

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Semi-Supervised Learning for Interstitial Lung Disease Finding Segmentation in High-Resolution Computed Tomography

Mariana Filipa Alves dos Santos



Master in Biomedical Engineering

Supervisor: Prof. João Pedrosa

Co-Supervisor: Prof. Tânia Melo

July 15, 2026

Semi-Supervised Learning for Interstitial Lung Disease Finding Segmentation in High-Resolution Computed Tomography

Mariana Filipa Alves dos Santos

Master in Biomedical Engineering

July 15, 2026

Abstract

Interstitial lung disease (ILD) comprises a diverse group of disorders that primarily affect the lung parenchyma. High-resolution computed tomography (HRCT) is the reference imaging modality for ILD assessment, but its interpretation is challenging due to overlapping radiological patterns among disease subtypes. These complexities underscore the need for automated and reproducible computational methods capable of reliably segmenting ILD-related abnormalities.

Deep learning (DL) methods have shown promising results in medical image segmentation. However, most existing approaches for segmenting ILD findings rely on fully supervised learning (SL), which requires large amounts of reliable pixel-level annotations that are laborious to obtain. This presents a major limitation, as publicly available ILD datasets often contain incomplete annotations, leaving a significant portion of the data unlabelled. As a result, semi-supervised learning (SSL) has emerged as a promising strategy to use both labelled and unlabelled data, although it remains unclear which SSL methods are most effective for ILD segmentation.

In the first phase, a supervised baseline for multi-class ILD findings segmentation was established using the public MedGIFT database. A 2D nnU-Net framework was adopted to evaluate different strategies to handle sparse annotations and alternative network configurations. The results indicated that segmentation performance is significantly influenced by how non-annotated pixels are treated, with better results obtained when these regions were ignored during supervised training. Among the evaluated configurations, the reduced-patch U-Net configuration achieved the best supervised performance, with an average Dice Similarity Coefficient (DSC) of 43.12%.

In the second phase, SSL strategies were explored, including pseudo-labelling, entropy minimization, consistency regularization, FixMatch-based methods, and an ensemble-based pseudo-label generation pipeline. The ensemble-based strategy yielded the best results, improving the average DSC by 5.36 percentage points over the best supervised baseline. This illustrates that SSL can enhance segmentation performance, especially for classes that are sufficiently represented initially.

In conclusion, this work shows that SSL can improve ILD findings segmentation in HRCT images. However, its effectiveness remains strongly dependent on the labelled data used to guide training. In the public dataset considered, incomplete and highly imbalanced annotations limited the ability of SSL methods to improve the segmentation of poorly represented findings, revealing that unlabelled data alone cannot compensate for weak initial supervision. Furthermore, the results emphasise that annotation quality is as critical as annotation quantity, as coarse or noisy labels, particularly when pathological regions are mixed with normal lung tissue, continue to hinder model performance. Overall, the findings of this dissertation suggest that SSL is a promising direction for ILD segmentation, but its full potential depends on the availability of a small, yet reliable and clinically consistent labelled set.

Keywords: Interstitial Lung Disease, Medical Image Segmentation, Supervised Learning, Semi-Supervised Learning, Deep Learning.

Resumo

A doença pulmonar intersticial (ILD) engloba um grupo heterogêneo de patologias que afetam predominantemente o parênquima pulmonar. A tomografia computadorizada de alta resolução (HRCT) constitui a modalidade de imagem de referência para a avaliação da ILD. Contudo, a sua interpretação é desafiante devido à sobreposição de padrões radiológicos entre os diferentes subtipos da doença. Esta complexidade evidencia a necessidade de métodos computacionais automatizados e reprodutíveis, capazes de segmentar de forma fiável as alterações associadas à ILD.

As técnicas de *deep learning* (DL) têm demonstrado resultados promissores na segmentação de imagens médicas. No entanto, a maioria das abordagens existentes para a segmentação de achados de ILD baseiam-se em aprendizagem supervisionada (SL), que requer grandes quantidades de anotações ao nível do pixel fiáveis, cuja obtenção é trabalhosa. Isto constitui uma limitação importante, uma vez que as bases de dados de ILD disponíveis publicamente apresentam frequentemente anotações incompletas, deixando uma parte significativa dos dados sem anotação. Neste contexto, a aprendizagem semi-supervisionada (SSL) tem vindo a afirmar-se como uma estratégia promissora para aproveitar simultaneamente dados anotados e não anotados. Ainda assim, permanece por esclarecer quais os métodos de SSL mais adequados para a segmentação de ILD.

Numa primeira fase, foi estabelecida uma *baseline* supervisionada para a segmentação multiclasse de achados de ILD, recorrendo à base de dados pública MedGIFT. Foi utilizada uma arquitetura *nnU-Net* 2D para avaliar diferentes estratégias de tratamento de anotações esparsas e distintas configurações da rede. Os resultados demonstraram que o desempenho da segmentação é significativamente influenciado pela forma como os píxeis não anotados são tratados, tendo-se obtido melhores resultados quando estas regiões foram ignoradas durante o processo de treino supervisionado. Entre as configurações avaliadas, a *nnU-Net* com *patch* reduzido alcançou o melhor desempenho supervisionado, com um Coeficiente de Similaridade de Dice (DSC) médio de 43.12%.

Numa segunda fase, foram exploradas diferentes estratégias de SSL, incluindo *pseudo-labelling*, *entropy minimization*, *consistency regularization*, métodos baseados em *FixMatch* e uma *pipeline* de geração de *pseudo-labels* através de *ensemble*. A estratégia baseada em *ensemble* apresentou o melhor desempenho, aumentando o DSC médio em 5.36 pontos percentuais relativamente à melhor *baseline* supervisionada. Estes resultados demonstram que a SSL pode contribuir para melhorar o desempenho da segmentação, particularmente em classes que dispõem de representação inicial suficiente.

Em suma, este trabalho demonstra que a SSL pode melhorar a segmentação de achados de ILD em imagens de HRCT. Contudo, a sua eficácia depende fortemente dos dados anotados utilizados para guiar o treino. No conjunto de dados público considerado, as anotações incompletas e altamente desequilibradas limitaram a capacidade dos métodos de SSL em melhorar a segmentação de achados pouco representados, revelando que os dados não anotados por si só não conseguem compensar uma supervisão inicial fraca. Além disso, os resultados sublinham que a qualidade das anotações é tão crítica quanto a sua quantidade, uma vez que anotações imprecisas ou rui-

dosas, nomeadamente quando regiões patológicas confundem-se com tecido pulmonar normal, continuam a comprometer o desempenho do modelo. Concluindo, os resultados desta dissertação sugerem que a SSL é uma direção promissora para a segmentação eficiente de ILD, mas o seu potencial depende da disponibilidade de um conjunto anotado pequeno, mas fiável e clinicamente consistente.

Palavras-chave: Doença Pulmonar Intersticial, Segmentação de Imagem Médica, Aprendizagem Supervisionada, Aprendizagem Semi-Supervisionada, *Deep Learning*.

UN Sustainable Development Goals

The United Nations Sustainable Development Goals (SDGs) provide a global framework for achieving a better and more sustainable future for all. They include 17 goals¹ addressing major global challenges, including health, inequality, climate change, environmental sustainability, peace, and justice.

This dissertation is aligned with SDG 3, which aims to ensure healthy lives and promote well-being for all at all ages. In particular, this work contributes to research on computational methods for the analysis of ILD in HRCT images. By developing supervised and semi-supervised deep learning strategies for the segmentation of ILD findings, the dissertation addresses an important step towards more automated and reproducible characterisation of lung abnormalities.

Although the developed methods are not intended for direct clinical deployment, they may contribute to future decision-support tools by assisting in the localisation and quantification of pathological patterns in lung imaging. This is particularly relevant for chronic respiratory diseases, where imaging plays an important role in diagnosis, monitoring, and disease progression assessment.

The specific Sustainable Development Goal considered in this dissertation is:

SDG 3 Ensure healthy lives and promote well-being for all at all ages.

SDG	Target	Contribution	Performance Indicators and Metrics
3	3.4	Developed supervised and semi-supervised segmentation methods for ILD findings in HRCT images, contributing to automated analysis of respiratory diseases.	Improvement in diagnostic consistency. Increased ability to quantify disease extent and monitor progression.
	3.8	Explored strategies that reduce dependence on dense manual annotations by using sparse labels and unlabelled data for model training.	Reduction in clinician annotation workload. Increased scalability and reproducibility of lung imaging analysis workflows.

¹<https://sdgs.un.org/goals>

Agradecimentos

Ao concluir esta dissertação, chega também ao fim um percurso académico que ficou marcado por desafios, conquistas e aprendizagens que levarei comigo para o futuro. Por isso, gostaria de expressar a minha sincera gratidão a todos aqueles que me acompanharam ao longo desta jornada e que, de diferentes formas, contribuíram para a minha formação académica, pessoal e profissional.

Em primeiro lugar, gostaria de agradecer ao Professor João Pedrosa e à Professora Tânia Melo pela orientação, confiança e dedicação demonstrada ao longo deste percurso. A oportunidade de integrar o Laboratório de Imagem Biomédica do INESC TEC permitiu-me crescer não só a nível académico, mas também pessoal, tornando esta experiência particularmente enriquecedora. À Professora Ana Maria Mendonça e à Carla, deixo também uma palavra de apreço pela disponibilidade, pelas valiosas contribuições e pelo conhecimento partilhado, que também foram fundamentais para a concretização deste trabalho.

Aos meus pais, um agradecimento que nenhuma palavra conseguirá verdadeiramente traduzir. Por todo o apoio incondicional que me deram ao longo deste percurso, e por terem sido sempre o meu maior exemplo e a minha maior força. Nada disto teria sido possível sem vocês, e esta conquista é, em grande parte, vossa.

Ao Vasco, pela força, paciência e compreensão e por ter sido uma presença constante ao longo desta etapa. Por partilhar comigo os momentos mais desafiantes e por tornar cada conquista ainda mais especial. Sem ti, este percurso teria sido, sem dúvida, muito mais difícil.

Por último, gostaria de agradecer aos meus amigos por todos os momentos de amizade, diversão e descontração que tornaram esta caminhada mais leve e especial. Deixo também uma palavra muito especial à minha melhor amiga de 4 patas, que me acompanha há mais de 12 anos e que, sem nunca pronunciar uma única palavra, sempre soube estar presente. Entre dias bons e menos bons, foi uma companhia constante, leal e reconfortante, acompanhando de perto uma parte tão importante da minha vida. Por tudo isso, merece também um lugar nestes agradecimentos.

Mariana Filipa Alves dos Santos

Contents

1	Introduction	1
1.1	Context	1
1.2	Motivation	2
1.3	Objectives	2
1.4	Document Structure	2
2	Clinical Background	4
2.1	Interstitial Lung Disease	4
2.2	ILD Diagnosis	5
2.2.1	High-Resolution Computed Tomography	6
2.3	Radiological Findings in ILD	7
3	Literature Review	13
3.1	ILD Findings Segmentation	13
3.1.1	Search Methodology	13
3.1.2	Structuring and Overview of Previous Studies	14
3.1.3	Supervised Learning Approaches	18
3.1.4	Semi-Supervised Learning Approaches	22
3.1.5	Active and Reinforcement Learning Approaches	25
3.2	Semi-Supervised Learning for Medical Image Segmentation	27
3.2.1	Pseudo-Labelling-Based Methods	27
3.2.2	Consistency Regularization-Based Methods	29
3.2.3	Consistency Regularization Methods Supervised by Pseudo-Labels	32
3.2.4	Methods Combining Contrastive Learning and Consistency Regularization	34
3.3	Future Directions in ILD Findings Segmentation	35
4	Supervised Learning for ILD Findings Segmentation	37
4.1	Methodology	37
4.1.1	Data	37
4.1.2	Model Architecture and Training Details	40
4.1.3	Evaluation	44
4.1.4	Supervision Strategies for Sparse Annotations	45
4.1.5	Network Configurations and Variants	47
4.2	Results and Discussion	48
4.2.1	Supervision Strategies for Sparse Annotations	48
4.2.2	Network Configurations and Variants	53
4.3	Summary	56

5	Semi-Supervised Learning for ILD Findings Segmentation	58
5.1	Methodology	58
5.1.1	Data	59
5.1.2	SSL via Training-Time Signal Integration	60
5.1.3	SSL via Ensemble Pseudo-Label Generation	67
5.2	Results and Discussion	69
5.2.1	SSL via Training-Time Signal Integration	69
5.2.2	SSL via Ensemble Pseudo-Label Generation	74
5.3	Summary	80
6	Conclusions and Future Work	83
	References	86
A	Additional Results	95
A.1	Best Supervised Configuration	95
A.2	Best Training-Time SSL Method	95
A.3	Best Ensemble-Based Pseudo-Labeling Model	96

List of Figures

2.1	Comparison between healthy and abnormal lung tissue [13].	5
2.2	Comparison between a standard CT and HRCT of the lung parenchyma: (a) Standard CT; (b) HRCT [26].	6
2.3	HU scale, mapping values to the represented tissue, and an axial slice of a CT scan showing the inside of the lung [29].	7
2.4	Representative HRCT findings characteristic of pulmonary fibrosis: (a) reticulation [30]; (b) traction bronchiectasis [31]; (c) honeycombing [31]; (d) architectural distortion [32]; and (e) volume loss [33].	8
2.5	Representative HRCT findings commonly observed in ILD: (a) nodules [31]; (b) ground-glass opacity [27]; (c) consolidation [31]; (d) cysts [30]; (e) emphysema [26]; (f) Mosaic attenuation [34]; (g) Pleural effusion [37]; (h) Crazy paving [34] and (i) Pleural plaques [38].	9
2.6	HRCT appearances of (a) perilymphatic nodules; (b) centrilobular nodules [31]. .	11
3.1	Schematic overview of the categorisation of the studies included in the literature review.	14
3.2	Architecture of the AtlasNet framework [45].	19
3.3	The proposed combination model with the two training datasets [49].	19
3.4	Two stage training strategy: transfer learning of the pre-trained model and screening of ILD findings [5].	21
3.5	The overall network structure with DAC and RMP [51].	21
3.6	Overall workflow of semantic segmentation [52].	22
3.7	Overview of the segmentation propagation algorithm [54].	23
3.8	Proposed ESsegILD framework for ILD segmentation. It comprises MT training with labelled (D_l) and unlabelled (D_u) data, pseudo-label selection, and iterative retraining [57].	25
3.9	U-Net-RGD architecture and the active learning workflow [46].	26
3.10	Overall pipeline of the proposed CTSA framework. The model is mainly divided into three parts: the IMG module, the DRL module and the constraint segmentation module [47].	27
3.11	Venn diagram summarizing representative semi-supervised segmentation methods and their intersections across main methodological categories.	28
3.12	Illustration of the curriculum SSL strategy. Adapted from [84].	29
3.13	The classic architectures used in consistency learning. (a) Π -Model [87]; (b) Temporal ensembling [88]; (c) The mean-teacher architecture [89].	31
3.14	Overview of the PixelFixMatch framework for semi-supervised medical image segmentation [95].	32

3.15	Overview of the MLRPL framework: (a) pseudo-label reliability assessed via intra-class feature similarity; (b) selection of reliable pseudo-labels through comparison of confidence scores from decoders (C_a and C_b), retaining higher-confidence predictions. Adapted from [96].	33
3.16	Overview of the BCP framework within the MT architecture. Adapted from [98].	33
3.17	Overview of semi-supervised segmentation framework with weak-to-strong feature perturbations and edge-aware contrastive learning. [99].	35
4.1	Representative examples of annotations in the MedGIFT ILD database, illustrating its partially labelled nature.	39
4.2	The architecture of the specific 2D nnU-Net network used.	42
4.3	Qualitative comparison of the segmentation masks obtained with the four supervision strategies, including representative HRCT slices, the corresponding GT annotations, and the predicted masks generated by each strategy.	51
4.4	Pixel-wise confusion matrix normalized by true label for the best performing model configuration.	55
5.1	Overview of the consistency regularization framework adopted.	63
5.2	Overview of the FixMatch framework adopted.	65
5.3	Overview of the ensemble-based pseudo-label generation pipeline. Stage 0 corresponds to the original supervised models, while stages 1 and 2 represent the first and second rounds of pseudo-label generation and retraining, respectively, applied to the MedGIFT and HSJ datasets.	68
5.4	Qualitative comparison of the segmentation masks obtained with the supervised baseline and the best-performing variant of each training-time SSL method, including representative HRCT slices, the corresponding GT annotations, and the predicted masks generated by each strategy.	72
5.5	Examples of pseudo-labels generated by the ensemble-based pipeline for slices without original manual annotations. Examples are shown for (a) MedGIFT data and (b) HSJ data.	77
5.6	Qualitative comparison of the segmentation masks obtained with configuration 2 from the supervised setting, the first stage and the second stage model of the ensemble-based pseudo-labelling pipeline, including representative HRCT slices, the corresponding GT annotations, and the predicted masks generated by each model.	78
5.7	Pixel-wise confusion matrix normalized by true label for the second stage configuration 2 model of the ensemble-based pseudo-labelling pipeline.	80

List of Tables

3.1	Summary of DL approaches for ILD findings segmentation. "NI", "Fib", "Hc", "Mno", "Ro", "Bronch", "Ggo", "Cons", "Emph" and "Pne" represent normal lung, fibrosis, honeycombing, micronodules, reticular opacity, bronchiectasis, ground-glass opacity, consolidation, emphysema, pneumothorax, respectively. NR indicates not reported in the study and "B-Acc" indicates balanced accuracy.	16
4.1	Summary of the original MedGIFT labels associated with each class, together with the number of studies, axial slices, and percentage of pixels.	38
4.2	Distribution of patients across pathological findings and folds.	40
4.3	Distribution of slices across pathological findings and folds.	40
4.4	Summary of the four supervision strategies for sparse annotations.	47
4.5	Segmentation performance of different strategies measured by the DSC. Values are reported as mean $\pm$ standard deviation (%) across five data splits. The average DSC is computed only over the major findings. The cysts class is omitted, as it achieved $0.00 \pm 0.00\%$ across all experiments.	49
4.6	Segmentation performance of different configurations measured by the DSC. Values are reported as mean $\pm$ standard deviation (%) across five data splits. The "Config" refers to the experimental configuration used in each setup. The cysts class is omitted, as it achieved $0.00 \pm 0.00\%$ across all experiments.	53
4.7	Comparison of segmentation performance measured by DSC (%) between the best-performing configuration proposed in this dissertation and selected state-of-the-art studies.	56
5.1	Segmentation performance of the supervised baseline and the training-time SSL methods and their variants, measured by the DSC. Values are reported as mean $\pm$ standard deviation (%) across five data splits. Bold values indicate the best result per class and shaded grey cells indicate improvement over the supervised baseline. The cysts class is omitted, as it achieved $0.00 \pm 0.00\%$ across all experiments. . .	70
5.2	Segmentation performance of the models obtained at different stages of the ensemble-based pseudo-labelling pipeline, measured by the DSC. Values are reported as mean $\pm$ standard deviation (%) across five data splits. The term "Config" refers to the experimental configuration used in each setup. Bold values indicate the best result per class, and shaded cells indicate improvement over the supervised baseline of the corresponding configuration, with grey, cyan, and orange representing configurations 1, 2, and 3, respectively. The cysts class is omitted, as it achieved $0.00 \pm 0.00\%$ across all experiments.	74

A.1	Additional segmentation performance metrics for the best supervised configuration measured by recall, precision, and specificity. Values are reported as mean $\pm$ standard deviation (%) across five data splits.	95
A.2	Additional segmentation performance metrics for the best training-time SSL method measured by recall, precision, and specificity. Values are reported as mean $\pm$ standard deviation (%) across five data splits.	95
A.3	Additional segmentation performance metrics for the best ensemble-based pseudo-labelling model measured by recall, precision, and specificity. Values are reported as mean $\pm$ standard deviation (%) across five data splits.	96

List of Acronyms

2D	Two-Dimensional
3D	Three-Dimensional
AL	Active Learning
CAD	Computer-Aided Diagnosis
CED	Convolutional Encoder-Decoder
CNN	Convolutional Neural Networks
CRF	Conditional Random Field
CT	Computed Tomography
DAC	Dense Atrous Convolution
DL	Deep Learning
DSC	Dice Similarity Coefficient
EMA	Exponential Moving Average
FP	False Positives
FRF	Fibrosis-related Findings
GT	Ground Truth
HRCT	High-Resolution Computed Tomography
HSJ	São João Hospital
HU	Hounsfield Units
ILD	Interstitial Lung Disease
IPF	Idiopathic Pulmonary Fibrosis
MLP-Mixers	Multi-Layer Perceptron Mixers
MT	Mean Teacher
PFT	Pulmonary Function Test
RL	Reinforcement Learning
RMP	Residual Multi-kernel Pooling
ROI	Regions of Interest
SL	Supervised Learning
SSL	Semi-Supervised Learning
SSc	Systemic Sclerosis
SVM	Support Vector Machine
VLM	Vision Language Models

Chapter 1

Introduction

1.1 Context

Interstitial lung disease (ILD) are a heterogeneous group of more than 200 disorders that primarily affect the pulmonary interstitium, leading to inflammation and fibrosis. The resulting thickening of the alveolar walls disrupts the normal lung architecture, impairing gas exchange and significantly reducing patients' quality of life [1]. ILD is an increasingly significant global health issue, with a reported prevalence ranging from 6.3 to 71 cases per 100,000 people [2].

The diagnosis of ILD is partially dependent on medical history, physical examination, and pulmonary function tests (PFT). Yet, high-resolution computed tomography (HRCT) is the key to the diagnosis, often eliminating the need for invasive lung biopsy [3]. By acquiring thin-slice images with high spatial resolution, HRCT provides detailed visualization of parenchymal abnormalities. However, most findings are non-specific, with considerable overlap in radiological appearance between different diseases, and even the same disease may exhibit variable manifestations [4].

As a result, examining all HRCT sections with consistent accuracy and efficiency can be an exhausting task, even for experienced radiologists [5]. Moreover, visual assessment of ILD findings remains largely qualitative and subject to inter- and intra-observer variability, limiting diagnostic reproducibility. Consequently, automated and quantitative approaches could offer valuable support, assisting radiologists in achieving a more reliable and stratified differential diagnosis [5–7].

In response to this need, deep learning (DL) techniques have been increasingly investigated for the automated segmentation of pulmonary abnormalities in ILD. In particular, convolutional neural networks (CNN) have demonstrated strong capability in learning complex textural patterns [3].

