 34, pp. 11303–11314, 2021.

[151] W. Kim, A. Kanezaki, and M. Tanaka, "Unsupervised learning of image segmentation based on differentiable feature clustering," *IEEE Transactions on Image Processing*, vol. 29, pp. 8055–8068, 2020.

[152] A. Tsai, W. Wells, C. Tempany, E. Grimson, and A. Willsky, "Mutual information in coupled multi-shape model for medical image segmentation," *Medical image analysis*, vol. 8, no. 4, pp. 429–445, 2004.

[153] X. Ji, J. F. Henriques, and A. Vedaldi, "Invariant information clustering for unsupervised image classification and segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9865–9874, 2019.

[154] X. Xia and B. Kulis, "W-net: A deep model for fully unsupervised image segmentation," *arXiv preprint arXiv:1711.08506*, 2017.

[155] M. Chen, T. Artières, and L. Denoyer, "Unsupervised object segmentation by redrawing," *Advances in neural information processing systems*, vol. 32, 2019.

[156] Y. Gandelsman, A. Shocher, and M. Irani, ""double-dip": unsupervised image decomposition via coupled deep-image-priors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11026–11035, 2019.

[157] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen, "Deep autoencoding gaussian mixture model for unsupervised anomaly detection," in *International conference on learning representations*, 2018.

[158] A. Van den Oord and B. Schrauwen, "Factoring variations in natural images with deep gaussian mixture models," *Advances in neural information processing systems*, vol. 27, 2014.

[159] F. G. Zanjani, S. Zinger, B. E. Bejnordi, J. A. van der Laak, *et al.*, "Histopathology stain-color normalization using deep generative models," in *1st Conference on Medical Imaging with Deep Learning (MIDL 2018)*, pp. 1–11, 2018.

[160] S. Sun, Z. Cao, H. Zhu, and J. Zhao, "A survey of optimization methods from a machine learning perspective," *IEEE transactions on cybernetics*, vol. 50, no. 8, pp. 3668–3681, 2019.

[161] K. Mardia, H. Southworth, and C. Taylor, "On bias in maximum likelihood estimators," *Journal of statistical planning and inference*, vol. 76, no. 1-2, pp. 31–39, 1999.

[162] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

[163] W. M. Association *et al.*, "World medical association declaration of helsinki. ethical principles for medical research involving human subjects.," *Bulletin of the World Health Organization*, vol. 79, no. 4, p. 373, 2001.

[164] J. Wang and J. Jiang, "Unsupervised deep clustering via adaptive gmm modeling and optimization," *Neurocomputing*, vol. 433, pp. 199–211, 2021.

[165] L. Zhou and W. Wei, "Dic: deep image clustering for unsupervised image segmentation," *Ieee Access*, vol. 8, pp. 34481–34491, 2020.

[166] X. Li, Y. Wang, Q. Tang, Z. Fan, and J. Yu, "Dual u-net for the segmentation of overlapping glioma nuclei," *Ieee Access*, vol. 7, pp. 84040–84052, 2019.

[167] K. Tian, J. Zhang, H. Shen, K. Yan, P. Dong, J. Yao, S. Che, P. Luo, and X. Han, "Weakly-supervised nucleus segmentation based on point annotations: A coarse-to-fine self-stimulated learning strategy," in *Medical Image Computing and Computer Assisted*

Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23, pp. 299–308, Springer, 2020.

[168] H. Qu, P. Wu, Q. Huang, J. Yi, Z. Yan, K. Li, G. M. Riedlinger, S. De, S. Zhang, and D. N. Metaxas, "Weakly supervised deep nuclei segmentation using partial points annotation in histopathology images," *IEEE transactions on medical imaging*, vol. 39, no. 11, pp. 3655–3666, 2020.

[169] D. Liu, D. Zhang, Y. Song, H. Huang, and W. Cai, "Panoptic feature fusion net: a novel instance segmentation paradigm for biomedical and biological images," *IEEE Transactions on Image Processing*, vol. 30, pp. 2045–2059, 2021.

[170] D. Liu, D. Zhang, Y. Song, F. Zhang, L. O'Donnell, H. Huang, M. Chen, and W. Cai, "Pdam: A panoptic-level feature alignment framework for unsupervised domain adaptive instance segmentation in microscopy images," *IEEE Transactions on Medical Imaging*, vol. 40, no. 1, pp. 154–165, 2020.

[171] D. A. Lynch, N. Sverzellati, W. D. Travis, K. K. Brown, T. V. Colby, J. R. Galvin, J. G. Goldin, D. M. Hansell, Y. Inoue, T. Johkoh, *et al.*, "Diagnostic criteria for idiopathic pulmonary fibrosis: a fleischner society white paper," *The Lancet respiratory medicine*, vol. 6, no. 2, pp. 138–153, 2018.

[172] Q. Han, Q. Luo, J.-X. Xie, L.-L. Wu, L.-Y. Liao, X.-X. Zhang, and R.-C. Chen, "Diagnostic yield and postoperative mortality associated with surgical lung biopsy for evaluation of interstitial lung diseases: a systematic review and meta-analysis," *The Journal of thoracic and cardiovascular surgery*, vol. 149, no. 5, pp. 1394–1401, 2015.

[173] J. Utz, J. Ryu, W. Douglas, T. Hartman, H. Tazelaar, J. Myers, M. Allen, and D. Schroeder, "High short-term mortality following lung biopsy for usual interstitial pneumonia," *European Respiratory Journal*, vol. 17, no. 2, pp. 175–179, 2001.

[174] M. Zhang, Y. Wu, H. Zhang, Y. Qin, H. Zheng, W. Tang, C. Arnold, C. Pei, P. Yu, Y. Nan, *et al.*, "Multi-site, multi-domain airway tree modeling," *Medical Image Analysis*, vol. 90, p. 102957, 2023.

[175] A. Pavao, I. Guyon, A.-C. Letournel, D.-T. Tran, X. Baro, H. J. Escalante, S. Escalera, T. Thomas, and Z. Xu, "Codalab competitions," *Journal of Machine Learning Research*, vol. 24, pp. 1–6, 2023.

[176] M. Sonka, W. Park, and E. A. Hoffman, "Rule-based detection of intrathoracic airway trees," *IEEE transactions on medical imaging*, vol. 15, no. 3, pp. 314–326, 1996.

[177] K. Mori, J.-i. Hasegawa, Y. Suenaga, and J.-i. Toriwaki, "Automated anatomical labeling of the bronchial branch and its application to the virtual bronchoscopy system," *IEEE Transactions on medical imaging*, vol. 19, no. 2, pp. 103–114, 2000.

[178] Y. Wu, S. Zhao, S. Qi, J. Feng, H. Pang, R. Chang, L. Bai, M. Li, S. Xia, W. Qian, *et al.*, "Two-stage contextual transformer-based convolutional neural network for airway extraction from ct images," *Artificial Intelligence in Medicine*, vol. 143, p. 102637, 2023.