Despite these advances, most state-of-the-art DL-based segmentation methods rely on fully supervised learning (SL) paradigms, which require large volumes of pixel-level annotated data. In the medical domain, and especially for ILD, such annotations are costly, time-consuming, and subject to observer variability, resulting in limited and often incomplete datasets that restrict model scalability [8, 9].

To address this constraint, semi-supervised learning (SSL) strategies have emerged as a promising alternative, seeking to reduce annotation requirements by jointly leveraging labelled and unlabelled data during training. Although SSL has shown encouraging results in other medical imaging applications, its adoption for ILD findings segmentation remains relatively limited.

1.2 Motivation

Although DL-based methods have shown encouraging results for ILD findings segmentation, these have often been reported on private datasets with more extensive annotation protocols. In contrast, the publicly available ILD dataset contains sparse and incomplete annotations, leaving large portions of the lung unlabelled.

This limitation motivates the exploration of learning strategies that can operate effectively under incomplete supervision and make use of the information contained in unlabelled image regions. Such approaches have the potential to improve segmentation performance while reducing the dependence on exhaustive manual annotation, which remains one of the main barriers to the development of scalable medical image analysis systems.

1.3 Objectives

In this context, the main goal of this work is to develop an automatic segmentation method for ILD findings in HRCT images, employing both supervised and semi-supervised DL techniques. The specific goals are:

1. Review the state of the art in ILD segmentation and SSL approaches applied to medical imaging.
2. Develop a supervised segmentation model to establish a performance baseline and obtain preliminary results.
3. Test and compare various semi-supervised strategies that integrates unannotated data, aiming to enhance model generalization and accuracy.
4. Validate the developed algorithms.

1.4 Document Structure

In addition to the introduction, this report contains five more chapters.

Chapter 2 introduces the fundamental concepts related to ILD. It presents an overview of the disease, its diagnostic methods, including HRCT, and the main radiological findings.

Chapter 3 provides a comprehensive review of previous studies on ILD findings segmentation using DL. The chapter first focuses on SL approaches and then expands to SSL strategies applied to

medical image segmentation. Finally, it discusses emerging trends and potential future directions for ILD findings segmentation.

Chapter 4 presents the SL stage developed for ILD findings segmentation. It describes the MedGIFT ILD database, the adopted class grouping, and the data splitting strategy. The chapter then introduces the supervised segmentation methodology, including the supervision strategies designed to address sparse annotations and the network configurations evaluated. Finally, it presents the quantitative and qualitative results obtained with the supervised experiments, establishing the baseline for the subsequent semi-supervised stage.

Chapter 5 presents the SSL stage of this dissertation. It describes the additional unlabelled data used and introduces the SSL methods explored, including pseudo-labelling, entropy minimization, consistency regularization, FixMatch-based approaches, and ensemble-based pseudo-label generation. The chapter then reports and discusses the corresponding results, comparing the semi-supervised models with the supervised baselines and analysing the impact and limitations of exploiting unlabelled data.

Chapter 6 concludes the dissertation by summarising the main findings and contributions of the work. It also discusses the limitations of the proposed approaches and outlines possible directions for future research in ILD findings segmentation.

Chapter 2

Clinical Background

This chapter introduces the fundamental concepts related to ILD. It presents an overview of the disease, its diagnostic methods, and the main radiological findings that serve as the foundation for this study.

2.1 Interstitial Lung Disease

The lungs are essential components of the respiratory system, primarily responsible for facilitating gas exchange between the environment and the bloodstream [10]. However, like any complex biological system, the lungs are vulnerable to a variety of diseases that can impair their normal function and jeopardize overall health. Their delicate structure makes them particularly susceptible to conditions such as ILD [11]. ILD, also known as diffuse parenchymal lung diseases, represents a multitude of pathologies, many—but not all—of which primarily affect the pulmonary interstitium [12]. Bounded on one side by the alveolar epithelium and on the other by the capillary endothelium, the interstitial space contains lymphatic vessels, occasional fibroblasts, and extracellular matrix proteins, such as collagen. In healthy individuals, the interstitium provides structural support to the alveoli while remaining only a few micrometers thick, allowing efficient gas exchange between air and blood. In cases of ILD, inflammation and fibrosis within this space disrupt the normal architecture of the lungs, leading to impaired gas transfer. This impairment can result in breathlessness and, in severe cases, respiratory failure or even death [1]. This structural difference between healthy and diseased lung tissue is illustrated in Figure 2.1, which compares the normal alveolar interstitium with the thickened interstitium characteristic of ILD.

More than 200 distinct disorders can give rise to ILD. These conditions vary from extremely rare disorders, such as lymphangioleiomyomatosis, to multisystem diseases like systemic sclerosis (SSc) and rheumatoid arthritis, where ILD often develops as a manifestation of the disease. One of the more common forms of ILD is idiopathic pulmonary fibrosis (IPF) [14–17].

ILD is a growing global health issue, with an 86% increase in ILD-related mortality and predicted to account for 0.26% of all-cause mortality worldwide by 2027 [18, 19]. A recent study found that the prevalence of ILD ranged from 6.3 to 71 per 100,000 people, and the incidence

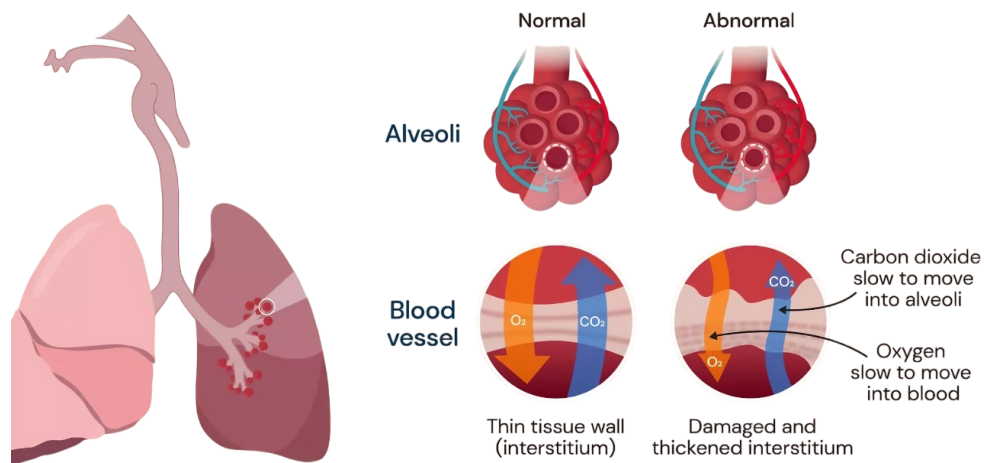


Figure 2.1: Comparison between healthy and abnormal lung tissue [13].

rate of ILD ranged from 1 to 31.5 per 100,000 people per year [2]. Another study from France reported sarcoidosis and IPF as the most commonly reported ILDs in Europe [20]. Despite this, ILD accounts for 15% of all cases seen by pulmonologists [21].

No universally accepted classification of ILD exists. Various schemes have been proposed based on different parameters, such as clinical presentation (for example, acute *versus* chronic), histopathologic findings, radiologic patterns and response to corticosteroid therapy. Perhaps the most practical classification of ILD for clinicians is a scheme based on cause, those with known *versus* unknown cause. Known causes include inhaled organic and inorganic substances, drugs, radiation, and systemic disorders such as connective tissue diseases. In contrast, some ILD have no identifiable cause despite the presence of well-recognized clinical and pathological syndromes, such as IPF and sarcoidosis [22].

The most common symptoms at the onset of the disease include dyspnoea, coughing, and fatigue. Some individuals may remain asymptomatic for several months or even years before receiving a diagnosis. During a physical examination, clinicians often identify a distinctive sound known as "velcro crackles," which is associated with this condition [23]. The clinical course and overall quality of life for individuals with ILD can be influenced by a variety of intrapulmonary and extrapulmonary comorbidities. These may include pulmonary hypertension, chronic obstructive pulmonary disease, lung cancer, gastroesophageal reflux disease, obstructive sleep apnoea syndrome, and cardiovascular disease [24].

2.2 ILD Diagnosis

Early detection and accurate diagnosis of these disorders are crucial for effective care and improved patient outcomes [11]. Diagnosing ILD requires a multidimensional approach that combines clinical, radiological, physiological, and, in some cases, histological findings. A comprehensive clinical history and a thorough physical examination are essential initial steps in confirming

the presence of ILD and guiding the diagnostic process [1].

The initial indication of ILD is often identified through computed tomography (CT), which typically shows diffuse parenchymal infiltrates. In some cases, abnormal PFTs, such as reduced lung volumes, decreased diffusing capacity, or oxygen desaturation during exercise, may serve as the first signs of the disease. It is important to note that about 10% of patients with ILD may have a normal CT scan, especially in the early stages of the disease. Therefore, a normal CT should not discourage clinicians from considering ILD as a possible diagnosis when the clinical context suggests it [22].

2.2.1 High-Resolution Computed Tomography

The diagnosis and classification of ILD have been revolutionized by HRCT, often obviating the need for more invasive procedures, such as surgical lung biopsy [1, 12]. HRCT is indicated in several key scenarios: to confirm the presence of lung disease in patients with clinical suspicion but normal CT, to provide a more precise characterization of previously detected abnormalities, identifying morphological patterns, to evaluate disease activity and potential treatment options, to guide the selection of the most appropriate anatomical site and procedure for biopsy and to monitor disease progression over time [25]. Unlike standard CT, HRCT is designed to enhance spatial resolution, improving the detection of small structures and subtle abnormalities [22, 26]. A comparative example of a standard CT and an HRCT slice is shown in Figure 2.2. In over two-thirds of cases, interpretation of HRCT patterns alongside clinical findings allows for a confident diagnosis [1].

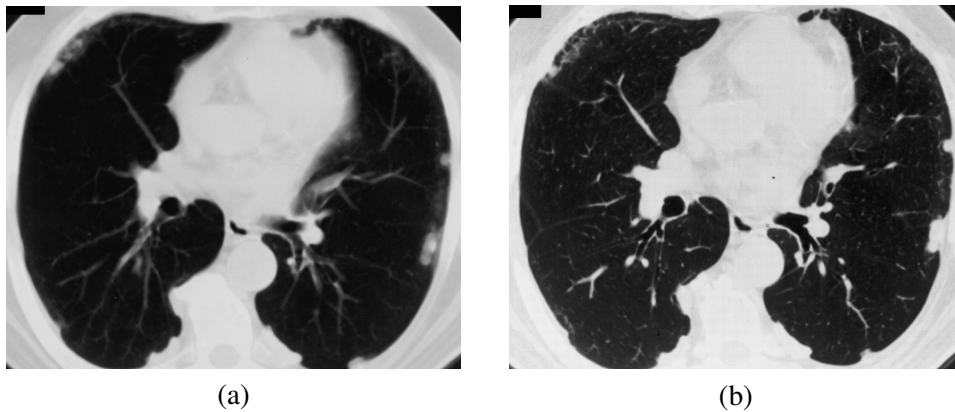


Figure 2.2: Comparison between a standard CT and HRCT of the lung parenchyma: (a) Standard CT; (b) HRCT [26].

From a technical point of view, the two consistent components of HRCT technique include the use of thin collimation, typically less than 2 mm thick, usually ranging from 1 to 1.5 mm, coupled with a high-spatial-frequency reconstruction algorithm [25, 26]. Thicker slices can hinder the assessment of small structures, while thinner slices, although more detailed, may increase image noise and the radiation dose to the patient [25]. High-spatial-frequency reconstruction enhances the definition of parenchymal images, rendering fine structures more sharply delineated [25]. To

ensure optimal image acquisition, HRCT scans are typically performed with the patient in a supine position during full inspiration [27].

Windowing parameters are essential for effectively visualizing and interpreting HRCT images. In CT imaging, pixel intensities are quantified in Hounsfield Units (HU), a standard scale representing tissue attenuation values [28]. A visual representation of the HU scale (Figure 2.3) further clarifies how grey-level values correspond to densities from air, -1000 HU, to bone, +1000 HU [29]. For evaluating lung parenchyma, the recommended settings usually include a window level between -600 and -700 HU, along with a window width ranging from 1000 to 1500 HU. These settings provide an optimal balance of detail and contrast. Wider windows, such as those set to 2000 HU, can enhance the visualization of high-contrast interfaces, like peripheral pleural abnormalities. Conversely, narrower settings are better for detecting subtle attenuation differences, such as areas with air trapping. When assessing the mediastinum, a window level of approximately 40-50 HU and a window width of 350-450 HU are typically used [25, 30].

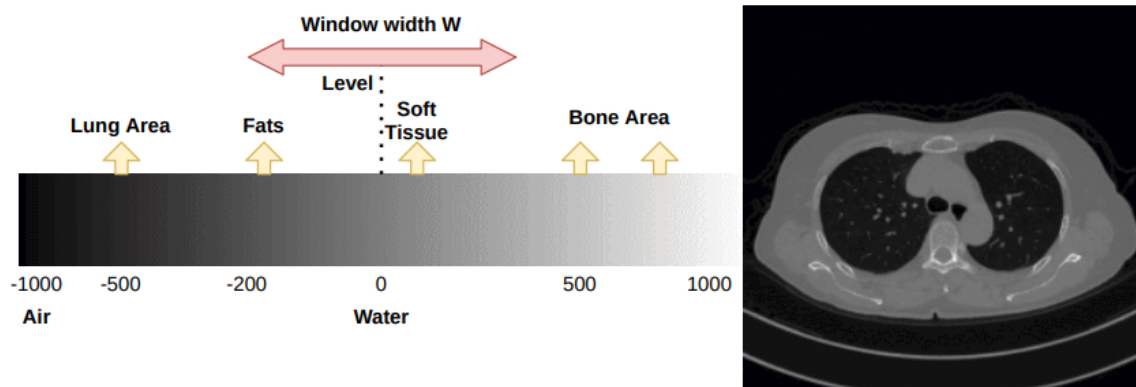


Figure 2.3: HU scale, mapping values to the represented tissue, and an axial slice of a CT scan showing the inside of the lung [29].

2.3 Radiological Findings in ILD

The diagnostic value of HRCT in ILD lies in the ability to recognize characteristic patterns of parenchymal abnormality. These radiological patterns, together with their distribution within the lungs and associated findings, provide essential clues for narrowing differential diagnoses and distinguishing between ILD subtypes [22]. In clinical practice, diagnosis often depends not on a single abnormality, but on specific combinations of HRCT findings that collectively define distinct disease patterns [31]. The subsequent analysis focuses on the principal HRCT findings in ILD and their corresponding pathological features.

Among the various HRCT findings, those indicative of pulmonary fibrosis are characterized by irreversible reticulation, traction bronchiectasis, honeycombing, architectural distortion, and volume loss [32]. Representative HRCT examples of these key fibrotic findings are illustrated in Figure 2.4.

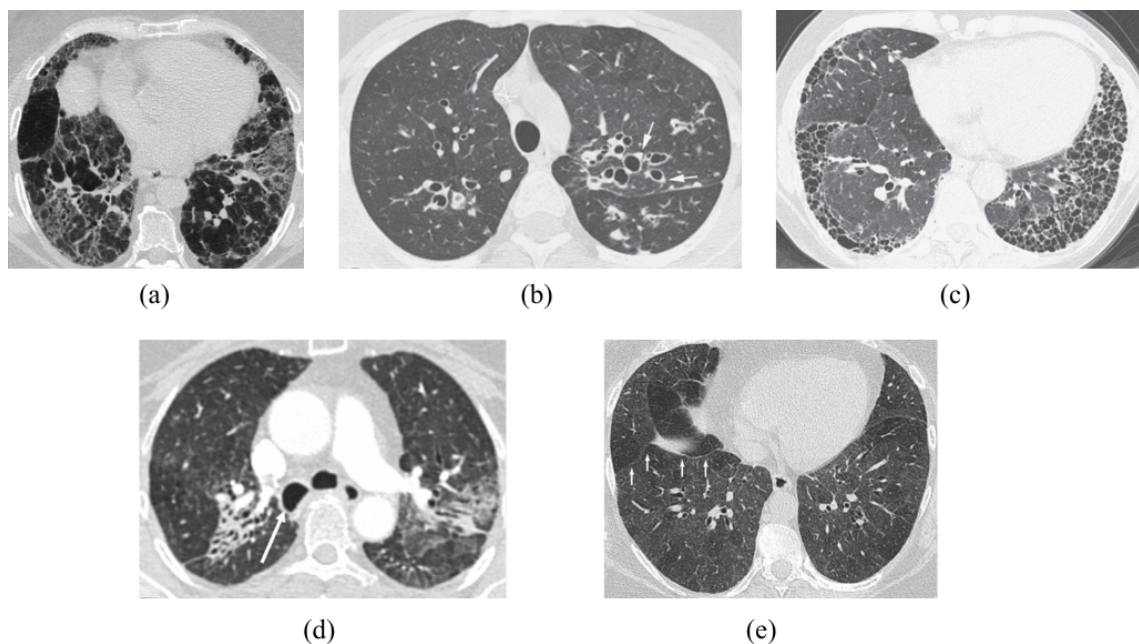


Figure 2.4: Representative HRCT findings characteristic of pulmonary fibrosis: (a) reticulation [30]; (b) traction bronchiectasis [31]; (c) honeycombing [31]; (d) architectural distortion [32]; and (e) volume loss [33].

For clarity, each of these key HRCT findings is described individually as follows:

- **Reticulation** – An early HRCT feature of fibrosis, which appears as a network of fine or coarse linear opacities forming an interlacing “mesh” or “spider web” morphology [25, 32]. These irregular intersecting lines represent both interlobular septal thickening and intralobular interstitial thickening [34].
- **Traction bronchiectasis and bronchiolectasis** – Commonly observed in areas of pulmonary fibrosis with a reticular pattern, they represent irreversible dilatation of the bronchi and bronchioles, respectively, caused by fibrotic retraction of the surrounding lung parenchyma. On HRCT, these airways typically have an irregular varicose appearance, reflecting the distortion induced by the fibrosis [31, 35].
- **Honeycombing** – Represents the end stage of pulmonary fibrosis and is frequently associated with traction bronchiectasis, as both arise from progressive parenchymal destruction and fibrotic retraction of the lung tissue. Pathologically, it is defined by the presence of small air-containing cystic spaces, generally lined by bronchiolar epithelium and having thickened walls composed of dense fibrous tissue. On HRCT, these cystic spaces typically range from 3 mm to 1 cm in diameter, although they can be as large as several centimetres in diameter [31].
- **Architectural distortion** – Refers to the displacement and irregular configuration of bronchi, pulmonary vessels, and fissures caused by fibrotic remodelling of the lung parenchyma [32, 36].

- **Volume loss** – Although non-specific, it can serve as a useful indicator of fibrotic lung disease, particularly when honeycombing is absent or traction bronchiectasis is subtle. It is often inferred from shifts in the position of the oblique fissures on axial HRCT images: in healthy lungs, the oblique fissure extends posteriorly at the level of the aortic arch and nearly reaches the anterior chest wall at the diaphragm, while in fibrotic lungs, volume loss in the lower or upper lobes alters this position, making the fissure appear displaced [33].

In addition to fibrotic findings, several other parenchymal abnormalities are commonly encountered in ILD. These include nodules, ground-glass opacity, consolidation, cysts, emphysema, mosaic attenuation and crazy paving. Representative HRCT examples of these findings are illustrated in Figure 2.5.

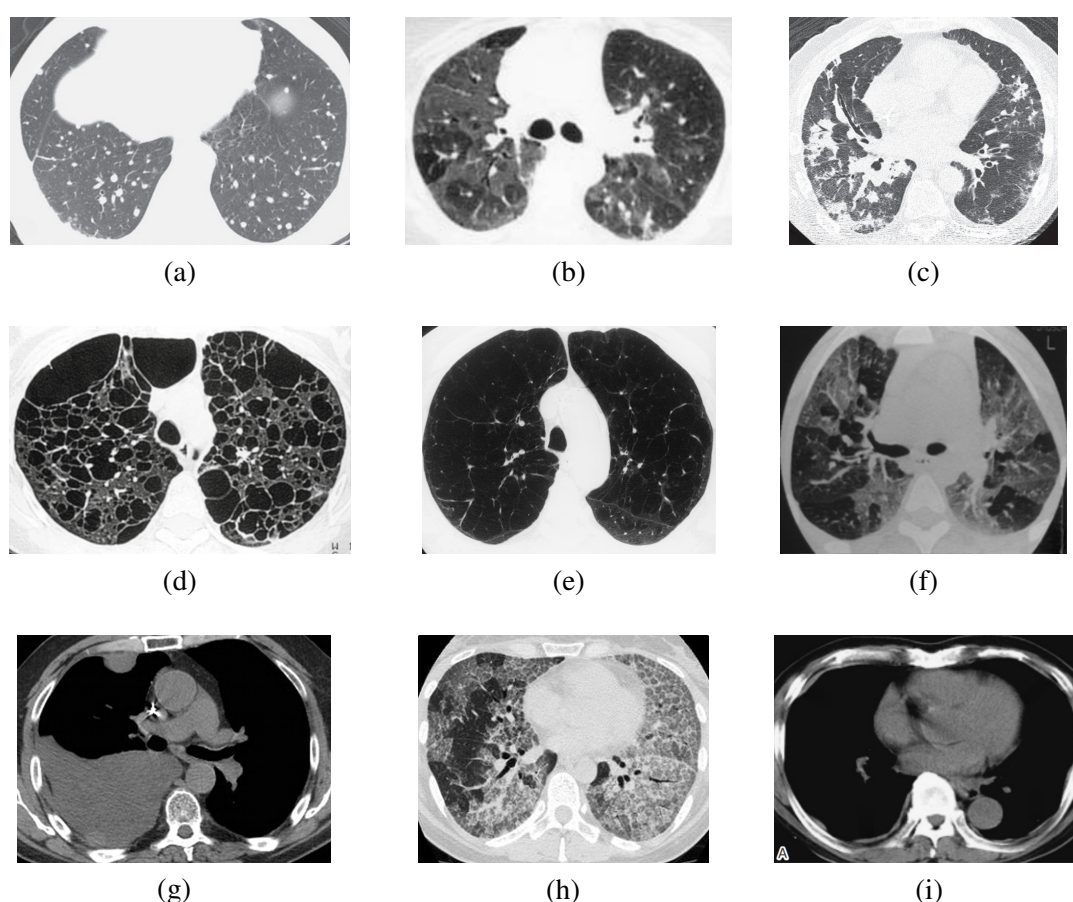


Figure 2.5: Representative HRCT findings commonly observed in ILD: (a) nodules [31]; (b) ground-glass opacity [27]; (c) consolidation [31]; (d) cysts [30]; (e) emphysema [26]; (f) Mosaic attenuation [34]; (g) Pleural effusion [37]; (h) Crazy paving [34] and (i) Pleural plaques [38].

These findings can be detailed as follows:

- **Nodules** – Defined as round or ovoid, discrete parenchymal opacities within the lung. They are typically characterized according to their size, attenuation, margin definition, and localisation. Based on size, they are generally classified as micronodules (<3 mm in diameter),

nodules (3–30 mm), and masses (>30 mm) [30]. In general, the anatomical location and distribution of nodules play a crucial role in the differential diagnosis. When assessed in relation to the anatomy of the secondary pulmonary lobule, nodules are typically classified into three major distribution categories:

- **Perilymphatic or Lymphatic distribution** – Several diseases are characteristically associated with nodules that occur predominantly in relation to pulmonary lymphatics. These diseases are termed perilymphatic or lymphatic. In patients with a perilymphatic distribution, nodules predominate, both histologically and on HRCT, involving the perihilar peribronchovascular interstitium, the interlobular septa, the subpleural regions, and the centrilobular peribronchovascular interstitium [31].
- **Random distribution** – Characterised by uniform involvement throughout the lungs, without preference for specific anatomical structures. The involvement tends to be bilateral and symmetrical. A nodular pattern with a random distribution in relation to the leading structures of the secondary lobule is indicative of a disease process with hematogenous spread as seen in miliary tuberculosis [30].
- **Centrilobular distribution** – Nodules are confined to the centrilobular regions and do not contact the fissures or pleura. They originate from the respiratory bronchioles, terminal bronchioles, or peripheral branches of the pulmonary artery. The nodules maintain a distance of several millimetres from the interlobular septa and fissures, and 5–10 mm from the pleural surface [30]. As indicated earlier, centrilobular nodules can be seen in patients having a perilymphatic or lymphatic distribution of disease. Pulmonary lymphatics are located in the peribronchovascular interstitial compartment in the centrilobular region. However, in patients with a perilymphatic distribution, nodules will also be seen in other locations (subpleural regions or interlobular septa) [31].

A centrilobular distribution of nodules may be associated with an important finding, termed tree-in-bud, which is of great value in differential diagnosis [39]. This finding typically reflects the presence of dilated centrilobular bronchioles filled with mucus, fluid, or pus, and is often associated with peribronchiolar inflammation. On HRCT, tree-in-bud is generally easy to recognize and may appear in various forms, either alone or combined. In the lung periphery, it often presents as small branching appearance, with the most peripheral nodules located several millimetres from the pleura, or as centrilobular clusters of nodules [31].

Representative HRCT examples of these nodular distributions are shown in Figure 2.6.

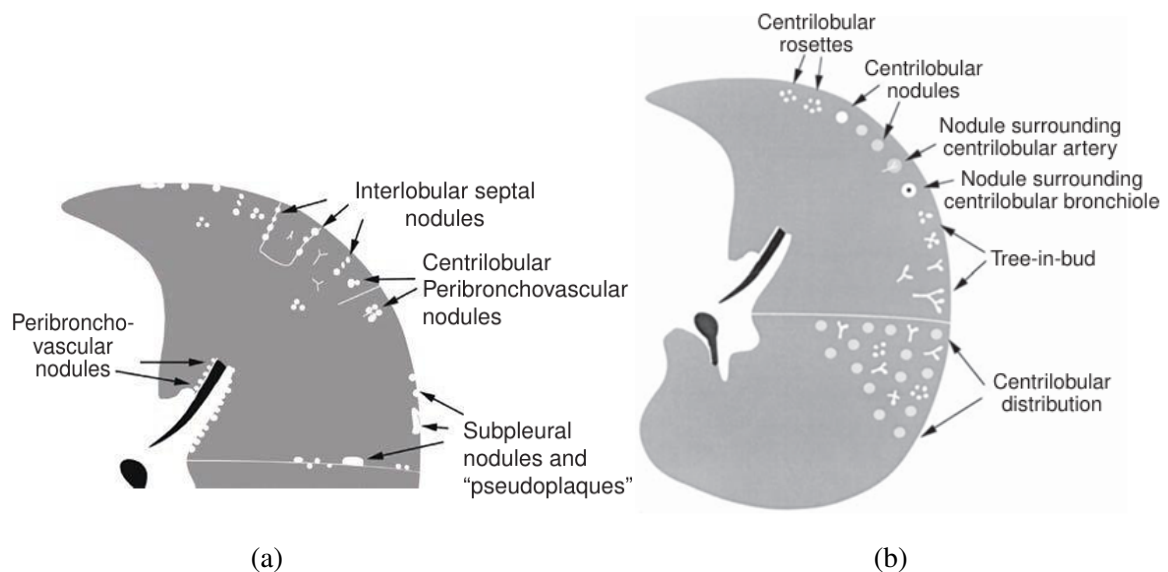


Figure 2.6: HRCT appearances of (a) perilymphatic nodules; (b) centrilobular nodules [31].

- **Ground-glass opacity** – Refers to a hazy increase in lung attenuation on HRCT that does not obscure the underlying bronchial and vascular margins. This finding usually reflects partial filling of the alveolar spaces or thickening of the interstitium [27].
- **Consolidation** – Observed when diseases replace alveolar air with fluid, cells, tissue, or other substances, leading to obscuration of underlying pulmonary vessels [31].
- **Cysts** – Used to refer to a well-defined and well-circumscribed rounded or irregular lesion, with a visible wall, that may be uniform or variable in thickness, but which is usually thin (less than 2–3 mm). A cyst usually contains air but may also contain liquid, semi-solid, or solid material [31].
- **Emphysema** – Defined as a permanent, abnormal enlargement of airspaces distal to the terminal bronchiole and accompanied by the destruction of the walls of the involved airspaces [31]. On HRCT, it appears as areas of abnormally low attenuation without definable walls, which distinguishes emphysema from the cystic lung diseases. The most common form, centrilobular emphysema, typically occurs as a result of cigarette smoking and is upper-lobe-predominant in distribution. Early centrilobular emphysema is characterized by round black holes that may appear in the central portion of the secondary pulmonary nodule around the centrilobular artery [26].
- **Mosaic attenuation** – Characterised by a patchwork of intermingled areas of increased and decreased attenuation [40].
- **Pleural effusion** – Defined as the abnormal accumulation of fluid in the pleural space and typically appears on imaging as a homogeneous opacity in the dependent portions of the thorax [41].

- **Crazy paving** – Consists of ground-glass opacity with a superimposed reticular pattern. In crazy paving areas, the ground-glass component may reflect airspace or interstitial abnormalities, while the reticular component may represent interlobular septal thickening, intralobular interstitial thickening or irregular areas of fibrosis [31, 42].
- **Pleural plaques** – A distinct type of pleural thickening, consisting of localised areas of thickening with shelf-like edges [41].
- **Pneumothorax** – Appears when air enters the pleural cavity. It can be broadly classified as traumatic/iatrogenic or spontaneous [43].

Given the wide range of HRCT findings that may occur in ILD, the interpretation of these images often presents considerable challenges. In some cases, diagnosis and classification can be straightforward, whereas in others, HRCT findings may overlap among different ILD subtypes, leading to ambiguity and inter-/intra-observer variation [3]. Moreover, examining all HRCT sections with consistent accuracy and efficiency is an exhausting and time-consuming task even for experienced radiologists [5]. These limitations underscore the potential value of tools capable of supporting radiologists in achieving more consistent and reliable assessments [44].

Chapter 3

Literature Review

This chapter reviews previous studies on ILD findings segmentation using DL. The following section expands the scope to semi-supervised methods in medical image segmentation. Finally, the chapter concludes by discussing potential future directions for ILD findings segmentation, highlighting emerging trends and avenues for further research.

3.1 ILD Findings Segmentation

In the current state-of-the-art research on ILD findings segmentation, relevant studies were identified through a structured search. The following subsection details the search methodology, as well as the criteria for inclusion and exclusion of the selected articles. Subsequent subsections provide an overview of the reviewed studies and present the methods employed, organized by learning strategy.

3.1.1 Search Methodology

The literature search was conducted following a structured and systematic approach to identify the most relevant studies addressing DL-based segmentation of ILD and related pulmonary findings. The search query was formulated as: (ILD OR “interstitial lung disease” OR “fibrotic lungs” OR “pulmonary fibrosis” OR “lung fibrosis”) AND segmentation AND (“deep learning” OR CNN OR “neural network” OR “machine learning” OR “artificial intelligence”).

An advanced search was performed, restricted to titles and abstracts, across five different databases: PubMed with 54 articles, IEEE Xplore with 18 articles, ScienceDirect with 22 articles, Web of Science with 69 articles, and Scopus with 203 articles, resulting in a total of 366 retrieved publications. To ensure the inclusion of recent and relevant research, only studies published from 2017 onwards and written in English were considered. Articles without full-text accessibility were excluded from the review process.

After removing 156 duplicates, the remaining articles were screened based on their titles and abstracts, according to the established inclusion and exclusion criteria. This preliminary screening led to the exclusion of 242 papers, leaving 28 studies for full-text evaluation. During the full-text

analysis, 13 additional articles were excluded for not meeting the defined criteria, resulting in a final selection of 15 studies.

Initially, the inclusion criteria focused on works addressing segmentation of various ILD related findings. However, due to the limited number of studies, the scope was subsequently broadened to include research focusing on segmentation of individual findings associated with IPF. This final set of selected papers provides the foundation for the state-of-the-art analysis of ILD findings segmentation.

3.1.2 Structuring and Overview of Previous Studies

The reviewed studies can be broadly categorized according to the segmentation type and the learning paradigm adopted, as illustrated in Figure 3.1. Regarding segmentation type, the papers can be divided into binary and multi-class segmentation approaches. In binary segmentation [45–48], models are typically designed either to distinguish between healthy and diseased lung tissue or to identify the presence or absence of a specific finding, most commonly fibrosis. In contrast, multi-class segmentation methods [5, 44, 49–57] aim to delineate multiple regions or categories. These approaches vary considerably among studies, encompassing both the differentiation of several pathological findings and the segmentation of distinct anatomical or structural regions. The number and type of segmented classes typically range from three to eight.

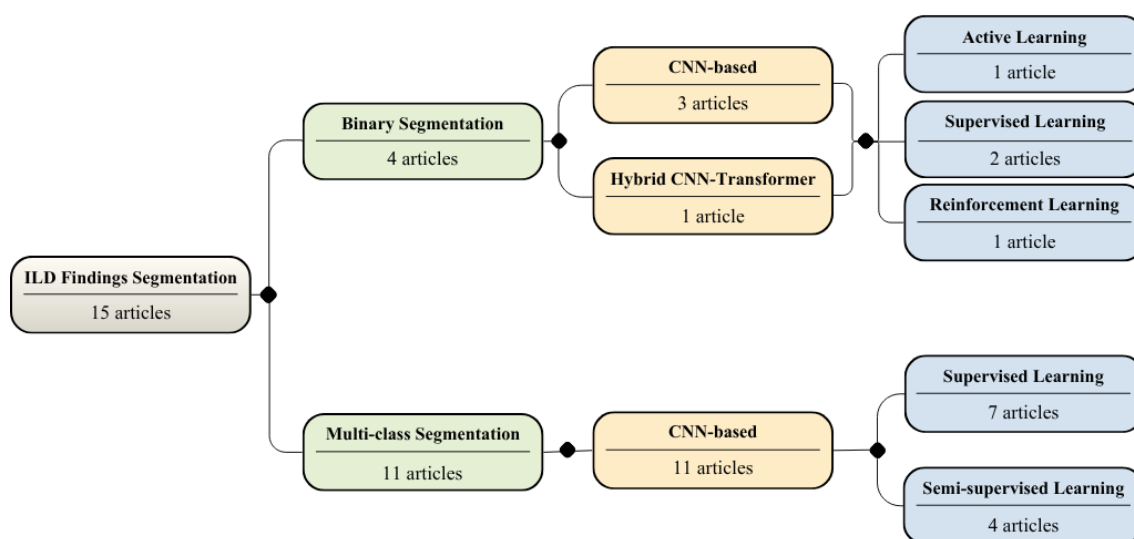


Figure 3.1: Schematic overview of the categorisation of the studies included in the literature review.

From a methodological perspective, the selected papers also differ in the type of learning strategy adopted. The majority employ SL approaches [5, 44, 45, 48–53], while a smaller number explore SSL strategies [54–57]. Additionally, a few studies implement reinforcement learning (RL) [47] or active learning (AL) [46] frameworks. Most of these studies relied on CNNs, with only one study incorporating a transformer-based solution in combination with CNN.

A summary of the reviewed literature on DL-based segmentation of ILD is presented in Table 3.1, including the authors, learning paradigms, segmentation tasks, targeted findings, network architectures, datasets, and quantity of data, as well as the reported Dice similarity coefficient (DSC) or, when unavailable, the balanced accuracy.

Following Table 3.1, it is worth briefly discussing the databases employed in the reviewed studies. Many of the articles relied on private datasets, which are not publicly available. Among the public resources, the MedGIFT ILD database [58] is frequently used. This database provides HRCT image series, including annotated regions of pathological lung tissue and lung parenchyma, along with clinical parameters from 113 patients. However, the annotations in this database are limited and incomplete, which often leads authors to supplement it with other databases to enhance performance. For instance, Hernández-Vázquez et al. [44] were the only authors to rely exclusively on the MedGIFT ILD database [58].

Less frequently, the reviewed studies also make use of other databases, namely the Lung Tissue Research Consortium (LTRC) [59] and the Inselspital ILD (INSEL) databases [56]. The LTRC database [59], developed by the National Heart, Lung, and Blood Institute (NHLBI), contains more than 100 cases with confirmed ILD diagnoses and provides annotations of the lung parenchyma, airways, and pathological tissue. Access to the database is granted upon request following approval from the LTRC data coordinating centre. Similarly, INSEL database offers two parts: INSEL-DB-Seg, with 60 HRCT lung scans with annotated tissue from two radiologists, and INSEL-DB-Diag, with 105 HRCT lung scans provided by the ILD board of Bern University Hospital, in which pneumologists, radiologists, and pathologists have diagnosed the patients according to international guidelines [35].

In the study by Lai et al. [47], two public external databases were utilized: the COVID-19 Collection [60] and Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) [61], although neither is specific to ILD.

Table 3.1: Summary of DL approaches for ILD findings segmentation. "Nl", "Fib", "Hc", "Mno", "Ro", "Bronch", "Ggo", "Cons", "Emph" and "Pne" represent normal lung, fibrosis, honeycombing, micronodules, reticular opacity, bronchiectasis, ground-glass opacity, consolidation, emphysema, pneumothorax, respectively. NR indicates not reported in the study and "B-Acc" indicates balanced accuracy.

Authors	Paradigm	Task	Findings	Architecture	Database	Samples	DSC (%)	B-acc (%)
Chassagnon et al. [45]	SL	Binary	Lesion	AtlasNet variant with six SegNet-based CNNs	Private	208 patients	74.50	NR
Vakalopoulou et al. [49]	SL	Multi-class	Lesion, Nl	Patch-based CNN followed by SegNet	Private, MedGIFT ILD [58]	80,000 patches and 3,950 slices	72.20	NR
Hernández-Vázquez et al. [44]	SL	Multi-class	Lung, Fib	U-Net	MedGIFT ILD [58]	120 patients	78.15	NR
Thillai et al. [48]	SL	Binary	Fib	3D U-Net	Private	446 patients	90.00	NR
Agarwala et al. [5]	SL	Multi-class	Fib, Con, Emph	P-Net	MedGIFT ILD [58], Private	480 images	NR	NR
Lai et al. [51]	SL	Multi-class	Hc, Ro, Ggo, Con, Emph	RDRNet	Private	410 patients	81.44	NR
Ham et al. [52]	SL	Multi-class	Nl, Hc, Ro, Ggo, Con, Emph	U-Net, nnU-Net	Private	912 patients and 92 images	36.28	NR

Continuation of Table 3.1

Authors	Paradigm	Task	Findings	Architecture	Database	Samples	DSC (%)	B-acc (%)
Fontanellaz et al. [53]	SL	Multi-class	Nl, Hc, Ro, Bron, Ggo, Emph	2D CNN–MLP Mixer	LTRC [59], INSEL [56]	338 cases	NR	73.21
Gao et al. [54]	SSL	Multi-class	Hc, Mno, Ggo, Emph	AlexNet CNN with fully connected CRF	MedGIFT ILD [58]	120 patients	NR	NR
Anthimopoulos et al. [55]	SSL	Multi-class	Hc, Ro, Mno, Ggo, Cons	FCN	MedGIFT ILD [58], INSEL [56]	172 scans	NR	81.80
Cai et al. [57]	SSL	Multi-class	Nl, Hc, Ro, Ggo, Con, Pne, Linear, Cyst	HRNet	Private	67 scans (11,000 images)	86.71	NR
Liu et al. [46]	AL	Binary	Fib	U-Net	Private	160 scans (1,600 images)	67.20	NR
Lai et al. [47]	RL	Binary	Lesion	Dual U-Net	Private, COVID-19 Collection [60], LIDC-IDRI [61]	506 patients (84,389 images)	94.90	NR

3.1.3 Supervised Learning Approaches

Initially, DL-based medical image segmentation predominantly employed SL methods, relying on large amounts of accurately annotated medical image data for training [62]. SL has yielded remarkable results in medical image processing, achieving performance levels comparable to those of human experts [63]. This paradigm requires the availability of a ground truth (GT) dataset and prior knowledge about the expected output, with the primary goal of learning the relationship and underlying structure between the input data and the corresponding output to make accurate predictions [64].

Among the reviewed studies employing SL, different segmentation objectives were identified. Some studies focused on distinguishing between healthy and diseased lung tissue [45, 49], while others aimed specifically at identifying fibrotic regions [44, 48]. In contrast, several works aimed to simultaneously segment multiple pathological findings [5, 50–53].

The study conducted by Chassagnon et al. [45] proposed a DL-based framework for the automated segmentation and quantification of ILD extent in patients with SSc. The approach is based on a modified version of AtlasNet architecture, presented in Figure 3.2, which incorporates six different deep CNNs, each trained on a specific anatomical template, providing a form of natural data augmentation. Each network follows a SegNet-inspired encoder-decoder structure with two decoding branches: one dedicated to lung segmentation and the other to disease segmentation, ensuring spatial coherence between both anatomical regions. After training the six CNNs, AtlasNet combines their predictions through an ensemble strategy using a majority voting scheme. The framework enables the model to differentiate healthy from diseased lung tissue and was compared against a conventional U-Net architecture trained under identical conditions. In terms of performance, AtlasNet slightly outperformed U-Net ($\approx 3\%$ improvement) and showed strong agreement with experts, with median DSCs comparable to inter-radiologist agreement. Finally, to assess its clinical relevance, the study evaluated the correlation between the disease extent and PFT measurements.

In a related work, Vakalopoulou et al. [49] also investigated patients with SSc, proposing a segmentation framework based on the combination of a patch-based CNN and a fully convolutional encoder-decoder (CED) network implemented using the SegNet architecture. The patch-based CNN is trained on sparsely annotated regions to capture local textural patterns of ILD, producing probabilistic maps for each class. These probability maps are subsequently used as additional input channels to the CED, which is pre-trained on fully annotated slices to perform pixel-wise dense segmentation. The framework, illustrated in Figure 3.3, integrates two complementary datasets: one containing CT images from patients with scleroderma disease, including both healthy and diseased lung tissue, and another composed of ILD-specific slices. The model differentiates between three regions: healthy lung, lung with disease, and background. Since the objective was to segment ILD tissue from scleroderma patients, all indicated disease abnormalities were merged into a single pathological class. Overall, the combined architecture produced noticeably smoother and

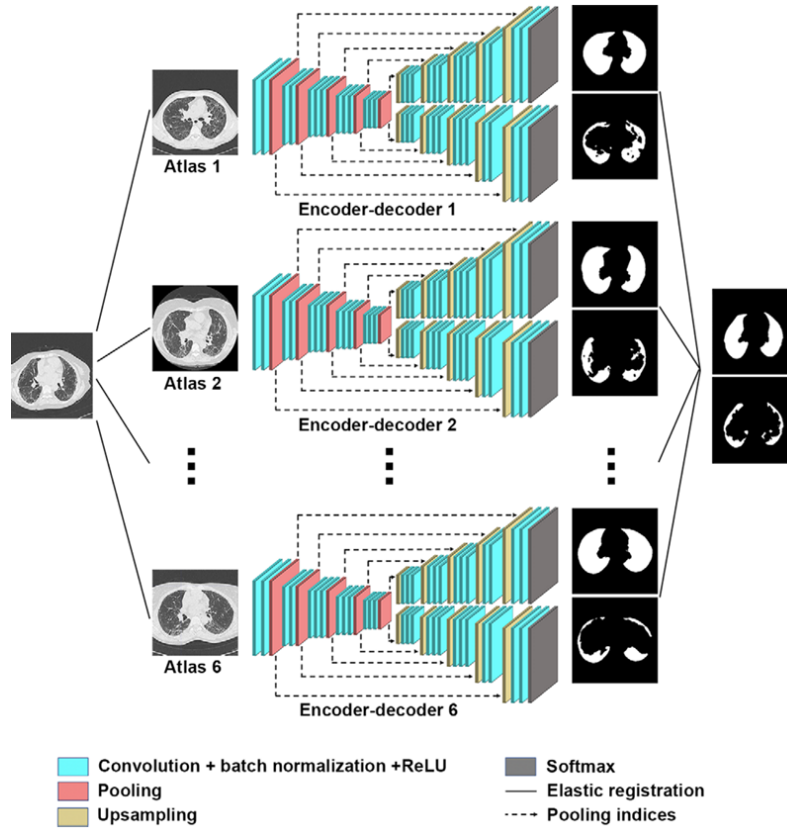


Figure 3.2: Architecture of the AtlasNet framework [45].

more spatially coherent segmentations, improving boundary definition compared to either component model alone.

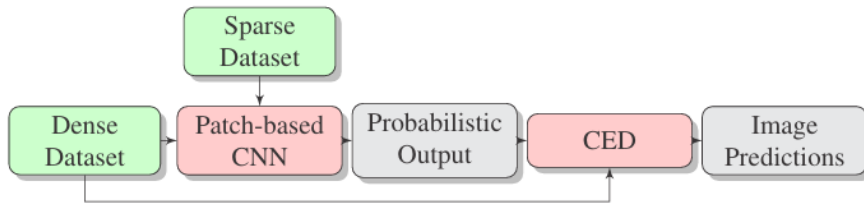


Figure 3.3: The proposed combination model with the two training datasets [49].