[179] G. Raghu, M. Remy-Jardin, J. L. Myers, L. Richeldi, C. J. Ryerson, D. J. Lederer, J. Behr, V. Cottin, S. K. Danoff, F. Morell, *et al.*, "Diagnosis of idiopathic pulmonary fibrosis. an official ats/ers/jrs/alat clinical practice guideline," *American journal of respiratory and critical care medicine*, vol. 198, no. 5, pp. e44–e68, 2018.

[180] G. Raghu, H. R. Collard, J. J. Egan, F. J. Martinez, J. Behr, K. K. Brown, T. V. Colby, J.-F. Cordier, K. R. Flaherty, J. A. Lasky, *et al.*, "An official ats/ers/jrs/alat statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management," *American journal of respiratory and critical care medicine*, vol. 183, no. 6, pp. 788–824, 2011.

[181] A. C. Best, A. M. Lynch, C. M. Bozic, D. Miller, G. K. Grunwald, and D. A. Lynch, "Quantitative ct indexes in idiopathic pulmonary fibrosis: relationship with physiologic impairment," *Radiology*, vol. 228, no. 2, pp. 407–414, 2003.

[182] J. Jacob, B. J. Bartholmai, S. Rajagopalan, A. L. Brun, R. Egashira, R. Karwoski, M. Kokosi, A. U. Wells, and D. M. Hansell, "Evaluation of computer-based computer tomography stratification against outcome models in connective tissue disease-related

interstitial lung disease: a patient outcome study," *BMC medicine*, vol. 14, no. 1, pp. 1–13, 2016.

[183] J. G. Nam, Y. Choi, S.-M. Lee, S. H. Yoon, J. M. Goo, and H. Kim, "Prognostic value of deep learning–based fibrosis quantification on chest ct in idiopathic pulmonary fibrosis," *European Radiology*, vol. 33, no. 5, pp. 3144–3155, 2023.

[184] S. M. Humphries, J. J. Swigris, K. K. Brown, M. Strand, Q. Gong, J. S. Sundy, G. Raghu, M. I. Schwarz, K. R. Flaherty, R. Sood, *et al.*, "Quantitative high-resolution computed tomography fibrosis score: performance characteristics in idiopathic pulmonary fibrosis," *European Respiratory Journal*, vol. 52, no. 3, 2018.

[185] B. Irving, P. Taylor, and A. Todd-Pokropek, "3d segmentation of the airway tree using a morphology based method," in *Proceedings of 2nd international workshop on pulmonary image analysis*, pp. 297–07, 2009.

[186] R. Pinho, S. Luyckx, and J. Sijbers, "Robust region growing based intrathoracic airway tree segmentation," in *Proc. of Second International Workshop on Pulmonary Image Analysis*, pp. 261–271, 2009.

[187] C. S. Mendoza, B. Acha, and C. Serrano, "Maximal contrast adaptive region growing for ct airway tree segmentation," in *Proc. of Second International Workshop on Pulmonary Image Analysis*, pp. 285–295, 2009.

[188] S. L. Walsh, S. M. Humphries, A. U. Wells, and K. K. Brown, "Imaging research in fibrotic lung disease; applying deep learning to unsolved problems," *The Lancet Respiratory Medicine*, vol. 8, no. 11, pp. 1144–1153, 2020.

[189] P. A. Yushkevich, J. Piven, H. C. Hazlett, R. G. Smith, S. Ho, J. C. Gee, and G. Gerig, "User-guided 3d active contour segmentation of anatomical structures: significantly improved efficiency and reliability," *Neuroimage*, vol. 31, no. 3, pp. 1116–1128, 2006.

[190] J. Hofmanninger, F. Prayer, J. Pan, S. Röhrich, H. Prosch, and G. Langs, "Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem," *European Radiology Experimental*, vol. 4, no. 1, pp. 1–13, 2020.

[191] W. T. Clarke, O. Mougin, I. D. Driver, C. Rua, A. T. Morgan, M. Asghar, S. Clare, S. Francis, R. G. Wise, C. T. Rodgers, *et al.*, "Multi-site harmonization of 7 tesla mri neuroimaging protocols," *Neuroimage*, vol. 206, p. 116335, 2020.

[192] D. Delbeke, R. E. Coleman, M. J. Guiberteau, M. L. Brown, H. D. Royal, B. A. Siegel, D. W. Townsend, L. L. Berland, J. A. Parker, K. Hubner, *et al.*, "Procedure guideline for tumor imaging with 18f-fdg pet/ct 1.0," *Journal of nuclear Medicine*, vol. 47, no. 5, pp. 885–895, 2006.

[193] J. Simon, D. Li, A. Traboulsee, P. Coyle, D. Arnold, F. Barkhof, J. Frank, R. Grossman, D. Paty, E. Radue, *et al.*, "Standardized mr imaging protocol for multiple sclerosis: Consortium of ms centers consensus guidelines," *American Journal of Neuroradiology*, vol. 27, no. 2, pp. 455–461, 2006.

[194] B.-M. Schmidt, C. J. Colvin, A. Hohlfeld, and N. Leon, "Defining and conceptualising data harmonisation: a scoping review protocol," *Systematic reviews*, vol. 7, pp. 1–6, 2018.

[195] B.-M. Schmidt, C. J. Colvin, A. Hohlfeld, and N. Leon, "Definitions, components and processes of data harmonisation in healthcare: a scoping review," *BMC Medical Informatics and Decision Making*, vol. 20, pp. 1–19, 2020.

[196] M. S. Pinto, R. Paolella, T. Billiet, P. Van Dyck, P.-J. Guns, B. Jeurissen, A. Ribbens, A. J. den Dekker, and J. Sijbers, "Harmonization of brain diffusion mri: Concepts and methods," *Frontiers in Neuroscience*, vol. 14, p. 396, 2020.

[197] S. Gitto, R. Cuocolo, D. Albano, F. Morelli, L. C. Pescatori, C. Messina, M. Imbriaco, and L. M. Sconfienza, "Ct and mri radiomics of bone and soft-tissue sarcomas: a systematic review of reproducibility and validation strategies," *Insights Into Imaging*, vol. 12, no. 1, pp. 1–14, 2021.

[198] S. A. Mali, A. Ibrahim, H. C. Woodruff, V. Andrearczyk, H. Müller, S. Primakov, Z. Salahuddin, A. Chatterjee, and P. Lambin, "Making radiomics more reproducible across scanner and imaging protocol variations: a review of harmonization methods," *Journal of personalized medicine*, vol. 11, no. 9, p. 842, 2021.

[199] T. Chen, M. Philip, K.-A. Lê Cao, and S. Tyagi, "A multi-modal data harmonisation approach for discovery of covid-19 drug targets," *Briefings in bioinformatics*, vol. 22, no. 6, p. bbab185, 2021.

[200] C. M. Tax, F. Grussu, E. Kaden, L. Ning, U. Rudrapatna, C. J. Evans, S. St-Jean, A. Leemans, S. Koppers, D. Merhof, *et al.*, "Cross-scanner and cross-protocol diffusion mri data harmonisation: A benchmark database and evaluation of algorithms," *NeuroImage*, vol. 195, pp. 285–299, 2019.

[201] M. Hotta, R. Minamimoto, Y. Gohda, K. Miwa, K. Otani, T. Kiyomatsu, and H. Yano, "Prognostic value of 18f-fdg pet/ct with texture analysis in patients with rectal cancer treated by surgery," *Annals of Nuclear Medicine*, vol. 35, no. 7, pp. 843–852, 2021.

[202] M. V. Mattoli, M. L. Calcagni, S. Taralli, L. Indovina, B. S. Spottiswoode, and A. Giordano, "How often do we fail to classify the treatment response with [18 f] fdg pet/ct acquired on different scanners? data from clinical oncological practice using an automatic tool for suv harmonization," *Molecular imaging and biology*, vol. 21, pp. 1210–1219, 2019.

[203] H. Mirzaalian, L. Ning, P. Savadjiev, O. Pasternak, S. Bouix, O. Michailovich, G. Grant, C. E. Marx, R. A. Morey, L. A. Flashman, *et al.*, "Inter-site and inter-scanner diffusion mri data harmonization," *NeuroImage*, vol. 135, pp. 311–323, 2016.

[204] T. Zhu, R. Hu, X. Qiu, M. Taylor, Y. Tso, C. Yiannoutsos, B. Navia, S. Mori, S. Ekholm, G. Schifitto, *et al.*, "Quantification of accuracy and precision of multi-center dti measurements: a diffusion phantom and human brain study," *Neuroimage*, vol. 56, no. 3, pp. 1398–1411, 2011.

[205] J. Jovicich, M. Marizzoni, B. Bosch, D. Bartrés-Faz, J. Arnold, J. Benninghoff, J. Wiltfang, L. Roccatagliata, A. Picco, F. Nobili, *et al.*, "Multisite longitudinal reliability of tract-based spatial statistics in diffusion tensor imaging of healthy elderly subjects," *Neuroimage*, vol. 101, pp. 390–403, 2014.

[206] P. Leo, G. Lee, N. N. Shih, R. Elliott, M. D. Feldman, and A. Madabhushi, "Evaluating stability of histomorphometric features across scanner and staining variations: prostate

cancer diagnosis from whole slide images," *Journal of medical imaging*, vol. 3, no. 4, pp. 047502–047502, 2016.

[207] A. Jha, S. Mithun, V. Jaiswar, U. Sherkhane, N. Purandare, K. Prabhash, V. Rangarajan, A. Dekker, L. Wee, and A. Traverso, "Repeatability and reproducibility study of radiomic features on a phantom and human cohort," *Scientific reports*, vol. 11, no. 1, p. 2055, 2021.

[208] N. Emaminejad, M. W. Wahi-Anwar, G. H. J. Kim, W. Hsu, M. Brown, and M. McNitt-Gray, "Reproducibility of lung nodule radiomic features: multivariable and univariable investigations that account for interactions between ct acquisition and reconstruction parameters," *Medical physics*, vol. 48, no. 6, pp. 2906–2919, 2021.

[209] M. Kim, S. C. Jung, J. E. Park, S. Y. Park, H. Lee, and K. M. Choi, "Reproducibility of radiomic features in sense and compressed sense: impact of acceleration factors," *European Radiology*, vol. 31, pp. 6457–6470, 2021.

[210] R. Yamashita, T. Perrin, J. Chakraborty, J. F. Chou, N. Horvat, M. A. Koszalka, A. Midya, M. Gonen, P. Allen, W. R. Jarnagin, *et al.*, "Radiomic feature reproducibility in contrast-enhanced ct of the pancreas is affected by variabilities in scan parameters and manual segmentation," *European radiology*, vol. 30, pp. 195–205, 2020.

[211] S. Fiset, M. L. Welch, J. Weiss, M. Pintilie, J. L. Conway, M. Milosevic, A. Fyles, A. Traverso, D. Jaffray, U. Metser, *et al.*, "Repeatability and reproducibility of mri-based radiomic features in cervical cancer," *Radiotherapy and Oncology*, vol. 135, pp. 107–114, 2019.

[212] E. Saeedi, A. Dezhkam, J. Beigi, S. Rastegar, Z. Yousefi, L. A. Mehdipour, H. Abdollahi, and K. Tanha, "Radiomic feature robustness and reproducibility in quantitative bone radiography: a study on radiologic parameter changes," *Journal of Clinical Densitometry*, vol. 22, no. 2, pp. 203–213, 2019.

[213] M. Meyer, J. Ronald, F. Vernuccio, R. C. Nelson, J. C. Ramirez-Giraldo, J. Solomon, B. N. Patel, E. Samei, and D. Marin, "Reproducibility of ct radiomic features within

the same patient: influence of radiation dose and ct reconstruction settings," *Radiology*, vol. 293, no. 3, pp. 583–591, 2019.

[214] T. Perrin, A. Midya, R. Yamashita, J. Chakraborty, T. Saidon, W. R. Jarnagin, M. Gonen, A. L. Simpson, and R. K. Do, "Short-term reproducibility of radiomic features in liver parenchyma and liver malignancies on contrast-enhanced ct imaging," *Abdominal Radiology*, vol. 43, pp. 3271–3278, 2018.

[215] A. Midya, J. Chakraborty, M. Gönen, R. K. Do, and A. L. Simpson, "Influence of ct acquisition and reconstruction parameters on radiomic feature reproducibility," *Journal of Medical Imaging*, vol. 5, no. 1, pp. 011020–011020, 2018.

[216] B. A. Altazi, G. G. Zhang, D. C. Fernandez, M. E. Montejo, D. Hunt, J. Werner, M. C. Biagioli, and E. G. Moros, "Reproducibility of f18-fdg pet radiomic features for different cervical tumor segmentation methods, gray-level discretization, and reconstruction algorithms," *Journal of applied clinical medical physics*, vol. 18, no. 6, pp. 32–48, 2017.

[217] P. Hu, J. Wang, H. Zhong, Z. Zhou, L. Shen, W. Hu, and Z. Zhang, "Reproducibility with repeat ct in radiomics study for rectal cancer," *Oncotarget*, vol. 7, no. 44, p. 71440, 2016.

[218] J. Choe, S. M. Lee, K.-H. Do, G. Lee, J.-G. Lee, S. M. Lee, and J. B. Seo, "Deep learning–based image conversion of ct reconstruction kernels improves radiomics reproducibility for pulmonary nodules or masses," *Radiology*, vol. 292, no. 2, pp. 365–373, 2019.

[219] A. N. Primak, C. H. McCollough, M. R. Bruesewitz, J. Zhang, and J. G. Fletcher, "Relationship between noise, dose, and pitch in cardiac multi–detector row ct," *Radiographics*, vol. 26, no. 6, pp. 1785–1794, 2006.