In a later study, Hernández-Vázquez et al. [44] introduced a multi-class semantic segmentation framework to identify fibrosis-related findings (FRF). Masks for the lungs and FRF regions were constructed to generate three-class label maps distinguishing background, lung tissue, and fibrotic areas, using data from the MedGIFT ILD database [58]. A U-Net with a VGG-16 backbone was employed for segmentation, supported by data augmentation to improve generalization. Model evaluation was conducted over the entire lung field, assessing both lung segmentation (lungEv) and fibrotic region segmentation (FRFEv) by directly comparing model outputs with the available

GT masks. Due to incomplete and inconsistent annotation of fibrotic regions in the dataset, where areas of fibrosis are frequently labelled as normal lung, the reported performance for FRF segmentation was relatively low, with a mean DSC of 38.80%. To mitigate the impact of missing annotations, the authors additionally proposed an ROI-delimited evaluation (R-FRFEv), where only the predicted fibrotic regions overlapping annotated GT fibrosis regions were considered during evaluation. Using this strategy, the segmentation performance increased substantially, reaching a mean DSC of 78.15%. These results highlight the strong influence of annotation quality on fibrosis segmentation performance and the challenges associated with evaluating partially annotated datasets.

Building upon this line of research, Thillai et al. [48] proposed a DL framework for segmentation of fibrotic regions, airways, and blood vessels in patients with IPF. Fibrotic regions and airways were segmented using supervised three-dimensional (3D) U-Net architectures with encoder-decoder pathways and skip connections. The lung parenchyma was segmented using an open-source model [65], while the intrapulmonary vessels were extracted by applying the Frangi enhancement filter [66]. This combination of methods allowed the extraction of quantitative volumetric biomarkers, which demonstrated strong correlations with lung function measures and prognostic outcomes, highlighting their potential to monitor disease progression.

In contrast with the studies discussed above, Agarwala et al. [5] proposed a multi-class segmentation framework focused on identifying three common ILD findings (emphysema, consolidation, and fibrosis) at a slice level. The method is based on P-Net [67], a fully convolutional architecture that replaces max-pooling and down-sampling with dilated convolutions to extract multi-scale contextual features, preserving spatial resolution of feature maps and enlarging the receptive field. To overcome the scarcity of annotated medical images, the authors adopted a two-stage training strategy, represented in Figure 3.4: first, a pre-trained model was developed using the large-scale PASCAL VOC natural-image dataset, selecting three abundant classes to initialise the network weights and, subsequently, transfer learning was applied to fine-tune all layers of the model using annotated HRCT sections from the MedGIFT ILD database [58] and a smaller private dataset. In general, the framework demonstrated robust segmentation performance for fibrosis and emphysema, while consolidation proved more challenging, often being confused with pleural regions due to contrast similarity.

Expanding on multi-class segmentation approaches, Lai et al. [51] proposed a DL framework, named RDRNet, designed to identify five ILD lesion types, quantify the extracted lesion regions, and subsequently employ an assessment model to predict ILD severity. It is worth noting that the framework proposed by Lai et al. [51] is aligned with previous work by Hou et al. [50], whose study employed the same RDRNet architecture for the same purpose. The network, illustrated in Figure 3.5, employs a modified U-Net architecture with a ResNet34 backbone, incorporating shortcut connections to accelerate convergence and mitigate vanishing gradients. In addition, Dense Atrous Convolution (DAC) blocks and Residual Multi-kernel Pooling (RMP) modules were integrated into the encoder-decoder architecture. DAC blocks enhance multi-scale feature extraction by cascading atrous convolutions with progressively larger dilation rates, while RMP modules

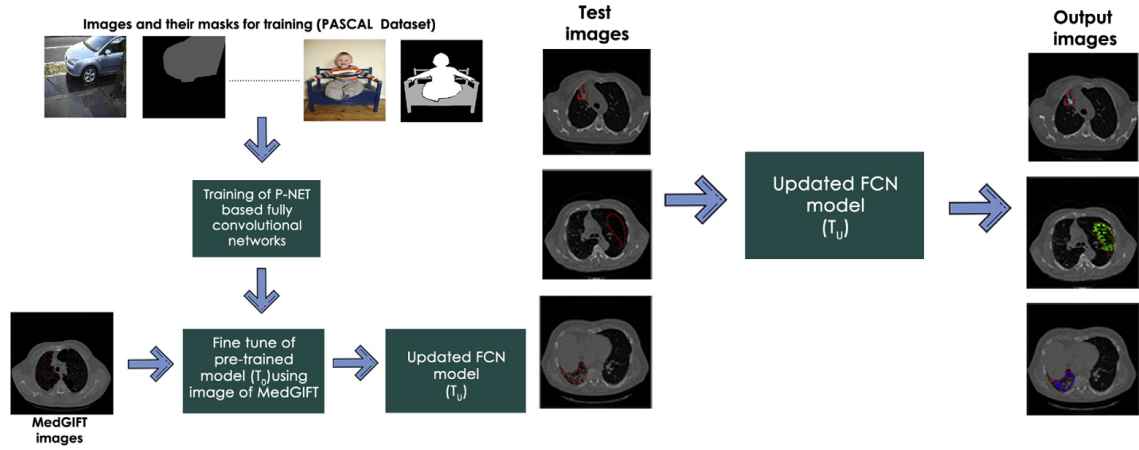


Figure 3.4: Two stage training strategy: transfer learning of the pre-trained model and screening of ILD findings [5].

aggregate global contextual information through parallel pooling operations with multiple kernel sizes. Using these quantitative lesion features, the authors demonstrated strong predictive performance for ILD severity, with key indicators such as honeycombing, reticular opacities, and fibrosis reliably distinguishing between disease stages.

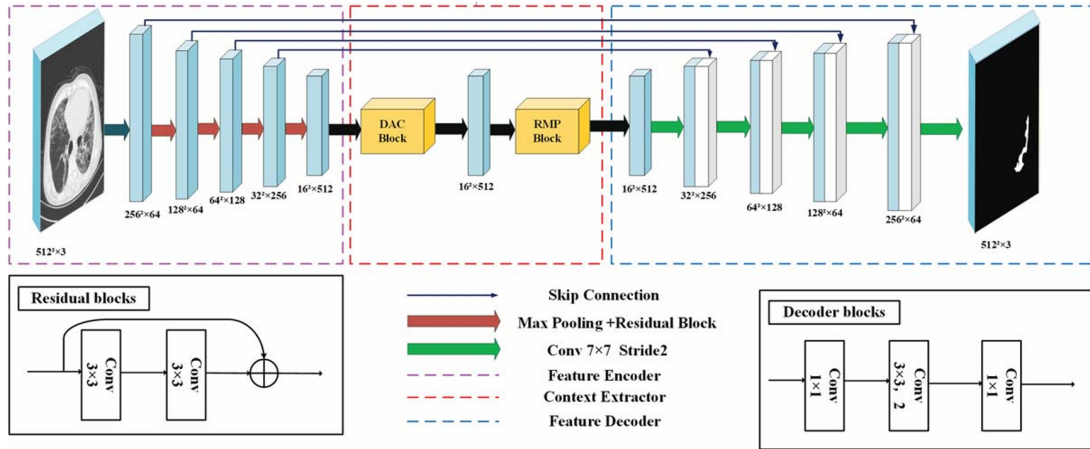


Figure 3.5: The overall network structure with DAC and RMP [51].

In another study, Ham et al. [52] proposed a framework designed to identify six radiological findings, and the overall pipeline is summarized in Figure 3.6. Expert-annotated patch-level regions of interest (ROIs) were first used to train a support vector machine (SVM), which was then applied to whole-lung HRCT scans to produce preliminary pixel-level segmentation maps. These SVM-derived labels subsequently served as supervision for training a U-Net and a no-new-U-Net (nnU-Net) segmentation network, later fine-tuned with manually delineated slice-level labels to reduce class imbalance and correct biases of the SVM classifier. In summary, fine-tuning markedly improved segmentation accuracy, reaching performance comparable to inter-radiologist agreement. External validation demonstrated that nnU-Net generally outperformed U-Net, par-

ticularly for normal lung parenchyma, while emphysema remained the most challenging finding. Nonetheless, the authors noted that the framework remains constrained by its reliance on expert manual annotations, which limits its scalability.

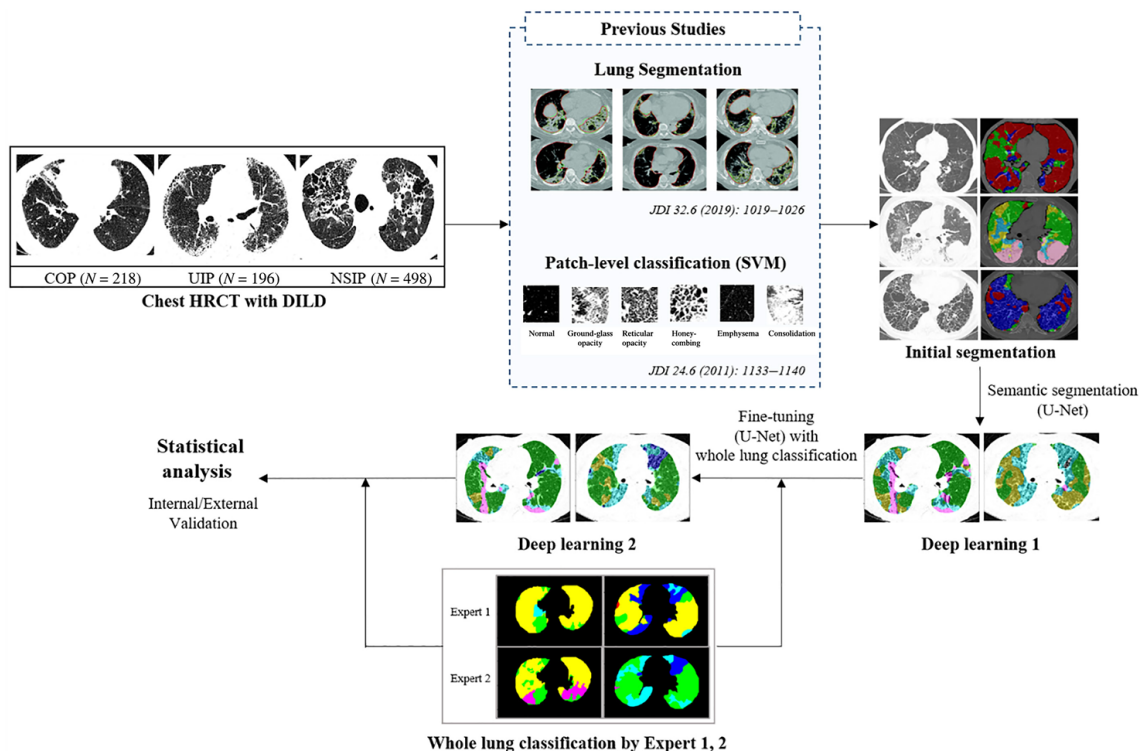


Figure 3.6: Overall workflow of semantic segmentation [52].

Following these advances in ILD analysis, Fontanellaz et al. [53] proposed a comprehensive computer-aided diagnosis (CAD) framework integrating segmentation of seven different ILD findings, quantification, radiomic feature extraction, and diagnosis. For the ILD findings segmentation component, the authors evaluated three architectures: two custom U-Net baselines (a two-dimensional (2D) U-Net and a 3D-U-Net), applied respectively to axial CT slices and 3D patches, and a 2D CNN multi-layer perceptron mixer (MLP-Mixer) hybrid model, designed to mitigate limitations associated with CNNs regarding receptive field and contextual representation. The performance of these models was compared against each other as well as against the 3D full-resolution nnU-Net. Overall, the nnU-Net and the hybrid model achieved similar performance, while the 2D- and 3D-U-Net showed lower balanced accuracy. For diffused ILD patterns, 2D segmentation performed comparably to 3D segmentation methods. In general, the MLP-Mixer architecture efficiently merges low-level texture and high-level global context features, however nnU-Net remained more robust than all three custom models.

3.1.4 Semi-Supervised Learning Approaches

With the development of DL technologies, SL based segmentation algorithms have made significant progress in medical image segmentation. However, since medical image annotation relies on

the knowledge and expertise of specialized doctors, the annotation process is both time-consuming and costly, resulting in a limited amount of labelled data available for training DL models [8, 9]. Data scarcity and high annotation costs have become major bottlenecks restricting the widespread application of SL methods in the field of medical image segmentation [68]. As a result, SSL methods have emerged as a prominent and highly promising research direction [69]. The goal of SSL is to leverage large amounts of unlabelled data, together with a small labelled subset, to improve the performance and generalization of supervised models [70].

Unlike supervised approaches, which vary between binary and multi-class segmentation, the reviewed semi-supervised studies consistently focused on multi-class ILD segmentation.

Employing a semi-supervised strategy, Gao et al. [54] proposed a label-propagation framework that expands sparse annotations from the MedGIFT ILD [58] dataset into dense multi-class ILD segmentations. Their approach, represented in Figure 3.7, treats segmentation as a maximum a posteriori (MAP) inference problem in a fully connected conditional random field (CRF). The unary terms come from a CNN patch-based classifier, while the pairwise terms connect every lung pixel to every other pixel, allowing long-range spatial consistency through efficient Gaussian message passing. The manually drawn annotations are used as hard constraints so they cannot be overridden, and connections between lung and non-lung regions are suppressed to avoid spreading labels into the wrong areas. With these constraints, the CRF is able to propagate the limited radiologist-drawn labels throughout the entire lung, producing coherent pixel-level ILD maps, which were then validated by an expert radiologist. Although the authors also introduced a secondary multi-label regression model for slice-level ILD detection, their main methodological contribution in the SSL context lies in the CRF-based label-propagation pipeline.

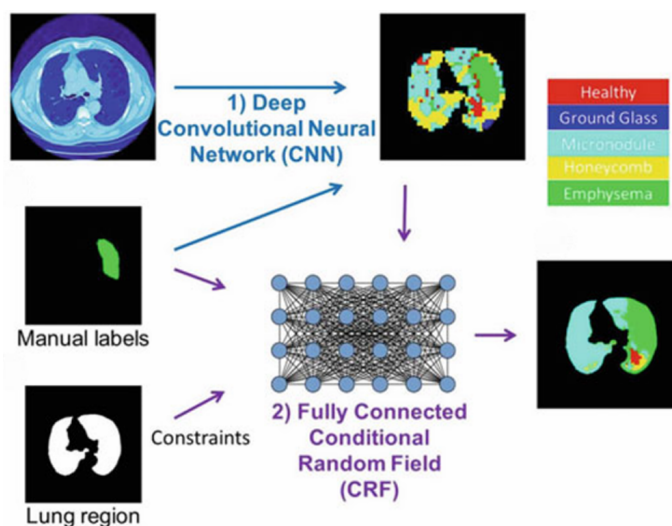


Figure 3.7: Overview of the segmentation propagation algorithm [54].

In another study, Anthimopoulos et al. [55] proposed a deep fully-convolutional network for ILD findings segmentation, employing dilated convolutions and an end-to-end semi-supervised training strategy. The network is composed exclusively by convolutional layers that use dilated

kernels, arranged to progressively increase the receptive field while preserving spatial resolution, with dilation rates selected according to the Fibonacci sequence to mitigate gridding artifacts. Feature maps from early layers are concatenated with the input and passed through dropout to aggregate multi-scale information and facilitate gradient flow, followed by additional convolutional layers that progressively reduce the feature dimensionality to the six output classes. Instance normalization layers are applied after each convolution, combined with skip connections to enhance feature contrast and improve invariance to intensity and contrast variations across different CT scanners. To address the sparsity and imbalance of the available annotations, the authors trained the network using a weighted categorical cross-entropy loss, where pixel weights were set inversely proportional to the number of samples of each class so that all categories contributed equally. In addition, an entropy minimization term was incorporated to exploit unlabelled image regions by penalising high-entropy predictions. This encourages the model to make confident decisions in non-annotated areas, effectively pushing decision boundaries away from densely populated clusters of labelled samples. The influence of the unsupervised term is controlled through a weighting factor that scales with the proportion of unlabelled pixels. The semi-supervised approach yielded an improvement of approximately 1.5%, although the method still showed difficulties in distinguishing similar textural patterns, such as reticulation versus honeycombing, and occasionally confused healthy tissue with reticulation. The same CNN design and training strategy were subsequently adopted by Christe et al. [56] for lung tissue characterisation, extending the approach to diagnostic classification of IPF cases by aggregating tissue distributions and clinical features.

In a further development, Cai et al. [57] introduced ESsegILD, an end-to-end SSL framework for the segmentation of ILD findings from CT images via self-training with selective re-training, as illustrated in Figure 3.8. The method builds upon a Mean Teacher (MT) architecture in which a student model is first trained in a fully supervised manner using the available labelled slices. A teacher model is then initialized as an exponential moving average (EMA) copy of the student to ensure stable guidance during self-training. At each iteration, the teacher generates pseudo-labels for the unlabelled images and assigns them a stability score. Only pseudo-labels with high stability are retained and added to the labelled set, following the selective re-training strategy. Next, the student is re-trained using both the GT annotations and the selected high-confidence pseudo-labels, optimized with supervised losses, dice and focal loss, together with a consistency loss that enforces agreement between student and teacher predictions under perturbations. The process (pseudo-label generation, stability-based filtering, and student re-training) is repeated iteratively until convergence. Both the student and teacher networks adopt high-resolution network (HRNet) with parallel multi-resolution representations to better capture the discriminative features of ILD findings. A total of 11,493 CT slices were used, including 2,132 fully annotated images. The results show that HRNet outperformed U-Net and dilated CNN. Adding unlabelled data further improved performance, yielding an average DSC gain of approximately 2.36 points. However, the method may still include low-quality pseudo-labels, which can constrain the model's maximum potential.

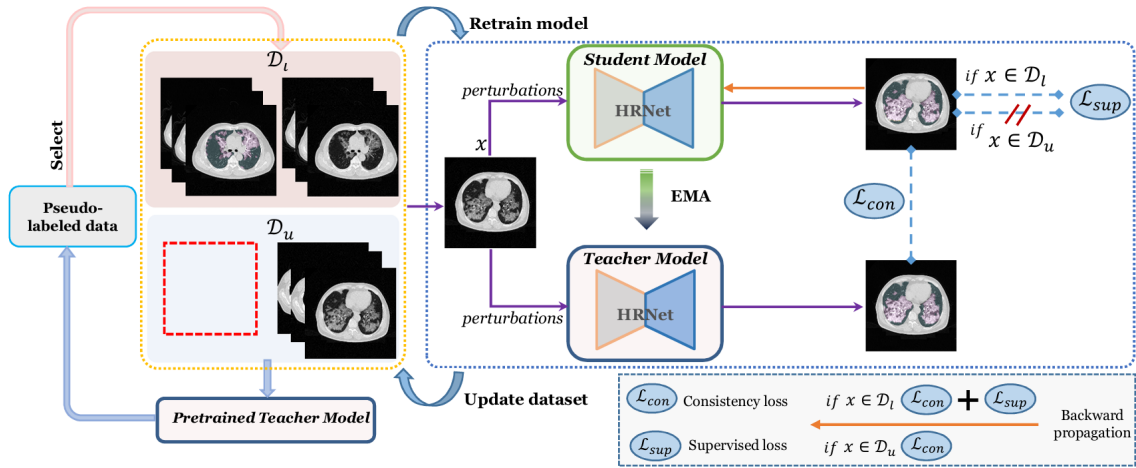


Figure 3.8: Proposed ESsegILD framework for ILD segmentation. It comprises MT training with labelled (D_l) and unlabelled (D_u) data, pseudo-label selection, and iterative retraining [57].

3.1.5 Active and Reinforcement Learning Approaches

In addition to supervised and semi-supervised strategies, AL and RL have emerged as alternative paradigms for ILD segmentation. These approaches address limitations inherent to conventional supervised pipelines by introducing mechanisms that reduce annotation burden or enable models to learn through interaction rather than static supervision [71–73].

AL is an approach that aims to maximise model performance while minimising the number of samples that require manual annotation. Instead of labelling an entire dataset, the model iteratively selects the most informative examples from a large pool of unlabelled data and requests annotations from the oracle (human annotator) only for those cases. The central premise is that a well-chosen subset can achieve performance comparable to that of a fully annotated dataset, substantially reducing the annotation burden [74, 75].

In this context, Liu et al. [46] introduced an AL framework focused on the binary segmentation of fibrotic tissue. The method is structured around three main components: the architecture of the segmentation network, uncertainty estimation in the image segmentation task, and the acquisition strategy. The segmentation network, termed U-Net-RGD and shown in Figure 3.9, is based on a customised U-Net and incorporates a ResNet-34 encoder with residual blocks. In the decoding path, channel-wise squeeze–excitation layers and dropout layers were applied to recalibrate feature responses and enable Bayesian approximation of uncertainty through Monte Carlo Dropout. Based on these uncertainty estimates, the DeepRISS acquisition strategy performs uncertainty-weighted clustering on feature embeddings extracted from the encoder, selecting the most informative and representative samples for annotation, thereby reducing redundancy and improving training efficiency in sequential CT slices. Evaluation on a private fibrosis segmentation dataset showed that U-Net-RGD outperformed the baseline U-Net model.

In contrast, RL is a feedback-driven learning paradigm in which an agent learns to make optimal decisions through trial-and-error interactions with an environment. Instead of relying on

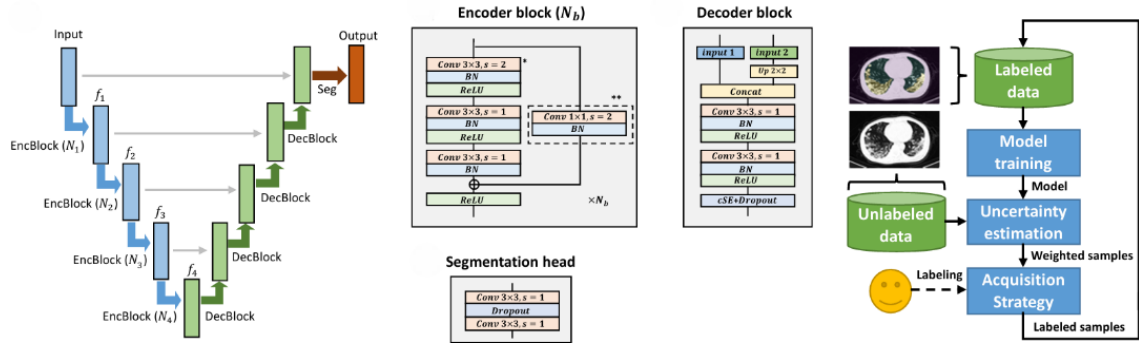


Figure 3.9: U-Net-RGD architecture and the active learning workflow [46].

fixed supervision, the agent observes the current state, performs an action, and receives a reward signal that indicates how beneficial that action was toward achieving a specific goal. Over time, the agent refines its policy to maximise the expected cumulative reward [72, 73, 76].