[220] D. S. Gierada, A. J. Bierhals, C. K. Choong, S. T. Bartel, J. H. Ritter, N. A. Das, C. Hong, T. K. Pilgram, K. T. Bae, B. R. Whiting, *et al.*, "Effects of ct section thickness and reconstruction kernel on emphysema quantification: relationship to the magnitude of the ct emphysema index," *Academic radiology*, vol. 17, no. 2, pp. 146–156, 2010.

[221] P.-Y. Tung, J. D. Blischak, C. J. Hsiao, D. A. Knowles, J. E. Burnett, J. K. Pritchard, and Y. Gilad, "Batch effects and the effective design of single-cell gene expression studies," *Scientific reports*, vol. 7, no. 1, p. 39921, 2017.

[222] T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija, "Comprehensive integration of single-cell data," *Cell*, vol. 177, no. 7, pp. 1888–1902, 2019.

[223] S. Vijh, M. Saraswat, and S. Kumar, "A new complete color normalization method for h&e stained histopatholgical images," *Applied Intelligence*, pp. 1–14, 2021.

[224] D. E. Chandler and R. W. Roberson, "Bioimaging: current concepts in light and electron microscopy," 2009.

[225] D. Sun, J. Wang, Y. Han, X. Dong, J. Ge, R. Zheng, X. Shi, B. Wang, Z. Li, P. Ren, *et al.*, "Tisch: a comprehensive web resource enabling interactive single-cell transcriptome visualization of tumor microenvironment," *Nucleic acids research*, vol. 49, no. D1, pp. D1420–D1430, 2021.

[226] M. Shafiq-ul Hassan, G. G. Zhang, K. Latifi, G. Ullah, D. C. Hunt, Y. Balagurunathan, M. A. Abdalah, M. B. Schabath, D. G. Goldgof, D. Mackin, *et al.*, "Intrinsic dependencies of ct radiomic features on voxel size and number of gray levels," *Medical physics*, vol. 44, no. 3, pp. 1050–1062, 2017.

[227] D. Mackin, X. Fave, L. Zhang, *et al.*, "Measuring ct scanner variability of radiomics features hhs public access," *Invest Radiol*, vol. 50, no. 11, pp. 757–765, 2015.

[228] G. Mårtensson, D. Ferreira, T. Granberg, L. Cavallin, K. Oppedal, A. Padovani, I. Rektorova, L. Bonanni, M. Pardini, M. G. Kramberger, *et al.*, "The reliability of a deep learning model in clinical out-of-distribution mri data: a multicohort study," *Medical Image Analysis*, vol. 66, p. 101714, 2020.

[229] S. Rathore, S. Bakas, H. Akbari, G. Shukla, M. Rozycki, and C. Davatzikos, "Deriving stable multi-parametric mri radiomic signatures in the presence of inter-scanner variations:

survival prediction of glioblastoma via imaging pattern analysis and machine learning techniques," in *Medical Imaging 2018: Computer-Aided Diagnosis*, vol. 10575, pp. 52–58, SPIE, 2018.

[230] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical bayes methods," *Biostatistics*, vol. 8, no. 1, pp. 118–127, 2007.

[231] Y. Zhang, G. Parmigiani, and W. E. Johnson, "Combat-seq: batch effect adjustment for rna-seq count data," *NAR genomics and bioinformatics*, vol. 2, no. 3, p. lqaa078, 2020.

[232] C. Müller, A. Schillert, C. Röthemeier, D.-A. Trégouët, C. Proust, H. Binder, N. Pfeiffer, M. Beutel, K. J. Lackner, R. B. Schnabel, *et al.*, "Removing batch effects from longitudinal gene expression-quantile normalization plus combat as best approach for microarray transcriptome data," *PloS one*, vol. 11, no. 6, p. e0156594, 2016.

[233] A. A. Shabalin, H. Tjelmeland, C. Fan, C. M. Perou, and A. B. Nobel, "Merging two gene-expression studies via cross-platform normalization," *Bioinformatics*, vol. 24, no. 9, pp. 1154–1160, 2008.

[234] I. Korsunsky, N. Millard, J. Fan, K. Slowikowski, F. Zhang, K. Wei, Y. Baglaenko, M. Brenner, P.-r. Loh, and S. Raychaudhuri, "Fast, sensitive and accurate integration of single-cell data with harmony," *Nature methods*, vol. 16, no. 12, pp. 1289–1296, 2019.

[235] L. Haghverdi, A. T. Lun, M. D. Morgan, and J. C. Marioni, "Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors," *Nature biotechnology*, vol. 36, no. 5, pp. 421–427, 2018.

[236] F. A. Wolf, P. Angerer, and F. J. Theis, "Scanpy: large-scale single-cell gene expression data analysis," *Genome biology*, vol. 19, pp. 1–5, 2018.

[237] K. Polański, M. D. Young, Z. Miao, K. B. Meyer, S. A. Teichmann, and J.-E. Park, "Bbknn: fast batch alignment of single cell transcriptomes," *Bioinformatics*, vol. 36, no. 3, pp. 964–965, 2020.

[238] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija, "Integrating single-cell transcriptomic data across different conditions, technologies, and species," *Nature biotechnology*, vol. 36, no. 5, pp. 411–420, 2018.

[239] D. Risso, F. Perraudeau, S. Gribkova, S. Dudoit, and J.-P. Vert, "A general and flexible method for signal extraction from single-cell rna-seq data," *Nature communications*, vol. 9, no. 1, p. 284, 2018.

[240] O. Alter, P. O. Brown, and D. Botstein, "Singular value decomposition for genome-wide expression data processing and modeling," *Proceedings of the National Academy of Sciences*, vol. 97, no. 18, pp. 10101–10106, 2000.

[241] Y. Lin, S. Ghazanfar, K. Y. Wang, J. A. Gagnon-Bartsch, K. K. Lo, X. Su, Z.-G. Han, J. T. Ormerod, T. P. Speed, P. Yang, *et al.*, "scmerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell rna-seq datasets," *Proceedings of the National Academy of Sciences*, vol. 116, no. 20, pp. 9775–9784, 2019.

[242] J. T. Leek, W. E. Johnson, H. S. Parker, A. E. Jaffe, and J. D. Storey, "The sva package for removing batch effects and other unwanted variation in high-throughput experiments," *Bioinformatics*, vol. 28, no. 6, pp. 882–883, 2012.

[243] J.-P. Fortin, E. M. Sweeney, J. Muschelli, C. M. Crainiceanu, R. T. Shinohara, A. D. N. Initiative, *et al.*, "Removing inter-subject technical variability in magnetic resonance imaging studies," *NeuroImage*, vol. 132, pp. 198–212, 2016.

[244] S. C. Karayumak, S. Bouix, L. Ning, A. James, T. Crow, M. Shenton, M. Kubicki, and Y. Rathi, "Retrospective harmonization of multi-site diffusion mri data acquired with different acquisition parameters," *Neuroimage*, vol. 184, pp. 180–200, 2019.

[245] H. Mirzaalian, L. Ning, P. Savadjiev, O. Pasternak, S. Bouix, O. Michailovich, S. Karmacharya, G. Grant, C. E. Marx, R. A. Morey, *et al.*, "Multi-site harmonization of diffusion mri data in a registration framework," *Brain imaging and behavior*, vol. 12, pp. 284–295, 2018.