Building on this paradigm, Lai et al. [47] introduced an RL framework, termed CT Smoother Agent (CTSA), designed to perform weakly supervised ILD lesion segmentation by exploiting contextual information across sequential CT slices while relying solely on image-level labels. The method, illustrated in Figure 3.10, operates cyclically and integrates three key components: an initial mask generation module based on a vision transformer, ViT-MSIL, a deep reinforcement learning module, and a constraint segmentation stage implemented with a dual U-Net. First, ViT-MSIL is applied to produce coarse lesion masks by leveraging multiple-instance learning over superpixel candidates derived from Felzenszwalb [77] algorithm. These initial masks are then used to train the deep reinforcement learning (DRL) agent, which incorporates optical-flow detection to identify regions exhibiting substantial morphological or positional variation across consecutive CT slices. The agent acts on these abnormal regions by adjusting pixel intensities through predefined actions, guided by a reward function that encourages more anatomically consistent lesion shapes. Subsequently, the optimized masks produced by the DRL module are combined with the initial masks and fed into a dual U-Net, which constrains training using both sources of supervision to generate enhanced segmentation outputs. This enhanced mask is then re-introduced into the DRL module, enabling a cyclic optimization process that progressively improves lesion delineation until convergence criteria are met. Experimental results show that CTSA improves segmentation accuracy, particularly in capturing subtle ILD lesions. Nonetheless, the method's generalization is constrained by its dependence on the quality of these initial masks, and it cannot be directly applied to datasets from different sources.

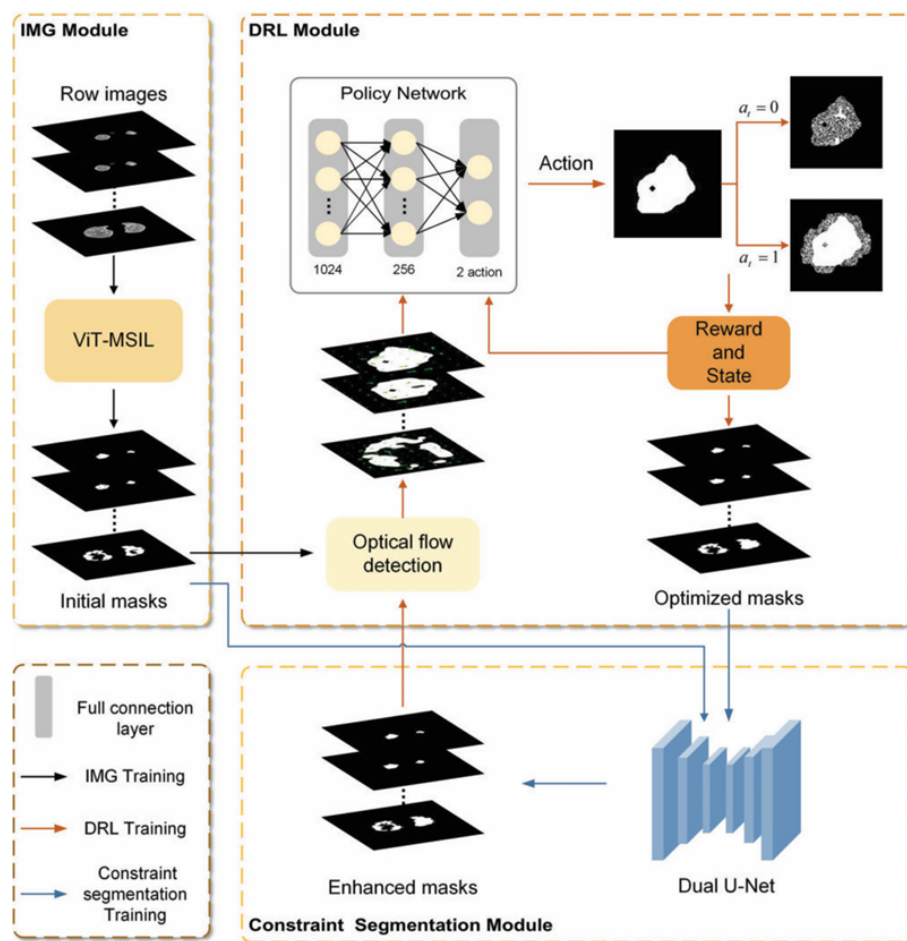


Figure 3.10: Overall pipeline of the proposed CTSA framework. The model is mainly divided into three parts: the IMG module, the DRL module and the constraint segmentation module [47].

3.2 Semi-Supervised Learning for Medical Image Segmentation

This section introduces representative semi-supervised methods applied to medical image segmentation. A substantial body of research has emerged in this area, with most methodologies focusing on two predominant paradigms: pseudo-labelling and consistency regularization. However, more recent strategies explore consistency regularization supervised by pseudo-labels and methods that combine contrastive learning with consistency regularization [78]. The following subsections discuss these four categories in detail.

To provide a clearer overview of their relationships, Figure 3.11 presents a Venn diagram that summarizes representative methods and their intersections across these methodological categories.

3.2.1 Pseudo-Labelling-Based Methods

Pseudo-labelling is one of the simplest and widely adopted strategies in semi-supervised segmentation [62, 79]. The core idea is to use the model's predictions on unlabelled data generating pseudo-labels, by applying thresholding or other selection mechanisms [69]. These pseudo-

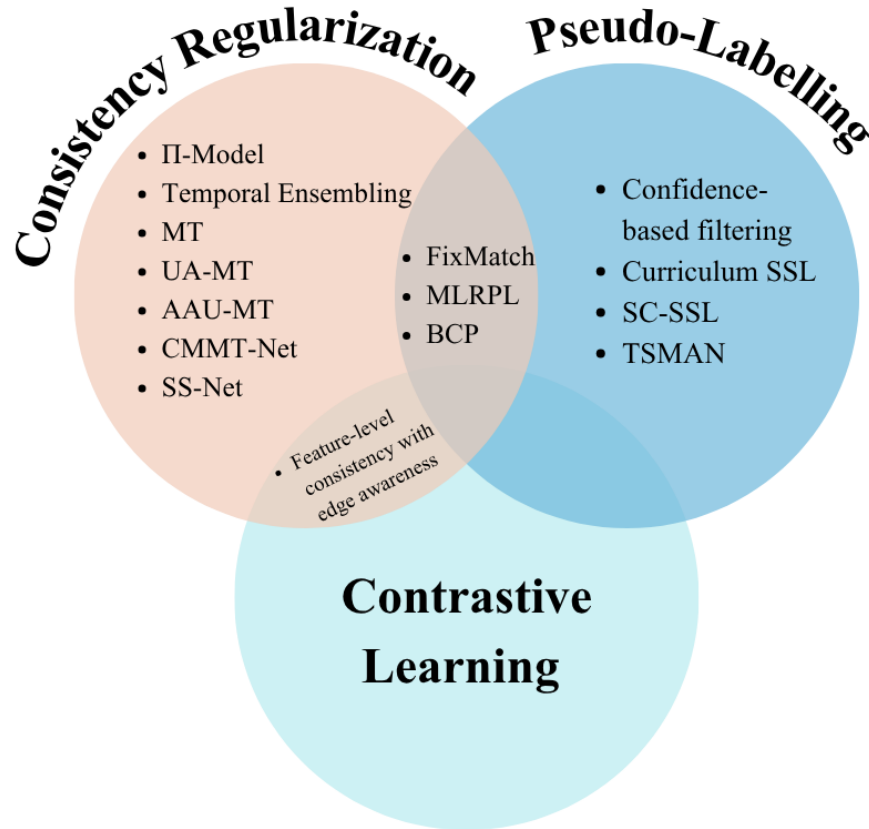


Figure 3.11: Venn diagram summarizing representative semi-supervised segmentation methods and their intersections across main methodological categories.

labelled samples are then combined with the manually annotated data to further train the model, usually in an iterative process that progressively improves the pseudo-mask quality [80]. The effectiveness of pseudo-labelling largely depends on ensuring that the generated pseudo-annotations are sufficiently reliable [62].

Early self-training methods [81] introduced confidence-based filtering, where each prediction is assigned a confidence score and only those exceeding a predefined threshold are retained as pseudo-labels. This strategy helps filter out low-confidence and potentially inaccurate predictions, but relying on a fixed threshold may still discard useful information. Although this approach allows the model to leverage unlabelled images, one of its main drawbacks is that errors in the initial pseudo-labels can be propagated and amplified during training, degrading segmentation performance [82]. Several techniques have been proposed to address this issue, such as co-training [83]. However, these approaches typically involve training multiple networks simultaneously, which can make model convergence slower or more difficult.

Kervadec et al. [84] approached the problem from a different perspective by introducing a curriculum SSL that regularizes the segmentation network using image-level regression tasks, such as estimating the target region size. This encourages predictions on unlabelled images to align with inferred label distributions. A distinct feature of this framework, illustrated in Figure 3.12, is

the use of inequality constraints, which provides greater flexibility by allowing uncertainty in the inferred knowledge, such as the regressed region size.

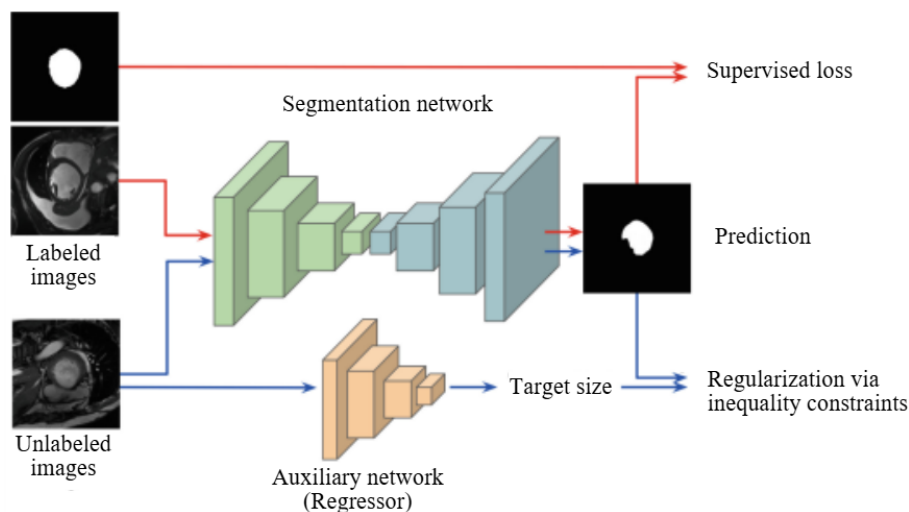


Figure 3.12: Illustration of the curriculum SSL strategy. Adapted from [84].

To overcome the limitations of noisy pseudo labels, some methods iteratively refine the quality of pseudo annotations and then reuse the improved labels to retrain the segmentation model, thereby reducing error propagation. For example, Miao et al. [85] introduced the self correction scheme (SC-SSL), which progressively adjusts the network’s targets toward distributions closer to the true labels, thus enhancing the exploitation of unlabelled data.

Along these lines, Min et al. [86] proposed a two-stream mutual attention network (TSMAN) designed to adaptively identify and mitigate errors in noisy pseudo-labels. The network exchanges multi-level features between the two streams and uses attention modules to highlight potentially incorrect labels, attenuating their influence during backpropagation. Despite its effectiveness, the approach remains sensitive to data augmentation strategies, and training two networks simultaneously increases computational cost and prolongs training time.

Overall, the use of pseudo-labelling in semi supervised medical image segmentation has evolved considerably, moving from early confidence based filtering strategies to more advanced frameworks such as dual network ensembles, curriculum approaches, and self-correcting mechanisms with attention modules. Nevertheless, current methods still depend heavily on the quality of the initial pseudo labels, show sensitivity to data augmentation and transformation choices, and often require substantial computational resources due to complex network designs.

3.2.2 Consistency Regularization-Based Methods

The fundamental principle of consistency regularization assumes that a model should yield stable predictions for unlabelled samples, even when those inputs undergo different perturbations or transformations. Enforcing consistency across augmented versions of the same data encourages

the network to learn more robust feature representations, less sensitive to input variations, thereby enhancing its generalization capability [62, 80].

Common implementations of this paradigm include the Π -Model and Temporal Ensembling. Sajjadi et al. [87] proposed the Π -Model, which applies different stochastic perturbations, such as data augmentation or dropout, to the same input sample and enforces the network to produce consistent predictions across multiple forward passes. Another influential strategy is Temporal Ensembling, proposed by Laine et al. [88], where consistency targets are obtained by maintaining an EMA of the model's predictions over successive training iterations. For each unlabelled sample, the EMA aggregates historical predictions, resulting in more stable targets that are closer to the underlying true labels. However, this approach introduces a considerable computational and memory overhead, as the prediction history must be stored and updated for the entire dataset throughout training [80].

To address this issue, Tarvainen et al. [89] propose the MT framework, which extends consistency regularization by introducing a dual-network architecture composed of a student and a teacher model, typically sharing an identical network structure. In this approach, the teacher network is not trained directly via backpropagation, instead, its parameters are updated as an EMA of the student's weights throughout training. Both models process differently perturbed versions of the same input, and a consistency loss is enforced between their predictions, encouraging the student to align its outputs with the more stable predictions of the teacher. For labelled data, the student is optimized using a combination of supervised loss and consistency loss, while for unlabelled data the training relies solely on the consistency constraint. By leveraging EMA-based weight averaging rather than maintaining prediction histories, MT effectively combines the advantages of the Π -Model and Temporal Ensembling while significantly reducing computational and memory overhead. Due to its simplicity, this framework has become one of the most widely adopted consistency-based methods in SSL. An overview of these three common consistency-based approaches is illustrated in Figure 3.13.

Meanwhile, many works extended the MT framework in different ways. One representative example is the Uncertainty-Aware Mean Teacher (UA-MT) proposed by Yu et al. [90], which addresses the limitation that teacher-generated targets can be noisy when labelled data are scarce, potentially leading the student model to learn incorrect information. UA-MT estimates the uncertainty of the teacher's predictions via Monte Carlo sampling and filtering out high-uncertainty targets to ensuring that only reliable outputs contribute to the consistency loss. Building upon this idea, the Anatomy-Aware Uncertainty framework (AAU-Net) [91] incorporates anatomical priors: a pre-trained denoising autoencoder maps teacher predictions to anatomically plausible reconstructions, and the discrepancy between these corrected outputs and the original predictions guides uncertainty estimation and consistency loss. Another extension, the Cross-head Mutual Mean-Teaching Network (CMMT-Net) [92], promotes model diversity using a shared encoder with dual decoders. Weakly augmented unlabelled samples are processed by the teacher to generate reliable predictions, which supervise the student under strong augmentations, enforcing consistency across perturbations and decoding paths. CMMT-Net also applies multi-level perturbations

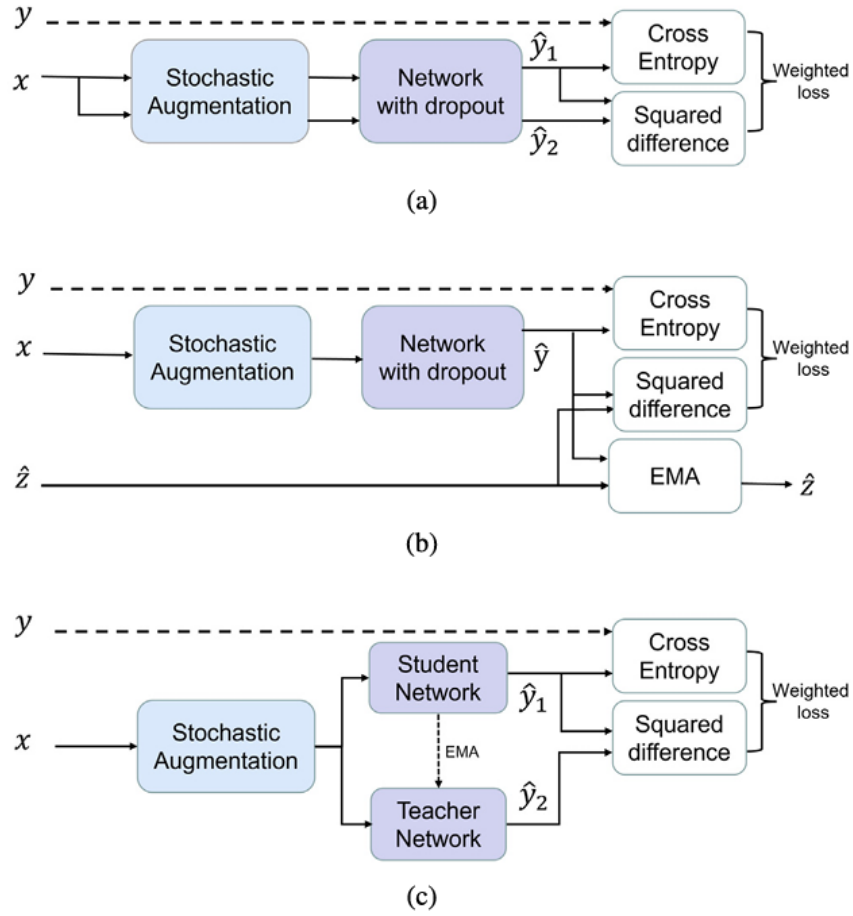


Figure 3.13: The classic architectures used in consistency learning. (a) Π -Model [87]; (b) Temporal ensembling [88]; (c) The mean-teacher architecture [89].

via adversarial noise and region-based mixing.

To further optimize the effectiveness of consistency regularization, Wu et al. [93] proposed SS-Net, a framework designed to simultaneously enforce pixel-level smoothness via strong perturbations and inter-class separation through a prototype-based strategy that encourages features to align with class prototypes while pushing different classes apart. However, when the dataset suffers from class imbalance, the model tends to overly focus on the majority class and neglect the minority class.

Overall, consistency regularization in semi-supervised medical image segmentation has evolved from simple perturbation-based strategies, such as the Π -Model and Temporal Ensembling, to more sophisticated frameworks like MT and its extensions, including UA-MT, AAU-Net, CMMT-Net, and SS-Net. Nevertheless, current approaches still face several challenges, including sensitivity to hyperparameters, variations in medical imaging data, and high computational resource requirements.

3.2.3 Consistency Regularization Methods Supervised by Pseudo-Labels

Hybrid methodologies integrating pseudo-labelling with consistency regularization have emerged as a particularly active research direction. This convergence aims to mitigate the principal limitations inherent to pseudo-labelling, as discussed in Section 3.2.1.

Among the most influential approaches combining pseudo-labelling and consistency regularization is FixMatch, proposed by Sohn et al. [94]. Originally developed for image classification, FixMatch generates pseudo-labels from weakly augmented unlabelled images and uses them to supervise predictions obtained from strongly augmented versions of the same samples, encouraging the model to produce consistent predictions under different perturbations while reducing the influence of unreliable pseudo-labels. To extend this concept to the field of medical image segmentation, Wang et al. [95] proposed PixelFixMatch, illustrated in Figure 3.14, which adapts the original framework to pixel-wise prediction tasks.

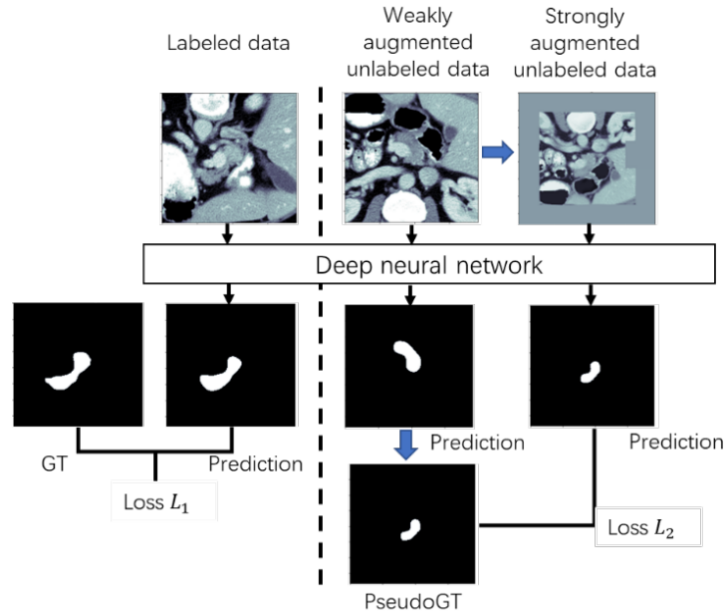


Figure 3.14: Overview of the PixelFixMatch framework for semi-supervised medical image segmentation [95].

To address pseudo-label unreliability, Su et al. [96] proposed the Mutual Learning with Reliable Pseudo-Label (MLRPL) framework, illustrated in Figure 3.15, which combines pseudo-labelling with consistency-driven mutual learning. Two decoders share a common encoder and generate independent predictions for unlabelled data as pseudo-label candidates. To suppress noisy supervision, MLRPL employs a dual reliability assessment: a mutual comparison selects the more confident prediction between decoders, while an intra-class consistency criterion evaluates pixel-wise reliability based on feature similarity to class prototypes. These cues guide knowledge transfer, ensuring learning relies primarily on high-quality pseudo-labels.

The effectiveness of the MLRPL framework is evident on the Left Atrium (LA) dataset [97],

where using only 20% labelled data, it outperforms the Bidirectional Copy–Paste (BCP) framework. BCP was proposed by Bai et al. [98] and operates by pasting regions cropped from a labelled image onto an unlabelled image, and vice versa, inserting regions from an unlabelled sample into a labelled one. To supervise these hybrid inputs, pseudo-labels generated by a teacher network for unlabelled images are combined with GT annotations using the same bidirectional copy–paste strategy, as illustrated in Figure 3.16.