[246] H. Mirzaalian, A. de Pierrefeu, P. Savadjiev, O. Pasternak, S. Bouix, M. Kubicki, C.-F. Westin, M. E. Shenton, and Y. Rathi, "Harmonizing diffusion mri data across multiple sites and scanners," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I 18*, pp. 12–19, Springer, 2015.

[247] Y. Zhang, D. F. Jenkins, S. Manimaran, and W. E. Johnson, "Alternative empirical bayes models for adjusting for batch effects in genomic studies," *BMC bioinformatics*, vol. 19, no. 1, pp. 1–15, 2018.

[248] K. M. Huynh, G. Chen, Y. Wu, D. Shen, and P.-T. Yap, "Multi-site harmonization of diffusion mri data via method of moments," *IEEE transactions on medical imaging*, vol. 38, no. 7, pp. 1599–1609, 2019.

[249] J. Wrobel, M. Martin, R. Bakshi, P. A. Calabresi, M. Elliot, D. Roalf, R. C. Gur, R. E. Gur, R. G. Henry, G. Nair, *et al.*, "Intensity warping for multisite mri harmonization," *NeuroImage*, vol. 223, p. 117242, 2020.

[250] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, D. Y. W. Solis, C. Molter, R. Duque, H. Bersini, and A. Nowé, "Geneshift: a nonparametric approach for integrating microarray gene expression data based on the inner product as a distance measure between the distributions of genes," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 10, no. 2, pp. 383–392, 2013.

[251] D. Mackin, X. Fave, L. Zhang, J. Yang, A. K. Jones, C. S. Ng, and L. Court, "Harmonizing the pixel size in retrospective computed tomography radiomics studies," *PloS one*, vol. 12, no. 9, p. e0178524, 2017.

[252] H. Moradmand, S. M. R. Aghamiri, and R. Ghaderi, "Impact of image preprocessing methods on reproducibility of radiomic features in multimodal magnetic resonance imaging in glioblastoma," *Journal of applied clinical medical physics*, vol. 21, no. 1, pp. 179–190, 2020.

[253] I. Pitas, *Digital image processing algorithms and applications*. John Wiley & Sons, 2000.

[254] R. T. Shinohara, E. M. Sweeney, J. Goldsmith, N. Shiee, F. J. Mateen, P. A. Calabresi, S. Jarso, D. L. Pham, D. S. Reich, C. M. Crainiceanu, *et al.*, "Statistical normalization techniques for magnetic resonance imaging," *NeuroImage: Clinical*, vol. 6, pp. 9–19, 2014.

[255] C. P. Loizou, V. Murray, M. S. Pattichis, I. Seimenis, M. Pantziaris, and C. S. Pattichis, "Multiscale amplitude-modulation frequency-modulation (am–fm) texture analysis of multiple sclerosis in brain mri images," *IEEE Transactions on Information Technology in Biomedicine*, vol. 15, no. 1, pp. 119–129, 2010.

[256] M. Shah, Y. Xiao, N. Subbanna, S. Francis, D. L. Arnold, D. L. Collins, and T. Arbel, "Evaluating intensity normalization on mris of human brain with multiple sclerosis," *Medical image analysis*, vol. 15, no. 2, pp. 267–282, 2011.

[257] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Computer graphics and applications*, vol. 21, no. 5, pp. 34–41, 2001.

[258] T. A. A. Tosta, P. R. de Faria, L. A. Neves, and M. Z. do Nascimento, "Color normalization of faded h&e-stained histological images using spectral matching," *Computers in biology and medicine*, vol. 111, p. 103344, 2019.

[259] A. M. Khan, N. Rajpoot, D. Treanor, and D. Magee, "A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution," *IEEE transactions on Biomedical Engineering*, vol. 61, no. 6, pp. 1729–1738, 2014.

[260] A. C. Ruifrok, D. A. Johnston, *et al.*, "Quantification of histochemical staining by color deconvolution," *Analytical and quantitative cytology and histology*, vol. 23, no. 4, pp. 291–299, 2001.

[261] M. Z. Hoque, A. Keskinarkaus, P. Nyberg, and T. Seppänen, "Retinex model based stain normalization technique for whole slide image analysis," *Computerized Medical Imaging and Graphics*, vol. 90, p. 101901, 2021.

[262] D. F. Swinehart, "The beer-lambert law," *Journal of chemical education*, vol. 39, no. 7, p. 333, 1962.

[263] V. Golkov, A. Dosovitskiy, J. I. Sperl, M. I. Menzel, M. Czisch, P. Sämann, T. Brox, and D. Cremers, "Q-space deep learning: twelve-fold shorter and model-free diffusion mri scans," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1344–1351, 2016.

[264] S. Cetin Karayumak, M. Kubicki, and Y. Rathi, "Harmonizing diffusion mri data across magnetic field strengths," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part III 11*, pp. 116–124, Springer, 2018.

[265] B. E. Dewey, C. Zhao, J. C. Reinhold, A. Carass, K. C. Fitzgerald, E. S. Sotirchos, S. Saidha, J. Oh, D. L. Pham, P. A. Calabresi, *et al.*, "Deepharmony: A deep learning approach to contrast harmonization across scanner changes," *Magnetic resonance imaging*, vol. 64, pp. 160–170, 2019.

[266] S. Park, S. M. Lee, K.-H. Do, J.-G. Lee, W. Bae, H. Park, K.-H. Jung, and J. B. Seo, "Deep learning algorithm for reducing ct slice thickness: effect on reproducibility of radiomic features in lung cancer," *Korean journal of radiology*, vol. 20, no. 10, pp. 1431–1440, 2019.

[267] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

[268] F. Zhao, Z. Wu, L. Wang, W. Lin, S. Xia, D. Shen, G. Li, and U. B. C. P. Consortium, "Harmonization of infant cortical thickness using surface-to-surface cycle-consistent adversarial networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 475–483, Springer, 2019.

[269] D. Moyer, G. Ver Steeg, C. M. Tax, and P. M. Thompson, "Scanner invariant representations for diffusion mri harmonization," *Magnetic resonance in medicine*, vol. 84, no. 4, pp. 2174–2189, 2020.

[270] N. Russkikh, D. Antonets, D. Shtokalo, A. Makarov, Y. Vyatkin, A. Zakharov, and E. Terentyev, "Style transfer with variational autoencoders is a promising approach to

rna-seq data harmonization and analysis," *Bioinformatics*, vol. 36, no. 20, pp. 5076–5085, 2020.

[271] N. Johansen and G. Quon, "scalign: a tool for alignment, integration, and rare cell identification from scrna-seq data," *Genome biology*, vol. 20, no. 1, pp. 1–21, 2019.

[272] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding.," *Journal of Machine Learning Research*, vol. 11, no. 1, 2010.

[273] S. St-Jean, P. Coupé, and M. Descoteaux, "Non local spatial and angular matching: Enabling higher spatial resolution diffusion mri datasets through adaptive denoising," *Medical image analysis*, vol. 32, pp. 115–130, 2016.

[274] X. Li, K. Wang, Y. Lyu, H. Pan, J. Zhang, D. Stambolian, K. Susztak, M. P. Reilly, G. Hu, and M. Li, "Deep learning enables accurate clustering with batch effect removal in single-cell rna-seq analysis," *Nature communications*, vol. 11, no. 1, p. 2338, 2020.

[275] T. Wang, T. S. Johnson, W. Shao, Z. Lu, B. R. Helm, J. Zhang, and K. Huang, "Bermuda: a novel deep transfer learning method for single-cell rna sequencing batch correction reveals hidden high-resolution cellular subtypes," *Genome biology*, vol. 20, no. 1, pp. 1–15, 2019.

[276] H. Guan, Y. Liu, E. Yang, P.-T. Yap, D. Shen, and M. Liu, "Multi-site mri harmonization via attention-guided deep domain adaptation for brain disorder identification," *Medical image analysis*, vol. 71, p. 102076, 2021.

[277] N. K. Dinsdale, M. Jenkinson, and A. I. Namburete, "Deep learning-based unlearning of dataset bias for mri harmonisation and confound removal," *NeuroImage*, vol. 228, p. 117689, 2021.

[278] S. Ge, H. Wang, A. Alavi, E. Xing, and Z. Bar-Joseph, "Supervised adversarial alignment of single-cell rna-seq data," *Journal of Computational Biology*, vol. 28, no. 5, pp. 501–513, 2021.

[279] Z. Rong, Q. Tan, L. Cao, L. Zhang, K. Deng, Y. Huang, Z.-J. Zhu, Z. Li, and K. Li, "Normae: deep adversarial learning model to remove batch effects in liquid chromatography mass spectrometry-based metabolomics data," *Analytical chemistry*, vol. 92, no. 7, pp. 5082–5090, 2020.

[280] M. Büttner, Z. Miao, F. A. Wolf, S. A. Teichmann, and F. J. Theis, "A test metric for assessing single-cell rna-seq batch correction," *Nature methods*, vol. 16, no. 1, pp. 43–49, 2019.

[281] D. Wang, S. Hou, L. Zhang, X. Wang, B. Liu, and Z. Zhang, "imap: integration of multiple single-cell datasets by adversarial paired transfer networks," *Genome biology*, vol. 22, no. 1, pp. 1–24, 2021.

[282] A. Kolaman and O. Yadid-Pecht, "Quaternion structural similarity: a new quality index for color images," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1526–1536, 2011.

[283] M. D. Zarella, C. Yeoh, D. E. Breen, and F. U. Garcia, "An alternative reference space for h&e color normalization," *PloS one*, vol. 12, no. 3, p. e0174489, 2017.

[284] Y. Zheng, Z. Jiang, H. Zhang, F. Xie, J. Shi, and C. Xue, "Adaptive color deconvolution for histological wsi normalization," *Computer methods and programs in biomedicine*, vol. 170, pp. 107–120, 2019.

[285] P. Maji and S. Mahapatra, "Rough-fuzzy circular clustering for color normalization of histological images," *Fundamenta Informaticae*, vol. 164, no. 1, pp. 103–117, 2019.

[286] T. A. A. Tosta, P. R. de Faria, J. P. S. Servato, L. A. Neves, G. F. Roberto, A. S. Martins, and M. Z. do Nascimento, "Unsupervised method for normalization of hematoxylin-eosin stain in histological images," *Computerized Medical Imaging and Graphics*, vol. 77, p. 101646, 2019.

[287] L. G. Nyúl, J. K. Udupa, and X. Zhang, "New variants of a method of mri scale standardization," *IEEE transactions on medical imaging*, vol. 19, no. 2, pp. 143–150, 2000.

[288] U. Pandey, J. Saini, M. Kumar, R. Gupta, and M. Ingalhalikar, "Normative baseline for radiomics in brain mri: evaluating the robustness, regional variations, and reproducibility on flair images," *Journal of Magnetic Resonance Imaging*, vol. 53, no. 2, pp. 394–407, 2021.

[289] R. Da-Ano, I. Masson, F. Lucia, M. Doré, P. Robin, J. Alfieri, C. Rousseau, A. Mervoyer, C. Reinhold, J. Castelli, *et al.*, "Performance comparison of modified combat for harmonization of radiomic features for multicenter studies," *Scientific Reports*, vol. 10, no. 1, p. 10248, 2020.

[290] X. Li and K. N. Plataniotis, "A complete color normalization approach to histopathology images using color cues computed from saturation-weighted statistics," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 7, pp. 1862–1873, 2015.

[291] S. Koppers, L. Bloy, J. I. Berman, C. M. Tax, J. C. Edgar, and D. Merhof, "Spherical harmonic residual network for diffusion signal harmonization," in *Computational Diffusion MRI: International MICCAI Workshop, Granada, Spain, September 2018 22*, pp. 173–182, Springer, 2019.

[292] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[293] P. Maji and S. Mahapatra, "Circular clustering in fuzzy approximation spaces for color normalization of histological images," *IEEE transactions on medical imaging*, vol. 39, no. 5, pp. 1735–1745, 2019.

[294] K. Wengler, C. Cassidy, M. van Der Pluijm, J. J. Weinstein, A. Abi-Dargham, E. van de Giessen, and G. Horga, "Cross-scanner harmonization of neuromelanin-sensitive mri for multisite studies," *Journal of Magnetic Resonance Imaging*, vol. 54, no. 4, pp. 1189–1199, 2021.

[295] A. Espín-Pérez, C. Portier, M. Chadeau-Hyam, K. van Veldhoven, J. C. Kleinjans, and T. M. de Kok, "Comparison of statistical methods and the use of quality control sam-

ples for batch effect correction in human transcriptome data," *PloS one*, vol. 13, no. 8, p. e0202947, 2018.

[296] S. Roy, S. Lal, and J. R. Kini, "Novel color normalization method for hematoxylin & eosin stained histopathology images," *IEEE Access*, vol. 7, pp. 28982–28998, 2019.

[297] A. Vahadane, T. Peng, A. Sethi, S. Albarqouni, L. Wang, M. Baust, K. Steiger, A. M. Schlitter, I. Esposito, and N. Navab, "Structure-preserving color normalization and sparse stain separation for histological images," *IEEE transactions on medical imaging*, vol. 35, no. 8, pp. 1962–1971, 2016.

[298] I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, pp. 255–268, 1989.

[299] A. Ibrahim, T. Refaee, R. T. Leijenaar, S. Primakov, R. Hustinx, F. M. Mottaghy, H. C. Woodruff, A. D. Maidment, and P. Lambin, "The application of a workflow integrating the variable reproducibility and harmonizability of radiomic features on a phantom dataset," *PLoS One*, vol. 16, no. 5, p. e0251147, 2021.