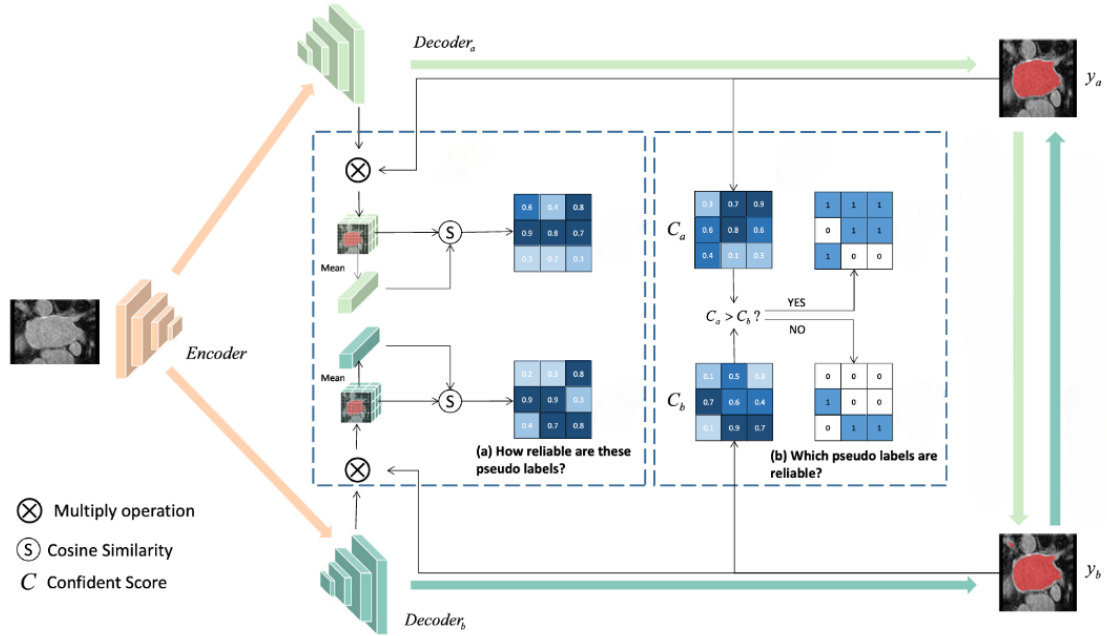


Figure 3.15: Overview of the MLRPL framework: (a) pseudo-label reliability assessed via intra-class feature similarity; (b) selection of reliable pseudo-labels through comparison of confidence scores from decoders (C_a and C_b), retaining higher-confidence predictions. Adapted from [96].

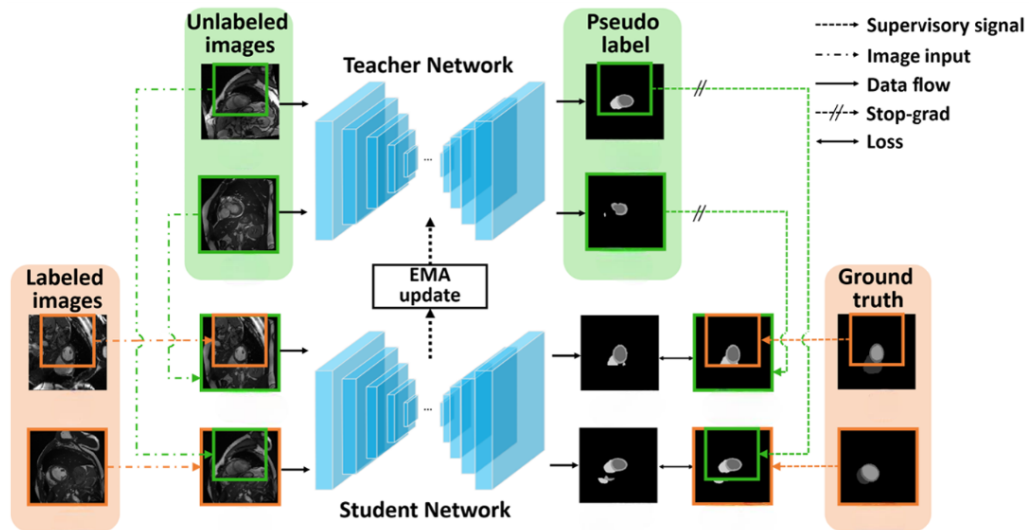


Figure 3.16: Overview of the BCP framework within the MT architecture. Adapted from [98].

3.2.4 Methods Combining Contrastive Learning and Consistency Regularization

To better leverage unlabelled data, recent research has increasingly focused on the joint use of contrastive learning and consistency regularization. Contrastive learning aims to enhance feature discriminability by encouraging representations of similar regions to be closer while separating dissimilar ones. By combining these two paradigms, constraints are simultaneously imposed in the feature representation space and in the output prediction space, leading to more robust and generalisable models [69].

The methodology proposed by Yang et al. [99] introduces a semi-supervised segmentation framework that jointly exploits feature-level consistency regularization and edge-aware contrastive learning to enhance segmentation accuracy, particularly in challenging edge regions. The architecture consists of a shared encoder and multiple decoders, each receiving progressively stronger perturbations applied directly to the feature space, guided by the statistics of the feature maps to perform controllable linear transformations. To leverage this structured perturbation effectively for learning from unlabelled data, the method incorporates a feature perturbation consistency loss, compelling the model to produce consistent predictions for the same input under varying perturbation strengths. Importantly, rather than simply averaging predictions from the different decoders, an uncertainty-weighted aggregation strategy is employed, fusing the outputs based on confidence levels derived from uncertainty estimates. Complementing this approach, an edge-aware contrastive learning strategy is applied in the feature space, where representations of boundary pixels are pulled towards stable class-specific prototypes and pushed away from other classes. By simultaneously constraining both the prediction space and the feature representation space, the framework enhances segmentation robustness, especially along object boundaries. An overview of this method is illustrated in Figure 3.17, highlighting the weak-to-strong perturbation and edge-aware contrastive learning strategy.

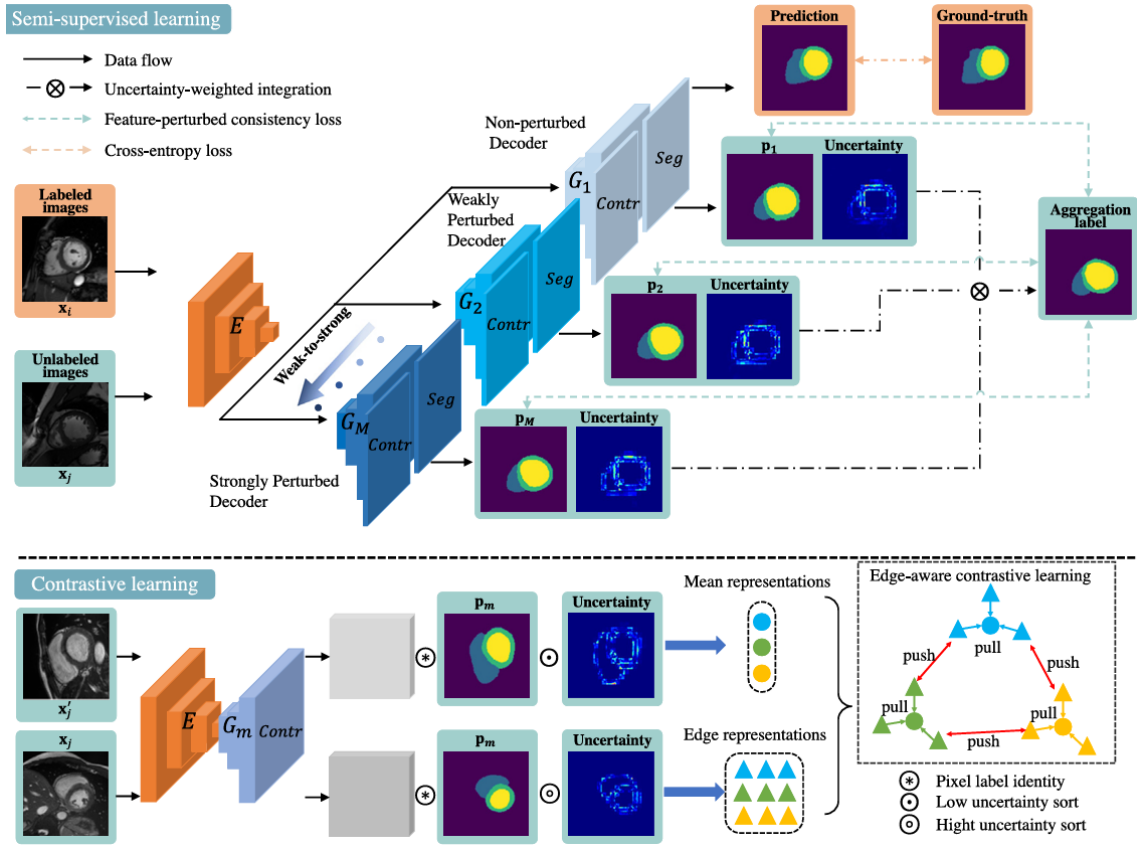


Figure 3.17: Overview of semi-supervised segmentation framework with weak-to-strong feature perturbations and edge-aware contrastive learning. [99].

SSL has become increasingly important in medical image segmentation, as it allows models to leverage the abundance of unlabelled data while minimizing the need for costly annotations. Rapid advances in DL have revitalised research in semi-supervised medical image segmentation. Approaches differ in how they exploit the available data, create pseudo-labels, and rely on labelled samples.

3.3 Future Directions in ILD Findings Segmentation

The current body of research on ILD findings segmentation demonstrates substantial progress in the application of DL techniques, particularly through SL approaches based on CNNs. These methods have achieved notable performance in both binary and multi-class segmentation tasks. Nevertheless, their strong dependence on large, fully annotated datasets continues to represent a critical bottleneck for scalability and the generalizability of existing models [64]. This limitation is especially pronounced in ILD research, where available datasets often contain incomplete or inconsistent pixel-level annotations, highly imbalanced annotations across different ILD findings, or rely on private collections that are partially annotated or weakly labelled. As a result, large

volumes of imaging data remain either unlabelled or weakly annotated, naturally motivating the adoption of SSL strategies capable of exploiting both labelled and unlabelled information.

Building upon these challenges, SSL strategies, discussed in Section 3.1.4, remain relatively underexplored in the context of ILD findings segmentation. Existing SSL approaches have demonstrated that incorporating unlabelled data during training can lead to measurable improvements in segmentation performance when compared to purely supervised baselines. Among these, the study by Cai et al. [57] stands out for its coverage of multiple types of lesion and its notably strong segmentation results. Nevertheless, the limited number of SSL studies focused on ILD, combined with the wide range of methodologies used in medical image segmentation literature, makes it challenging to establish general conclusions about the most effective strategies for this task.

Moving forward, an essential opportunity lies in systematically evaluating SSL strategies for ILD findings segmentation. By benchmarking different SSL approaches, it will be possible to clarify their relative effectiveness and limitations. This evaluation will provide practical guidance for selecting methods that make optimal use of both labelled and unlabelled data, ultimately leading to more robust and generalizable segmentation models for ILD.

Chapter 4

Supervised Learning for ILD Findings Segmentation

This chapter presents the SL approach developed as the initial stage of this dissertation for ILD findings segmentation in HRCT images. First, the primary database used in this dissertation is described, including the available annotations, the class grouping adopted, and the data splitting strategy used for training, validation, and testing. The chapter then introduces the supervised segmentation methodology, with particular focus on the strategies designed to handle sparse annotations and on the network configurations evaluated. Finally, the chapter presents and discusses the quantitative and qualitative results obtained with the supervised experiments.

4.1 Methodology

In the initial phase, a supervised learning approach was employed for ILD findings segmentation in HRCT images. The main objective of this supervised baseline method was to establish a reference performance for evaluating the proposed pipeline, providing insight into the expected capabilities of the model when trained exclusively with annotated data, and serving as a benchmark for comparison with more advanced SSL strategies. Given the partially labelled nature of the MedGIFT ILD database [58], different supervision strategies were investigated to identify the most suitable approach for handling sparse annotations during training, while also analysing their impact on segmentation performance. Furthermore, network variants were explored to assess their influence on the segmentation performance.

4.1.1 Data

The MedGIFT ILD database [58] was selected as the principal dataset for this dissertation. Developed at the University Hospitals of Geneva, it is a publicly available database comprising 119 studies acquired from 113 different patients. Each study consists of a volumetric HRCT acquisition represented by a set of axial slices, where each slice is stored as a 512×512 pixel matrix with a slice thickness ranging from 1 to 2 mm and an inter-slice spacing of approximately 10 to 15 mm.

The number of slices per study varies across patients, with an average of approximately 24 slices per scan.

For each CT slice, a corresponding lung binary mask is provided, with the same spatial resolution as the original image. These masks delineate the lung parenchyma and allow the exclusion of non-lung regions. In addition to the lung masks, the dataset includes ROI annotations describing pathological findings within the lung fields.

The MedGIFT ILD database [58] provides annotations for multiple pulmonary findings, including healthy tissue, emphysema, ground-glass opacities, FRF, nodular findings, consolidation, cysts, bronchial abnormalities, air trapping, and increased attenuation. For this dissertation, some of the original labels were grouped into broader clinically relevant categories, following a grouping scheme approved by an experienced thoracic radiologist with expertise in ILD. Specifically, the fibrosis class combined fibrosis, reticulation, bronchiectasis, and early fibrosis, while the nodules class grouped macronodules, micronodules, and peripheral micronodules. As a result, the final pathological classes considered were fibrosis, nodules, ground-glass opacities, consolidation, cysts, and emphysema. Moreover, under-represented labels, namely bronchial wall thickening, air trapping, increased attenuation, tuberculosis, and pcp, were merged into a single class named "others", whereas healthy tissue was kept as an independent category.

Table 4.1 summarises the mapping between the original MedGIFT annotations and the final set of classes used, together with the corresponding number of studies, axial slices, and percentage of pixels associated with each class. Additionally, the table includes the background class, corresponding to regions outside the lungs, and the unlabelled class, representing pixels inside the lung region that were not assigned to any category.

Table 4.1: Summary of the original MedGIFT labels associated with each class, together with the number of studies, axial slices, and percentage of pixels.

Original Labels	Class	Studies	Slices	% Pixels
–	Background	119	3020	87.07
Fibrosis, reticulation, bronchiectasis, early fibrosis	Fibrosis	55	405	0.26
Macronodules, micronodules, peripheral micronodules	Nodules	30	235	0.32
Ground-glass opacities	Ground-glass	37	242	0.13
Consolidation	Consolidation	14	116	0.02
Cysts	Cysts	3	7	< 0.01
Emphysema	Emphysema	8	66	0.06
Bronchial wall thickening, air trapping, increased attenuation, tuberculosis, pcp	Others	8	56	0.06
Healthy	Healthy	14	101	0.24
–	Unlabelled	119	2936	11.84

The partially labelled nature of the MedGIFT ILD database [58] is primarily a consequence of its annotation protocol. Only pulmonary findings that were clinically relevant to the underlying disease were annotated, rather than performing exhaustive pixel-wise labelling of all structures

present in the pulmonary parenchyma. Representative examples of the annotation strategy are shown in Figure 4.1. In addition, the annotation process was highly time-consuming, requiring an average of 3 hours and 30 minutes per case for annotation of the image series and associated clinical data, with a total of 1183.75 person-hours spent during database construction. Limitations also arise in the representation of the healthy class, as normal lung tissue was annotated in patients with a history of ILD. The absence of dedicated healthy subjects who undergo HRCT examinations is mainly related to practical and ethical constraints, including the associated high cost and radiation exposure. As a result, the available annotations reflect a limited representation of normal lung parenchyma.

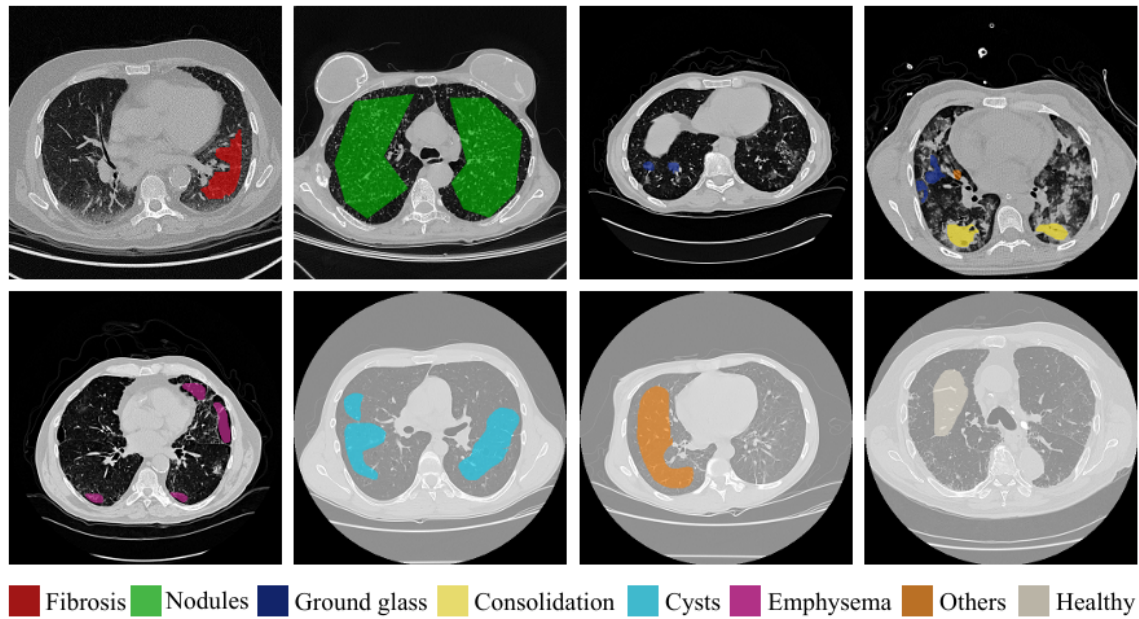


Figure 4.1: Representative examples of annotations in the MedGIFT ILD database, illustrating its partially labelled nature.

For the planned experiments, the database was divided to train and evaluate the segmentation models. A stratified approach was applied to create folds that aim to balance the distribution of the target pathological findings across the subsets, while excluding healthy tissue and "others" categories from the stratification process. Specifically, all slices from a given patient remained together in the same fold to avoid leakage of patient-specific information between training and validation sets, which could otherwise artificially inflate performance metrics. Furthermore, when a patient had more than one HRCT study, all corresponding studies were assigned to the same subset.

Initially, 80% of the studies were assigned to the training and validation pool, while the remaining 20% were reserved as a held-out test set. This test subset remained fixed throughout all experiments. The remaining 80% of the studies were subsequently divided into five folds, and a 5-fold cross-validation strategy was employed, where one fold was used for validation and the remaining folds for training at each iteration.

The resulting distribution of patients and axial slices across folds is reported in Tables 4.2 and 4.3, respectively. Fold 0 corresponds to the test set and therefore represents 20% of the complete dataset, whereas the remaining five folds were generated exclusively from the training/validation pool, corresponding to the remaining 80% of the dataset. Consequently, fold 0 is not directly comparable to the other folds in terms of sample distribution.

Table 4.2: Distribution of patients across pathological findings and folds.

Fold	Fibrosis	Nodules	Ground Glass	Consolidation	Cysts	Emphysema
0	10	7	8	3	1	2
1	7	3	5	2	0	1
2	7	6	6	2	0	1
3	8	5	4	2	0	1
4	7	3	6	1	0	1
5	10	6	4	3	0	2

Table 4.3: Distribution of slices across pathological findings and folds.

Fold	Fibrosis	Nodules	Ground Glass	Consolidation	Cysts	Emphysema
0	80	43	49	18	1	20
1	68	33	39	24	0	18
2	65	35	38	16	0	6
3	63	39	36	26	0	6
4	64	40	43	13	0	5
5	57	45	34	16	0	11

Given the sparse occurrence of certain findings, in particular cysts, special care was taken to ensure that these findings were represented in the training, validation, and test subsets. Although a stratified splitting strategy was applied to balance the distribution of pathological findings across folds, cysts remained highly under-represented. In the training/validation pool, only two studies contained cysts. Due to this limited number of cases, these studies could not be consistently assigned to specific fold. Instead, they were kept outside the fixed fold partitioning and dynamically included during each cross-validation iteration, with one case assigned to the training set and the other to the validation set. Consequently, although not evenly distributed across folds, cyst-related information was exposed to both training and validation processes across the cross-validation procedure.

4.1.2 Model Architecture and Training Details

To perform multi-class segmentation of ILD findings, the nnU-Net framework was applied [100]. nnU-Net is an extension of the original U-Net architecture [101] that provides a fully automated and self-configuring pipeline for medical image segmentation tasks. The framework follows a configuration strategy that begins with the extraction of dataset-specific properties, referred to as

the data fingerprint. This process analyses relevant characteristics such as image modality, spatial resolution, and intensity value distribution. Based on this information, nnU-Net automatically defines rule-based preprocessing steps, including intensity normalization. Subsequently, the network topology, input patch size, and batch size are determined based on the properties of the dataset and the available GPU memory. Training-related components, including the loss function, optimizer, learning rate, and number of training epochs, are also predefined within the framework. nnU-Net provides multiple configuration variants, including a 2D model, a low-resolution 3D model, and a cascaded architecture for high-resolution 3D segmentation.

The supervised segmentation framework adopted in this dissertation is based on the 2D configuration of the nnU-Net architecture. The adoption of this configuration was motivated by the acquisition characteristics of the MedGIFT ILD database described previously. Although each study is volumetric, the scans present a highly anisotropic structure, with high in-plane resolution but sparse sampling along the axial direction. Consequently, neighbouring slices do not provide a densely sampled anatomical continuation of the same structures, and the available GT annotations are not always continuous across slices. Under these conditions, 3D convolutional models could be affected by discontinuous through-plane information, while resampling the volumes to an isotropic spacing would require strong interpolation along the axial direction. A 2.5D approach using neighbouring slices as additional input channels was also considered less suitable, since the anatomical gap between slices could provide limited contextual benefit. For these reasons, a 2D configuration was considered more appropriate, allowing the model to focus on the high-resolution in-plane texture patterns that characterise ILD findings while avoiding unreliable through-plane context.

The selected 2D nnU-Net follows an encoder-decoder structure inspired by the original U-Net design. The network exhibits a symmetric U-shaped configuration, composed of a contracting path, the encoder, and an expansive path, the decoder, connected through skip connections at corresponding levels. A schematic representation of the specific 2D nnU-Net architecture is shown in Figure 4.2, comprising three main components:

- **Encoder** – The encoder comprises eight resolution stages responsible for progressively extracting high-level contextual information from the input image. Each stage contains two convolutional layers with kernel sizes of 3×3 , followed by Instance Normalization and Leaky ReLU activation functions. Starting from an input patch size of 512×512 pixels, downsampling is performed with a stride of 2×2 , progressively reducing the spatial resolution while increasing the number of feature channels. The feature dimensionality increases from 32 channels at the first stage to 512 channels in the deepest layers of the network. After seven successive downsampling operations, the spatial resolution is reduced to 4×4 pixels.
- **Decoder** – The decoder is responsible for progressively reconstructing the spatial resolution of the feature maps. It consists of seven decoding stages, each composed of two 3×3 convolutional layers followed by Instance Normalization and Leaky ReLU activation. Up-sampling is performed using transposed convolutions with a stride of 2×2 , progressively

restoring the spatial resolution from 4×4 back to 512×512 pixels. At each stage, the up-sampled feature maps are concatenated with the corresponding encoder feature maps via skip connections, allowing the network to combine low-level spatial details with high-level semantic information.

- **Final Layer** – At the final stage of the network, a 1×1 convolutional layer maps the decoded feature representations to the desired number of output channels corresponding to the segmentation classes. In this work, the output layer produces pixel-wise multi-class segmentation maps representing the different ILD findings.

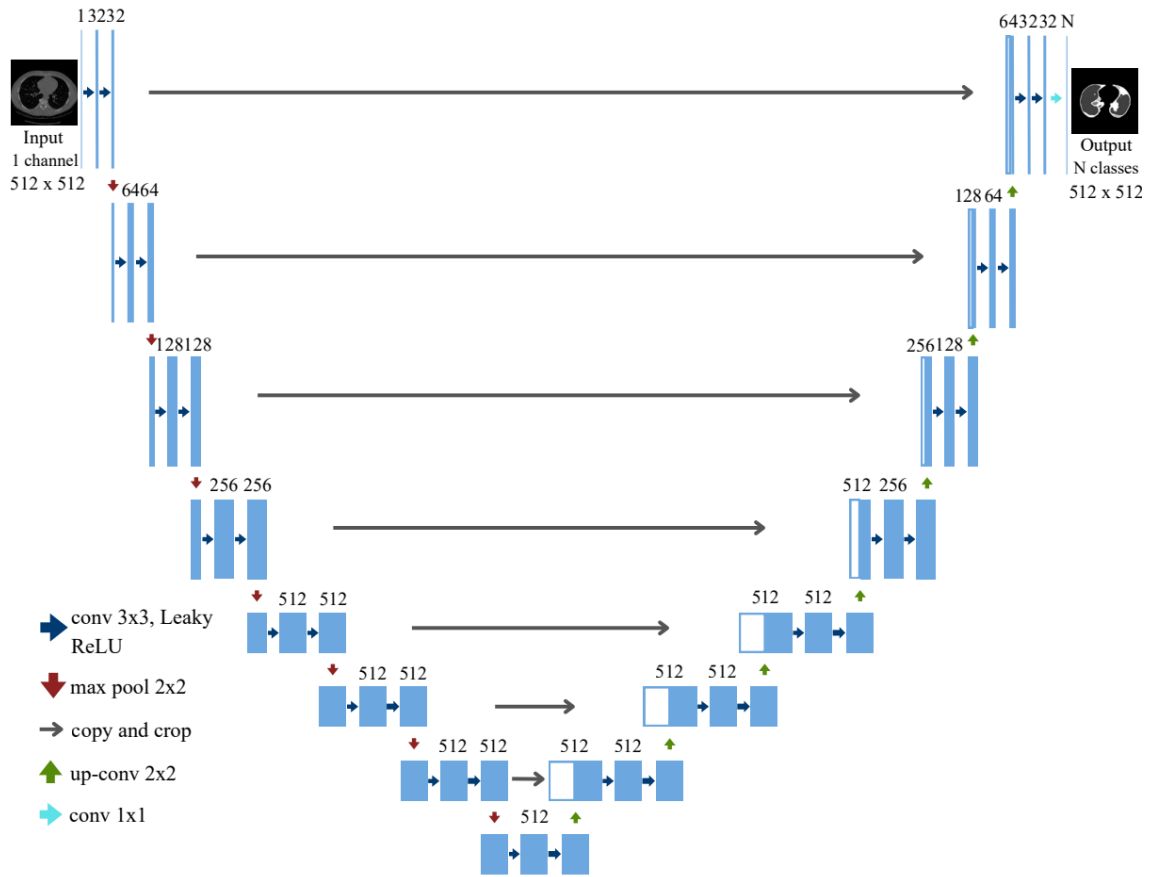


Figure 4.2: The architecture of the specific 2D nnU-Net network used.

Before training, all data underwent preprocessing to ensure consistency across the dataset. Image intensities were clipped and normalised according to dataset-specific intensity statistics, following the CT normalisation procedure of nnU-Net. Specifically, voxel intensities were limited to the 0.5th and 99.5th percentiles of the foreground, corresponding to approximately -884 and 1210 HU, respectively. Subsequently, intensities were z-score normalised using the mean and standard deviation computed from the dataset foreground, resulting in images with zero-centred intensities and unit variance.

Data augmentation techniques were applied to increase the variability of the data. These techniques included random rotations within ± 10 degrees, scaling operations, as well as intensity-based augmentations including Gaussian noise, Gaussian blurring, multiplicative brightness transformations, contrast adjustments, gamma transformations, and simulated low-resolution transformations.

Model training was performed for a maximum of 300 epochs using Stochastic Gradient Descent with Nesterov momentum for network optimisation. During training, a batch size of 12 samples was used, where six samples were constrained to contain at least one of the major pathological findings considered in this dissertation, namely fibrosis, nodules, ground-glass opacities, consolidation, cysts, and emphysema. The remaining six samples were randomly selected from the training set. The initial learning rate was set to 1×10^{-3} . To reduce unnecessary computational cost and mitigate overfitting, early stopping was applied with patience of 30 epochs based on the mean validation DSC across all segmentation classes. All experiments were implemented in PyTorch machine learning library and trained on an NVIDIA GeForce GTX 1080 GPU with 8.0 GB of VRAM.

The network was trained using the standard nnU-Net compound loss function, which combines cross-entropy loss with soft Dice loss. The cross-entropy component provides pixel-wise supervision by penalising incorrect class probabilities at each pixel, while the Dice component directly optimises the spatial overlap between the predicted segmentation and the GT mask. This combination is particularly suitable for medical image segmentation tasks. The total loss is defined as:

$$\mathcal{L} = \lambda_{CE} \mathcal{L}_{CE} + \lambda_D \mathcal{L}_D, \quad (4.1)$$

where $\mathcal{L}_{CE}$ denotes the cross-entropy loss, $\mathcal{L}_D$ denotes the soft Dice loss, and both components were assigned equal weights, with $\lambda_{CE} = 1$ and $\lambda_D = 1$.

For a batch of 2D images, let B denote the batch size, Ω the set of pixels in each image, and C the set of segmentation classes. The network outputs, for each pixel i of image b , a probability $p_{b,i,c}$ for each class $c \in C$ after the softmax operation. The GT class of that pixel is denoted by $y_{b,i}$. The cross-entropy loss is computed independently for each pixel and averaged over the pixels in the batch:

$$\mathcal{L}_{CE} = -\frac{1}{B|\Omega|} \sum_{b=1}^B \sum_{i \in \Omega} \log(p_{b,i,y_{b,i}}) \quad (4.2)$$

The Dice component was computed using the predicted softmax probabilities and the one-hot encoded GT masks. In accordance with the nnU-Net implementation, the background class was excluded from the Dice computation, so the overlap term focused only on the foreground segmentation classes. For each foreground class $c \in C \setminus \{0\}$, the soft Dice coefficient is given by:

$$DSC_c = \frac{2 \sum_{b=1}^B \sum_{i \in \Omega} p_{b,i,c} g_{b,i,c} + \epsilon}{\sum_{b=1}^B \sum_{i \in \Omega} p_{b,i,c} + \sum_{b=1}^B \sum_{i \in \Omega} g_{b,i,c} + \epsilon}, \quad (4.3)$$

where $g_{b,i,c}$ is the one-hot encoded GT label, equal to 1 when pixel i of image b belongs to class c and 0 otherwise. The constant $\varepsilon = 10^{-5}$ was used to avoid division by zero. The Dice loss is then defined as:

$$\mathcal{L}_D = 1 - \frac{1}{|C| - 1} \sum_{c \in C \setminus \{0\}} DSC_c \quad (4.4)$$

Additionally, deep supervision was used during training, as in nnU-Net. The same compound loss was computed at multiple decoder outputs with different spatial resolutions, and the resulting losses were combined using exponentially decreasing weights, assigning greater importance to higher-resolution predictions.

4.1.3 Evaluation

The performance of the segmentation models was evaluated using the DSC, precision, recall, and specificity. Since the dataset contains sparse and incomplete annotations, the evaluation was restricted to labelled pixels only. Therefore, unlabelled pixels were excluded from the metric computation, ensuring that the reported values reflected performance only in regions where GT information was available.

The DSC is a widely used metric for measuring the similarity between two sets. In this context, it was employed to quantify the overlap between the segmentation mask predicted by the network and the corresponding GT mask for each pathological finding. Mathematically, it is defined as:

$$DSC = \frac{2|A \cap B|}{|A| + |B|}, \quad (4.5)$$

where A represents the set of pixels in the GT mask for a given pathological finding, and B represents the set of pixels in the predicted mask for the same finding. The DSC ranges from 0, indicating no overlap, to 1, indicating perfect overlap.

In addition to DSC, precision and recall were computed to further characterise segmentation performance. Precision measures the proportion of pixels predicted as belonging to a given class that were correctly classified, and is defined as:

$$Precision = \frac{TP}{TP + FP}, \quad (4.6)$$

where TP denotes true positives and FP denotes false positives. A high precision indicates that the model produces few FP predictions for a given pathological finding.

Recall, also known as sensitivity, measures the proportion of GT pixels belonging to a given class that were correctly detected by the model:

$$Recall = \frac{TP}{TP + FN}, \quad (4.7)$$

where FN denotes false negatives. A high recall indicates that the model successfully detects most pixels belonging to the pathological finding.

Specificity was also computed to evaluate the ability of the model to correctly identify background pixels, corresponding to pixels not belonging to any pathological finding. It is defined as:

$$Specificity = \frac{TN}{TN + FP}, \quad (4.8)$$

where TN denotes the number of background pixels correctly classified as background, and FP denotes the number of background pixels incorrectly classified as a pathological finding. A high specificity therefore indicates that the model produces few FP pathological predictions in regions labelled as background. However, since background pixels represent the majority of the annotated area, specificity values are expected to be high and should be interpreted with caution, as this metric may be inflated by the strong class imbalance between background and the other classes.

All metrics were independently computed for each segmentation class and then summarised across classes to assess overall model performance. The average was also computed across segmentation classes as an unweighted class-wise macro-average, corresponding to the arithmetic mean of the metric values obtained for each class.

4.1.4 Supervision Strategies for Sparse Annotations

Although the supervised training pipeline follows the nnU-Net framework, the annotation setting considered in this dissertation differs from a fully annotated segmentation dataset. The available masks contain sparse annotations, where only a subset of pathological findings is explicitly labelled. Consequently, regions without annotation cannot always be assumed to correspond to the true background, since they may contain unlabelled pathological tissue or other structures not represented in the target classes. To address this limitation, different supervision strategies were investigated, varying how unlabelled pixels, healthy lung tissue, other pathological findings, and non-lung regions were represented during training.

Four supervision strategies were explored. In all strategies, the main target classes corresponded to the major pathological findings considered, namely fibrosis, nodules, ground-glass opacities, consolidation, cysts, and emphysema. To isolate the effect of the supervision strategy, all experiments were implemented using the same baseline 2D nnU-Net configuration, with a patch size of 512×512 pixels and a batch size of 12.

The first strategy was used as the fully supervised baseline. In this setting, the model was trained using the major pathological findings and a single background class. This background class included all pixels not belonging to one of the target findings, including unlabelled pixels, healthy lung tissue, non-target pathological findings grouped as "others", and regions outside the lungs. Therefore, all pixels contributed to the training loss, following the conventional assumption that every pixel has a reliable label.

The second strategy introduced the ignore label to exclude unlabelled pixels from the supervised loss computation. The target pathological findings remained explicitly labelled, while healthy lung tissue, non-target pathological findings, and non-lung regions were still represented

as background. This strategy aimed to reduce noisy supervision caused by treating unlabelled regions as true background. This was particularly relevant given the annotation protocol of the dataset, since unlabelled lung areas could still contain target pathological findings that were not explicitly delineated.

The third strategy further refined the label representation by separating non-lung regions from lung-related non-target regions. As in the second strategy, unlabelled pixels were assigned to the ignore label and excluded from the supervised loss. However, the previous single background category was divided into two classes: a background class representing regions outside the lungs, and an "others" class representing lung regions that did not correspond to the major target findings. This "others" class included healthy lung tissue and pathological findings not considered among the major classes.

The fourth strategy extended the previous representation by separating healthy lung tissue from non-target pathological findings. Unlabelled pixels were assigned to the ignore label, while the background class represented only regions outside the lungs. Lung parenchyma without pathological findings was represented by an independent healthy class, as explicitly modelling healthy tissue was considered important for improving the model's understanding of normal lung appearance. Non-target pathological findings were grouped into a separate "others" class. Overall, this strategy provided the most detailed label space, distinguishing between major pathological findings, healthy parenchyma, other abnormalities, and non-lung background regions.

In addition, an anatomical consistency constraint was introduced to discourage anatomically implausible predictions. Since ignored pixels were known to belong to the lung parenchyma, predicting the background class in these regions was penalized. This encouraged the model to classify lung regions as either one of the target pathological findings, the healthy class, or the "others" class, rather than incorrectly assigning them to non-lung background. This constraint was implemented as an additional term added to the standard nnU-Net compound loss. Furthermore, class weighting was applied to the cross-entropy component, assigning a lower weight to the background class and higher weights to the remaining classes, thereby increasing the contribution of lung-related and pathological classes during training.

Let Ω_{ign} denote the set of pixels assigned to the ignore label. The anatomical consistency penalty was defined as the mean predicted probability of the background class over these pixels:

$$\mathcal{L}_{bg} = \frac{1}{|\Omega_{ign}|} \sum_{(b,i) \in \Omega_{ign}} p_{b,i,0}, \quad (4.9)$$

where $p_{b,i,0}$ represents the predicted probability of the background class for pixel i of image b . This term increases when the model assigns high background probability to ignored pixels inside the lungs, encouraging these regions to instead be classified as one of the lung-related classes.

The final loss used in this strategy was therefore defined as:

$$\mathcal{L}_{final} = \mathcal{L} + \lambda_{bg} \mathcal{L}_{bg}, \quad (4.10)$$

where $\mathcal{L}$ corresponds to the standard nnU-Net compound loss defined in Equation 4.1, and $\lambda_{bg} = 2$ controls the contribution of the anatomical consistency penalty.

The four supervision strategies are summarised in Table 4.4, highlighting the differences in label representation and training constraints.

Table 4.4: Summary of the four supervision strategies for sparse annotations.

Strategy	Outside Lung	Unlabelled	Others	Healthy	Constraints
1	Background	Background	Background	Background	None
2	Background	Ignore	Background	Background	None
3	Background	Ignore	Others	Others	None
4	Background	Ignore	Others	Healthy	Anatomical consistency + class weighting

4.1.5 Network Configurations and Variants

In addition to the different supervision strategies, three network configurations were considered to evaluate the impact of input patch size, batch size, and architectural design on segmentation performance.

The first configuration corresponded to the baseline 2D nnU-Net plan generated by the framework after dataset preprocessing. This configuration used an input patch size of 512×512 pixels and a batch size of 12. Since the median image size in the dataset was also 512×512 , this setting allowed the network to process the full in-plane image context during training. This can be advantageous in ILD segmentation, as some findings are characterised not only by their appearance but also by their anatomical distribution within the lungs. For example, peripheral micronodules may be more reliably interpreted when their proximity to the pleural surface and their spatial arrangement within the lung fields are visible to the network. The baseline architecture followed the standard U-Net design, as shown in Figure 4.2.

The second configuration kept the same general U-Net architecture but reduced the input patch size to 320×320 pixels. This reduction decreased GPU memory consumption, allowing the batch size to be increased from 12 to 16. A larger batch size may provide more stable gradient estimates during training, while the smaller patch size increases the number of possible sampled regions from each image. However, reducing the patch size also reduces the amount of anatomical context available to the network in each training sample. Due to the smaller spatial dimensions of the input patches, the depth of the network was also reduced. While the baseline 512×512 configuration supported eight resolution stages, the 320×320 variant used seven stages. This adjustment was necessary because each downsampling operation reduces the spatial resolution by a factor of two. With a smaller input patch, maintaining the same number of downsampling stages would lead to non-integer feature map sizes in the deepest layers.

The third configuration was based on the reduced-patch setting but replaced the standard U-Net architecture with a residual encoder U-Net. This model used the same 320×320 patch size and

batch size of 16, while introducing residual blocks in the encoder pathway. Residual connections, originally proposed in ResNet by He et al. [102], allow the input of a convolutional block to be added to its output. This design improves gradient propagation during backpropagation and helps mitigate degradation problems commonly associated with increasing network depth [103]. In this configuration, the number of residual blocks per encoder stage increased with depth, increasing the representational capacity of the feature extractor while preserving the overall encoder-decoder structure of the segmentation network. For a given encoder stage l , with input feature map $x^{(l)}$, a residual block can be expressed as:

$$\text{ResidualBlock}^{(l)}(x^{(l)}) = F^{(l)}(x^{(l)}) + x^{(l)}, \quad (4.11)$$

where $F^{(l)}$ represents the sequence of convolutional operations applied to $x^{(l)}$ within the block.

Overall, these configurations were designed to compare the trade-off between spatial context, batch size, and representational capacity. The baseline configuration prioritised full-image context, the reduced-patch configuration increased the batch size while lowering memory requirements, and the residual encoder variant investigated whether a more expressive feature extractor could improve segmentation performance under the same reduced-patch setting.

4.2 Results and Discussion

This section presents and discusses the results obtained with the supervised segmentation experiments described in the previous section. The evaluation includes quantitative performance metrics as well as qualitative visual examples of the predicted segmentation masks.

The section is divided into two main parts. First, subsection 4.2.1 evaluates the impact of the different supervision strategies designed to address sparse annotations. Second, subsection 4.2.2 investigates the influence of different network configurations.

4.2.1 Supervision Strategies for Sparse Annotations

The first set of experiments focused on the different strategies used to train the model under sparse annotation conditions. These experiments aimed to assess how the different label representations affected the segmentation performance of each class.

Table 4.5 presents the DSC obtained for each supervision strategy, with the values reported separately for each class. Since the label space differs between strategies, some classes were not evaluated in all cases. In particular, the "others" and healthy classes were not included in strategies 1 and 2, while the healthy class was not included in strategy 3.

Table 4.5: Segmentation performance of different strategies measured by the DSC. Values are reported as mean $\pm$ standard deviation (%) across five data splits. The average DSC is computed only over the major findings. The cysts class is omitted, as it achieved $0.00 \pm 0.00\%$ across all experiments.

Strategy	Fibrosis	Nodules	Ground glass	Consolidation	Emphysema	Others	Healthy	Average
1	41.97 $\pm$ 13.76	20.96 $\pm$ 12.46	18.22 $\pm$ 8.37	11.62 $\pm$ 8.07	8.31 $\pm$ 11.33	–	–	16.85 $\pm$ 13.13
2	80.85 $\pm$ 4.57	50.15 $\pm$ 1.68	49.00$\pm$2.89	12.48 $\pm$ 9.02	47.00 $\pm$ 20.52	–	–	39.91 $\pm$ 26.64
3	80.42 $\pm$ 4.53	55.97$\pm$3.31	46.25 $\pm$ 4.74	17.07 $\pm$ 12.03	51.06 $\pm$ 23.34	43.81$\pm$6.38	–	41.79 $\pm$ 26.32
4	81.35$\pm$4.72	53.64 $\pm$ 8.69	44.84 $\pm$ 10.28	22.95$\pm$15.66	52.89$\pm$14.95	0.01 $\pm$ 0.02	30.66 $\pm$ 16.20	42.26$\pm$25.61

The results presented in Table 4.5 highlight a clear difference between the fully supervised baseline and the strategies designed to account for sparse annotations. Strategy 1 obtained the lowest average DSC among the major findings, with an average value of 16.85%. In contrast, strategies 2, 3 and 4 achieved substantially higher average DSC values, ranging from 39.91% to 42.61%. This indicates that treating all non-annotated pixels as background, as done in strategy 1, negatively affected the segmentation performance. Since the dataset contains incomplete annotations, many pixels assigned to the background during training may in fact correspond to healthy tissue or unlabelled pathological regions. As a result, the model may have been penalised for potentially correct predictions made in regions without explicit annotation, introducing noisy supervision and likely encouraging predictions biased toward the background class.

The introduction of the ignore label in strategy 2 led to a marked improvement in performance for most classes. The average DSC increased from 16.85% in strategy 1 to 39.91%, showing that excluding unlabelled pixels from the loss computation was beneficial. This improvement was particularly evident for fibrosis, nodules, ground glass, and emphysema, suggesting that removing uncertain regions from the supervised objective allowed the model to learn more effectively from reliable annotated pixels.

Strategy 3 further improved the average DSC to 41.79%, representing the best performance for nodules, and a slight improvement for consolidation compared with strategy 2. The improvement observed for nodules and consolidation suggests that separating regions inside the lungs from the external background helped the model learn a more appropriate distinction between pulmonary tissue and non-lung areas. The "others" class itself achieved a DSC of 43.81%, indicating that the model was able to learn this broader category to some extent. This result may be explained by the fact that, in this strategy, the class included both healthy tissue and non-target findings, providing a larger number of annotated pixels for training despite grouping visually heterogeneous patterns under the same label.

Strategy 4 achieved the highest average DSC among the major findings, with a value of 42.61%. This strategy obtained the best results for fibrosis, consolidation and emphysema. Compared with strategy 3, the gain in average performance was modest, but it suggests that explicitly modelling healthy tissue as an independent class can provide additional useful information dur-

ing training. This distinction is particularly relevant in multi-class segmentation, since the model must not only identify pathological findings, but also learn to separate abnormal tissue from normal lung parenchyma. However, the healthy class obtained a DSC of 30.66%, with a relatively high standard deviation, indicating variability across folds. Moreover, the "others" class obtained an almost null DSC, despite performing well in strategy 3. This suggests that once healthy tissue was separated into its own class, the remaining "other" category became too scarce to be learnt reliably.

Despite the improvements observed in strategies 2 to 4, the performance remained highly dependent on the class. Fibrosis was consistently the best segmented finding, reaching DSC values above 80% in all strategies where the loss was computed only on reliable pixels. This result is consistent with its high representation in the database, as fibrosis was among the most represented pathological findings and appeared in the largest number of studies and annotated slices. Nodules and ground-glass opacities also benefited from the sparse-annotation strategies, although their DSC values remained lower than fibrosis. Consolidation showed more variable and generally lower results, which may be related to its smaller number of annotated examples. In contrast, emphysema achieved relatively good DSC values despite its limited representation in the database. This may indicate that its visual pattern was more distinguishable from the other classes, making it easier for the model to learn compared with other under-represented findings.