[300] D. Liljequist, B. Elfving, and K. Skavberg Roaldsen, "Intraclass correlation–a discussion and demonstration of basic features," *PloS one*, vol. 14, no. 7, p. e0219854, 2019.

[301] J. C. Beer, N. J. Tustison, P. A. Cook, C. Davatzikos, Y. I. Sheline, R. T. Shinohara, K. A. Linn, A. D. N. Initiative, *et al.*, "Longitudinal combat: A method for harmonizing longitudinal multi-scanner imaging data," *Neuroimage*, vol. 220, p. 117129, 2020.

[302] J. Radua, E. Vieta, R. Shinohara, P. Kochunov, Y. Quidé, M. J. Green, C. S. Weickert, T. Weickert, J. Bruggemann, T. Kircher, *et al.*, "Increased power by harmonizing structural mri site differences with the combat batch adjustment method in enigma," *Neuroimage*, vol. 218, p. 116956, 2020.

[303] J.-P. Fortin, N. Cullen, Y. I. Sheline, W. D. Taylor, I. Aselcioglu, P. A. Cook, P. Adams, C. Cooper, M. Fava, P. J. McGrath, *et al.*, "Harmonization of cortical thickness measurements across scanners and sites," *Neuroimage*, vol. 167, pp. 104–120, 2018.

[304] J.-P. Fortin, D. Parker, B. Tunç, T. Watanabe, M. A. Elliott, K. Ruparel, D. R. Roalf, T. D. Satterthwaite, R. C. Gur, R. E. Gur, *et al.*, "Harmonization of multi-site diffusion tensor imaging data," *Neuroimage*, vol. 161, pp. 149–170, 2017.

[305] J. Lan, S. Cai, Y. Xue, Q. Gao, M. Du, H. Zhang, Z. Wu, Y. Deng, Y. Huang, T. Tong, *et al.*, "Unpaired stain style transfer using invertible neural networks based on channel attention and long-range residual," *IEEE Access*, vol. 9, pp. 11282–11295, 2021.

[306] C. Wachinger, A. Rieckmann, S. Pölsterl, A. D. N. Initiative, *et al.*, "Detect and correct bias in multi-site neuroimaging datasets," *Medical Image Analysis*, vol. 67, p. 101879, 2021.

[307] H.-D. Cheng, X. Cai, and R. Min, "A novel approach to color normalization using neural network," *Neural Computing and Applications*, vol. 18, pp. 237–247, 2009.

[308] Q. Tong, T. Gong, H. He, Z. Wang, W. Yu, J. Zhang, L. Zhai, H. Cui, X. Meng, C. W. Tax, *et al.*, "A deep learning–based method for improving reliability of multicenter diffusion kurtosis imaging with varied acquisition protocols," *Magnetic Resonance Imaging*, vol. 73, pp. 31–44, 2020.

[309] A. Jog, A. Carass, S. Roy, D. L. Pham, and J. L. Prince, "Mr image synthesis by contrast learning on neighborhood ensembles," *Medical image analysis*, vol. 24, no. 1, pp. 63–76, 2015.

[310] J. Zhong, Y. Wang, J. Li, X. Xue, S. Liu, M. Wang, X. Gao, Q. Wang, J. Yang, and X. Li, "Inter-site harmonization based on dual generative adversarial networks for diffusion tensor imaging: application to neonatal white matter development," *Biomedical engineering online*, vol. 19, no. 1, pp. 1–18, 2020.

[311] A. Jog, A. Carass, S. Roy, D. L. Pham, and J. L. Prince, "Random forest regression for magnetic resonance image synthesis," *Medical image analysis*, vol. 35, pp. 475–488, 2017.

[312] C. K. Stein, P. Qu, J. Epstein, A. Buros, A. Rosenthal, J. Crowley, G. Morgan, and B. Barlogie, "Removing batch effects from purified plasma cell gene expression microarrays with modified combat," *BMC bioinformatics*, vol. 16, no. 1, pp. 1–9, 2015.

[313] M. Ren, N. Dey, J. Fishbaugh, and G. Gerig, "Segmentation-renormalized deep feature modulation for unpaired image harmonization," *IEEE transactions on medical imaging*, vol. 40, no. 6, pp. 1519–1530, 2021.

[314] H. Beaumont, A. Iannessi, A.-S. Bertrand, J. M. Cucchi, and O. Lucidarme, "Harmonization of radiomic feature distributions: impact on classification of hepatic tissue in ct imaging," *European Radiology*, vol. 31, pp. 6059–6068, 2021.

[315] H. M. Whitney, H. Li, Y. Ji, P. Liu, and M. L. Giger, "Harmonization of radiomic features of breast lesions across international dce-mri datasets," *Journal of Medical Imaging*, vol. 7, no. 1, pp. 012707–012707, 2020.

[316] A. Llera, I. Huertas, P. Mir, and C. F. Beckmann, "Quantitative intensity harmonization of dopamine transporter spect images using gamma mixture models," *Molecular imaging and biology*, vol. 21, pp. 339–347, 2019.

[317] P. Chirra, P. Leo, M. Yim, B. N. Bloch, A. R. Rastinehad, A. Purysko, M. Rosen, A. Madabhushi, and S. E. Viswanath, "Multisite evaluation of radiomic feature reproducibility and discriminability for identifying peripheral zone prostate tumors on mri," *Journal of Medical Imaging*, vol. 6, no. 2, pp. 024502–024502, 2019.

[318] G. S. Ioannidis, E. Trivizakis, I. Metzakis, S. Papagiannakis, E. Lagoudaki, and K. Marias, "Pathomics and deep learning classification of a heterogeneous fluorescence histology image dataset," *Applied Sciences*, vol. 11, no. 9, p. 3796, 2021.

[319] M. Ingalhalikar, S. Shinde, A. Karmarkar, A. Rajan, D. Rangaprakash, and G. Deshpande, "Functional connectivity-based prediction of autism on site harmonized abide dataset," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 12, pp. 3628–3637, 2021.

[320] S. Kothari, J. H. Phan, T. H. Stokes, A. O. Osunkoya, A. N. Young, and M. D. Wang, "Removing batch effects from histopathological images for enhanced cancer diagnosis," *IEEE journal of biomedical and health informatics*, vol. 18, no. 3, pp. 765–772, 2013.

[321] C. T. Arendt, D. Leithner, M. E. Mayerhoefer, P. Gibbs, C. Czerny, C. Arnoldner, I. Burck, M. Leinung, Y. Tanyildizi, L. Lenga, *et al.*, "Radiomics of high-resolution computed tomography for the differentiation between cholesteatoma and middle ear inflammation: effects of post-reconstruction methods in a dual-center study," *European Radiology*, vol. 31, pp. 4071–4078, 2021.

[322] M. Lotfollahi, F. A. Wolf, and F. J. Theis, "scgen predicts single-cell perturbation responses," *Nature methods*, vol. 16, no. 8, pp. 715–721, 2019.

[323] B. Hie, B. Bryson, and B. Berger, "Efficient integration of heterogeneous single-cell transcriptomes using scanorama," *Nature biotechnology*, vol. 37, no. 6, pp. 685–691, 2019.

[324] U. Shaham, K. P. Stanton, J. Zhao, H. Li, K. Raddassi, R. Montgomery, and Y. Kluger, "Removal of batch effects using distribution-matching residual networks," *Bioinformatics*, vol. 33, no. 16, pp. 2539–2546, 2017.

[325] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information fusion*, vol. 58, pp. 82–115, 2020.

[326] G. Yang, Q. Ye, and J. Xia, "Unbox the black-box for the medical explainable ai via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond," *Information Fusion*, vol. 77, pp. 29–52, 2022.

[327] A. Holzinger, M. Dehmer, F. Emmert-Streib, R. Cucchiara, I. Augenstein, J. Del Ser, W. Samek, I. Jurisica, and N. Díaz-Rodríguez, "Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence," *Information Fusion*, vol. 79, pp. 263–278, 2022.

[328] F. J. López-González, J. Silva-Rodríguez, J. Paredes-Pacheco, A. Niñerola-Baizán, N. Efthimiou, C. Martín-Martín, A. Moscoso, Á. Ruibal, N. Roé-Vellvé, and P. Aguiar, "Intensity normalization methods in brain fdg-pet quantification," *Neuroimage*, vol. 222, p. 117229, 2020.

[329] H. Liu, R. Wang, S. Shan, and X. Chen, "Deep supervised hashing for fast image retrieval," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2064–2072, 2016.

[330] Z. Li, X. Zhang, H. Müller, and S. Zhang, "Large-scale retrieval for medical image analytics: A comprehensive review," *Medical image analysis*, vol. 43, pp. 66–84, 2018.

[331] C. Yan, B. Gong, Y. Wei, and Y. Gao, "Deep multi-view enhancement hashing for image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 4, pp. 1445–1451, 2020.

[332] R. Xia, Y. Pan, H. Lai, C. Liu, and S. Yan, "Supervised hashing for image retrieval via image representation learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 28, 2014.

[333] H. Lai, Y. Pan, Y. Liu, and S. Yan, "Simultaneous feature learning and hash coding with deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3270–3278, 2015.

[334] L. Yuan, T. Wang, X. Zhang, F. E. Tay, Z. Jie, W. Liu, and J. Feng, "Central similarity quantization for efficient image and video retrieval," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3083–3092, 2020.

[335] Z. Cao, M. Long, J. Wang, and P. S. Yu, "Hashnet: Deep learning to hash by continuation," in *Proceedings of the IEEE international conference on computer vision*, pp. 5608–5617, 2017.

[336] Y. Cao, M. Long, B. Liu, and J. Wang, "Deep cauchy hashing for hamming space retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1229–1237, 2018.

[337] L. Fan, K. W. Ng, C. Ju, T. Zhang, and C. S. Chan, "Deep polarized network for supervised learning of accurate binary hashing codes.," in *IJCAI*, pp. 825–831, 2020.

[338] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12116–12128, 2021.

[339] J. Wang, T. Zhang, N. Sebe, H. T. Shen, *et al.*, "A survey on learning to hash," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 769–790, 2017.

[340] X. Zheng, Y. Zhang, and X. Lu, "Deep balanced discrete hashing for image retrieval," *Neurocomputing*, vol. 403, pp. 224–236, 2020.

[341] Y. Chen, Z. Lai, Y. Ding, K. Lin, and W. K. Wong, "Deep supervised hashing with anchor graph," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9796–9804, 2019.

[342] R.-C. Tu, X.-L. Mao, C. Kong, Z. Shao, Z.-L. Li, W. Wei, and H. Huang, "Weighted gaussian loss based hamming hashing," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 3409–3417, 2021.

[343] Q.-Y. Jiang and W.-J. Li, "Asymmetric deep supervised hashing," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[344] Y. K. Jang and N. I. Cho, "Generalized product quantization network for semi-supervised image retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3420–3429, 2020.

[345] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2, pp. 1398–1402, Ieee, 2003.

[346] M. Amgad, H. Elfandy, H. Hussein, L. A. Atteya, M. A. Elsebaie, L. S. Abo Elnasr, R. A. Sakr, H. S. Salem, A. F. Ismail, A. M. Saad, *et al.*, "Structured crowdsourcing enables convolutional segmentation of histology images," *Bioinformatics*, vol. 35, no. 18, pp. 3461–3467, 2019.

[347] S. Graham, M. Jahanifar, A. Azam, M. Nimir, Y.-W. Tsang, K. Dodd, E. Hero, H. Sahota, A. Tank, K. Benes, *et al.*, "Lizard: a large-scale dataset for colonic nuclear instance segmentation and classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 684–693, 2021.

[348] X. Mei, Z. Liu, P. M. Robson, B. Marinelli, M. Huang, A. Doshi, A. Jacobi, C. Cao, K. E. Link, T. Yang, *et al.*, "Radimagenet: an open radiologic deep learning research dataset for effective transfer learning," *Radiology: Artificial Intelligence*, vol. 4, no. 5, p. e210315, 2022.

[349] Q. Li, Z. Sun, R. He, and T. Tan, "Deep supervised discrete hashing," *Advances in neural information processing systems*, vol. 30, 2017.

[350] Y. Li, W. Pei, J. van Gemert, *et al.*, "Push for quantization: Deep fisher hashing," *arXiv preprint arXiv:1909.00206*, 2019.

[351] Y. Chen and X. Lu, "Deep discrete hashing with pairwise correlation learning," *Neurocomputing*, vol. 385, pp. 111–121, 2020.

[352] N. Passalis and A. Tefas, "Deep supervised hashing using quadratic spherical mutual information for efficient image retrieval," *Signal Processing: Image Communication*, vol. 93, p. 116146, 2021.

[353] T. Li, Z. Zhang, L. Pei, and Y. Gan, "Hashformer: Vision transformer based deep hashing for image retrieval," *IEEE Signal Processing Letters*, vol. 29, pp. 827–831, 2022.

[354] Y. Chen, S. Zhang, F. Liu, Z. Chang, M. Ye, and Z. Qi, "Transhash: Transformer-based hamming hashing for efficient image retrieval," in *Proceedings of the 2022 International Conference on Multimedia Retrieval*, pp. 127–136, 2022.

[355] J. Fang, H. Fu, and J. Liu, "Deep triplet hashing network for case-based medical image retrieval," *Medical image analysis*, vol. 69, p. 101981, 2021.

[356] A. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," *IEEE Transactions on Medical imaging*, vol. 19, no. 3, pp. 203–210, 2000.

[357] E. Calabrese, J. E. Villanueva-Meyer, J. D. Rudie, A. M. Rauschecker, U. Baid, S. Bakas, S. Cha, J. T. Mongan, and C. P. Hess, "The university of california san francisco preoperative diffuse glioma mri dataset," *Radiology: Artificial Intelligence*, vol. 4, no. 6, p. e220058, 2022.

[358] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International journal of computer vision*, vol. 111, pp. 98–136, 2015.

[359] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[360] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021.