The cyst class was not segmented successfully by any strategy, with a DSC of 0.00% in all cases. This result is most likely explained by the extremely limited representation of cysts in the dataset, which included only a very small number of annotated slices. Under these conditions, the model had insufficient examples to learn a consistent visual representation of this class, even when cyst-containing cases were dynamically included during cross-validation.

Taken together, these results indicate that ignoring unlabelled pixels is essential when training on this partially labelled database. The transition from strategy 1 to strategy 2 produced the largest improvement, showing that the main limitation of the baseline approach was the incorrect treatment of unlabelled regions as background. The additional refinements introduced in strategies 3 and 4 led to smaller but relevant gains, with strategy 4 achieving the highest average DSC across the major findings. Therefore, strategy 4 was selected as the most suitable supervision strategy for the subsequent experiments with different network configurations.

To complement the quantitative analysis, qualitative examples were also examined to assess the spatial behaviour of the predicted masks and the ability to localize pathological regions. Figure 4.3 presents representative examples obtained with the different supervision strategies, illustrating the visual impact of the adopted label representations.

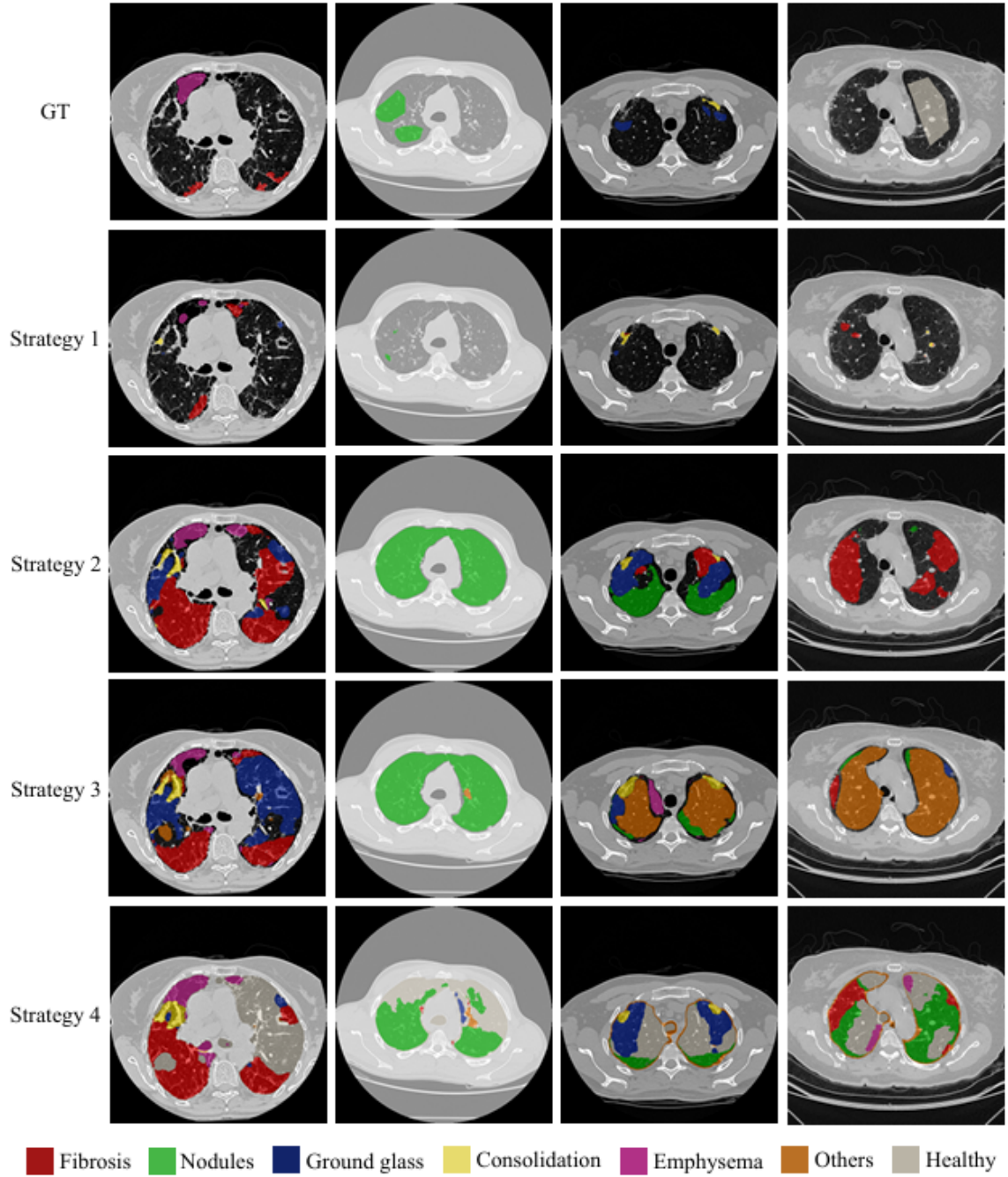


Figure 4.3: Qualitative comparison of the segmentation masks obtained with the four supervision strategies, including representative HRCT slices, the corresponding GT annotations, and the predicted masks generated by each strategy.

The qualitative examples presented in Figure 4.3 further illustrate the challenges associated with the sparse annotation structure of the MedGIFT database [58]. As can be observed, the GT masks often annotate only limited regions of the pathological findings, while other regions in the same slice with similar visual appearance remain unlabelled. Moreover, the annotation style is not entirely consistent across studies. In some cases, the annotated regions exhibit irregular

contours suggestive of manual delineation, whereas in other cases the annotations appear more geometrically shaped or coarse, reflecting variability in the annotation process.

In the first example, strategy 1 produced only limited pathological predictions, mainly restricted to small fibrosis and emphysema regions overlapping with the annotated areas. In contrast, strategies 2, 3, and 4 generated substantially larger segmentation masks, extending beyond the annotated GT regions into visually similar tissue. Since the evaluation was restricted to labelled pixels, predictions in unlabelled regions did not directly contribute to the quantitative metrics. Nevertheless, these additional regions are relevant for qualitative analysis, as several of them appear anatomically plausible and show consistent agreement among the different sparse-aware strategies, particularly for fibrosis, consolidation, and emphysema. However, due to the absence of complete annotations, it is not possible to determine whether all these additional predictions correspond to true pathological findings.

In the second example, strategy 1 again produced very limited predictions, whereas strategies 2 and 3 segmented large portions of the lung parenchyma as nodules. Strategy 4 generated a more heterogeneous prediction map, separating healthy tissue from pathological regions. Compared with strategies 2 and 3, the predictions obtained with strategy 4 appear visually more constrained and anatomically structured, suggesting that the explicit modelling of healthy tissue may have helped reduce over-segmentation.

The third example highlights the effect of introducing the "others" class. While strategies 1 and 2 mainly focused on the major pathological findings, strategy 3 additionally predicted extensive regions belonging to the "others" category. This behaviour is particularly visible in regions surrounding the annotated findings. In strategy 4, part of these regions was instead assigned to the healthy class. As observed in the previous examples, strategies 2, 3, and 4 also tended to produce larger segmentation masks for consolidation and ground-glass opacities, extending beyond the annotated GT regions. Nevertheless, all strategies were able to detect the consolidation region annotated in the GT, although with different spatial extents.

The final example focuses on a region annotated as healthy tissue. Since strategies 1 and 2 did not include a specific healthy class, this region was not expected to be segmented as an independent category. This was observed in strategy 1, which did not assign the annotated healthy region to any pathological finding. However, strategy 2 incorrectly assigned parts of this region to fibrosis. In strategy 3, the same region was segmented as "others", which is consistent with the label definition used in this strategy, where healthy tissue was included in the "others" class. In strategy 4, the model was able to assign most of the region to the healthy class, although part of it was still confused with nodules.

Overall, the qualitative results suggest that strategy 4 produced the most coherent segmentation maps among the evaluated strategies, particularly due to its ability to distinguish healthy tissue from pathological findings and to reduce some of the over-segmentation observed in strategy 2. However, a specific pattern can be observed near the outer contour of the lung masks. In several examples, the GT annotations do not cover peripheral regions that still belong to the lung parenchyma, especially near the lung boundaries. In strategies 2 and 3, these regions often

remained classified as background or were not assigned to any foreground class. In contrast, strategy 4 frequently assigned these peripheral areas to the "others" class, particularly in the last two examples. This behaviour is consistent with the anatomical consistency constraint used in this strategy, which discouraged background predictions inside the lung mask and encouraged lung pixels to be assigned to one of the available non-background classes. Although this can improve anatomical consistency, it may also lead to uncertain peripheral regions being classified as "others" when the available annotations do not provide enough information to determine their true class.

4.2.2 Network Configurations and Variants

After selecting the most suitable supervision strategy, the next set of experiments evaluated the effect of different network configurations on segmentation performance. Three setups were compared: configuration 1, corresponding to the baseline 2D nnU-Net with a 512×512 patch size, configuration 2, corresponding to the standard U-Net with a reduced patch size of 320×320 , and configuration 3, corresponding to the residual encoder U-Net using the same reduced-patch setting.

Table 4.6 presents the DSC obtained for each network configuration, with the values reported separately for each class. The remaining evaluation metrics, namely recall, precision and specificity, are provided in Appendix A.1 for the best performing configuration.

Table 4.6: Segmentation performance of different configurations measured by the DSC. Values are reported as mean $\pm$ standard deviation (%) across five data splits. The "Config" refers to the experimental configuration used in each setup. The cysts class is omitted, as it achieved $0.00 \pm 0.00\%$ across all experiments.

Config	Fibrosis	Nodules	Ground glass	Consolidation	Emphysema	Others	Healthy	Average
1	81.35 $\pm$ 4.72	53.64 $\pm$ 8.69	44.84 $\pm$ 10.28	22.95 $\pm$ 15.66	52.89 $\pm$ 14.95	0.01 $\pm$ 0.02	30.66 $\pm$ 16.20	35.79 $\pm$ 26.27
2	84.84$\pm$2.43	56.61$\pm$5.43	51.23 $\pm$ 9.14	24.92$\pm$8.31	64.71$\pm$22.13	0.01 $\pm$ 0.02	62.63$\pm$5.11	43.12$\pm$29.34
3	83.96 $\pm$ 2.36	55.30 $\pm$ 5.80	60.33$\pm$5.33	10.97 $\pm$ 4.76	61.61 $\pm$ 22.07	0.01 $\pm$ 0.02	62.13 $\pm$ 14.98	41.79 $\pm$ 30.72

The results presented in Table 4.6 demonstrate that the reduced-patch configurations improved the overall segmentation performance compared with the baseline configuration. Configuration 1 obtained an average DSC of 35.79%, whereas configurations 2 and 3 achieved higher average values of 43.12% and 41.79%, respectively. This indicates that reducing the patch size to 320×320 , had a positive effect on segmentation performance. Although the baseline configuration provided the model with the full spatial context of each axial slice, the reduced-patch setting may have allowed more effective sampling of annotated regions during training, which is particularly relevant given the sparse and localized nature of the annotations.

Configuration 2 achieved the highest average DSC and obtained the best results for most classes, including fibrosis, nodules, consolidation, emphysema, and healthy tissue. The improvement was especially marked for the healthy class, where the DSC increased from 30.66% to 62.63%. This suggests that the reduced-patch configuration improved the model's ability to learn

normal lung appearance, possibly because smaller patches increased the relative contribution of local texture information and reduced the dominance of large background regions. The remaining improvements were more moderate and varied across classes.

The comparison between configurations 2 and 3 suggest that the residual encoder did not lead to a consistent improvement over the standard U-Net architecture. Configuration 3 achieved the best result for ground-glass opacities, with a DSC of 60.33%, representing a clear improvement compared with both configuration 1 and 2. This may indicate that the residual encoder helped capture the more subtle texture patterns associated with ground-glass opacities. However, this improvement was not observed across all classes. In particular, consolidation decreased, and smaller reductions were also observed for fibrosis, nodules, emphysema, and healthy tissue. As a result, the average DSC of configuration 3 remained slightly lower than that of configuration 2.

The performance of the cysts and "others" classes remained very limited across all configurations. Cysts obtained a DSC of 0.00% in all variants, confirming that changing the network configuration was not sufficient to overcome the extremely low representation of this class. Similarly, the "others" class remained close to zero in all configurations.

Overall, these results suggest that the reduced-patch standard U-Net configuration provided the best trade-off between local detail, batch size, and segmentation performance. Therefore, configuration 2 was considered the most effective network configuration among the evaluated setups.

To further analyse the behaviour of the best-performing network configuration, Figure 4.4 presents the pixel-wise confusion matrix obtained for configuration 2. The matrix was normalized by the true class, meaning that each row represents the percentage distribution of predicted labels for pixels belonging to a given GT class. Thus, diagonal values represent correctly classified pixels, while remaining values indicate inter-class confusion.

Figure 4.4 confirms the trends observed in the DSC results for configuration 2. Fibrosis was the best recognized pathological finding, with 83.35% of fibrosis pixels correctly classified. The not lung class also showed a very high correct classification rate, although this result should be interpreted with caution due to the much larger number of pixels belonging to this class.

Healthy tissue was also mostly well identified, with 81.04% of healthy pixels correctly classified. However, 15.55% of healthy pixels were classified as nodules, while 27.34% of nodule pixels were classified as healthy. This mutual confusion suggests that the model still struggled to separate subtle nodular findings from normal lung texture. This may also be partly influenced by the annotation characteristics of the dataset, since some nodule segmentations encompass neighbouring healthy tissue due to the coarse delineation of lesion boundaries. Consequently, the distinction between healthy tissue and nodular findings is not always clearly represented in the GT annotations, which contribute to the observed confusion between these two classes.

Ground-glass opacities showed a more distributed pattern of errors. Although 49.11% of ground-glass pixels were correctly classified, relevant confusion was observed with fibrosis, healthy tissue, and nodules, which may reflect the diffuse appearance of this pattern and its overlap with



Figure 4.4: Pixel-wise confusion matrix normalized by true label for the best performing model configuration.

other parenchymal abnormalities. Consolidation and emphysema also showed confusion with both normal and abnormal lung tissue classes.

The poorest results were observed for cysts and the "others" class. Cysts were not correctly identified, in agreement with the DSC results, while "others" was mostly confused with healthy tissue and nodules. Overall, the confusion matrix indicates that the main remaining errors were associated with visual similarities between classes and class imbalance.

To contextualize the supervised results obtained, the performance of the best configuration was compared with selected studies from the literature. This comparison should be interpreted with caution, since the studies differ in terms of dataset, annotation protocol, target classes, network architecture, and evaluation strategy. Therefore, only classes with a clear correspondence between studies were considered.

Table 4.7 summarises the comparison between configuration 2 and the selected studies. Ham et al. [52] evaluated nnU-Net on private datasets for ILD findings, including normal lung parenchyma, ground-glass opacity, consolidation, and emphysema. Hernández-Vázquez et al. used only the

same MedGIFT ILD database [58], which makes it the most directly comparable study, although their work focused only on fibrotic region segmentation. Thillai et al. [48] and Lai et al. [51] are also included for broader context, although both used different datasets and lesion definitions.

Table 4.7: Comparison of segmentation performance measured by DSC (%) between the best-performing configuration proposed in this dissertation and selected state-of-the-art studies.

Setup	Fibrosis	Ground glass	Consolidation	Emphysema	Healthy
Configuration 2	84.84	51.23	24.92	64.71	62.63
Ham et al. [52]	–	30.80	39.60	12.70	74.40
Hernández-Vázquez et al. [44]	78.15	–	–	–	–
Thillai et al. [48]	90.00	–	–	–	–
Lai et al. [51]	–	78.40	86.20	82.50	–

As shown in Table 4.7, configuration 2 achieved competitive results for overlapping classes. For fibrosis, it obtained a DSC of 84.84%, outperforming the value reported by Hernández-Vázquez et al. [44], which is particularly relevant given that both studies used the same database and evaluated process using the available labelled regions. A higher fibrosis DSC was reported by Thillai et al. [48], although this result was obtained using a private dataset and a different annotation protocol.

Compared with Ham et al. [52], configuration 2 performed better for ground-glass opacities and emphysema, but worse for healthy tissue and consolidation. Lai et al. [51] reported higher DSC values for ground-glass opacities, consolidation, and emphysema. However, these results were obtained using a private dataset, where differences in class prevalence and annotation completeness may have contributed to improved segmentation performance.

Overall, this comparison should be interpreted as contextual rather than as a direct benchmark, since differences in datasets, annotation density, and evaluation protocols make direct comparisons difficult. Nevertheless, considering the use of a public database with sparse and incomplete annotations, the supervised model achieved results consistent with the literature. The performance obtained for fibrosis is particularly encouraging, especially when compared with previous work using the same database.

4.3 Summary

This chapter presented the supervised segmentation experiments developed as the baseline stage of this dissertation. A 2D nnU-Net framework was used for multi-class ILD findings segmentation in HRCT images, and two main aspects were investigated: the supervision strategy used to handle sparse annotations and the network configuration used for training.

The results showed that the way unlabelled pixels were treated had a substantial impact on segmentation performance. Treating all non-annotated pixels as background led to poor results, while excluding unlabelled pixels from the loss computation considerably improved performance.

Further separating healthy tissue and other non-target findings provided additional benefits, with strategy 4 achieving the best average DSC among the major pathological findings. This indicated that explicitly modelling healthy tissue and avoiding background predictions inside the lung mask were useful for this partially labelled setting.

The comparison between network configurations showed that the reduced-patch standard U-Net configuration achieved the best overall performance. This setup improved the average DSC compared with the baseline 512×512 configuration and outperformed the residual encoder variant in most classes. Nevertheless, the results also highlighted persistent limitations, particularly for rare classes such as cysts and "others", and relevant confusion between visually similar patterns such as nodules and healthy tissue.

Overall, these supervised experiments established a reference performance for ILD findings segmentation using the available labelled data. At the same time, they highlighted the limitations imposed by sparse annotations and strong class imbalance. These limitations motivate the next stage of this dissertation, where semi-supervised strategies are explored to make use of unlabelled lung regions and improve the model beyond what can be achieved with SL alone.

Chapter 5

Semi-Supervised Learning for ILD Findings Segmentation

This chapter presents the SSL approach developed to exploit unlabelled data for ILD findings segmentation in HRCT images. First, the additional data used in this stage are described and the corresponding data splitting strategy. The chapter then introduces the semi-supervised methodology, with particular focus on pseudo-labelling, entropy minimization, consistency regularization, and FixMatch-based approaches. Finally, the chapter presents and discusses the quantitative and qualitative results obtained with the semi-supervised experiments, including a direct comparison with the supervised baselines.

5.1 Methodology

In the second phase of this dissertation, SSL strategies were investigated for the segmentation of ILD findings in HRCT images. In the supervised setting, the optimisation was driven only by pixels with reliable GT annotations, while unlabelled pixels were either ignored or treated according to the supervision strategy being evaluated. In contrast, SSL explicitly incorporates these unlabelled regions to provide additional training signal, since many pixels inside the lung region do not have a manual label and discarding them entirely may underuse potentially informative image content. More specifically, the labelled pixels continued to contribute to the supervised segmentation loss, while unlabelled pixels were incorporated through SSL methods.

The semi-supervised methodology was built upon the supervised nnU-Net pipeline described in Chapter 4. The same 2D nnU-Net architecture, preprocessing procedure, data augmentation strategy, optimiser, and evaluation metrics were maintained. The main difference was the way unlabelled information was incorporated during training. In all semi-supervised experiments, the supervised component of the loss was kept unchanged and followed the fourth supervision strategy described in Section 4.1.4. The overall training objective was defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{sup} + \lambda_u \mathcal{L}_{unsup}, \quad (5.1)$$

where $\mathcal{L}_{sup}$ corresponds to the supervised loss, $\mathcal{L}_{unsup}$ denotes the unsupervised loss term specific to each SSL strategy, and λ_u controls the overall contribution of the unsupervised component. The main difference between the investigated methods was therefore the definition of $\mathcal{L}_{unsup}$, which was approached through two complementary strategies for incorporating unlabelled data.

The first comprises SSL methods that incorporate unlabelled signal directly during model optimisation, where $\mathcal{L}_{unsup}$ is computed and backpropagated at each training step. The second consists of an ensemble-based pseudo-label generation pipeline, which exploits the complementary predictions of multiple trained models to expand the available supervision prior to retraining. For the training-time SSL methods, only the baseline nnU-Net configuration was used, as it represents the standard nnU-Net setting and corresponds to configuration 1 evaluated in the supervised chapter. In contrast, the ensemble-based pseudo-labelling pipeline incorporated all three network configurations previously evaluated, in order to exploit their complementary behaviour when generating pseudo-labels. For the final retraining stages of this pipeline, only the configuration 2 was carried forward, as it had achieved the best performance in the supervised setting.

5.1.1 Data

In the semi-supervised setting, the MedGIFT ILD database [58] was also used as the principal data source, following the same data organisation and class definitions described in subsection 4.1.1.

In addition to the MedGIFT data, some experiments incorporated an additional private ILD dataset acquired at São João Hospital (HSJ). Unlike MedGIFT, this dataset did not include pixel-level annotations for ILD findings and was therefore used exclusively as unlabelled data. Instead, annotated radiology reports were available for a subset of the examinations, providing structured information regarding ILD findings, their anatomical locations, and their reported presence or absence.

The private dataset comprised 538 HRCT studies acquired from 249 different patients, making it substantially larger than the MedGIFT dataset in terms of number of scans. Each study consists of a volumetric HRCT acquisition represented by axial slices stored as 512×512 pixel matrices, with an average of approximately 30 slices per study. The scans were acquired with a slice thickness of 1mm and an inter-slice spacing of approximately 10mm.

For the semi-supervised experiments, the same data splitting strategy adopted in the supervised setting was maintained. Unlike the data structure assumed by many SSL methods in the literature, where training data are divided into fully labelled and fully unlabelled images, the MedGIFT database [58] contains sparse pixel-level annotations. In this dataset, each image may contain labelled pixels corresponding to annotated findings, while other pixels inside the lung region remain unlabelled. Therefore, the semi-supervised methods were designed to exploit the unlabelled regions within the same training images while preserving the validation and test splits.

For the experiments that incorporated the private HSJ dataset, no stratification according to pathological classes was required due to the absence of manual annotations. Nevertheless, patient-level separation was preserved. These additional studies were distributed across the five cross-validation iterations and used exclusively during training, never contributing to validation

or testing. On average, each iteration incorporated approximately 50 additional patients and 108 unlabelled HRCT studies.

5.1.2 SSL via Training-Time Signal Integration

This section details the training-time SSL methods introduced above. Four such methods were investigated within this category: dynamic pseudo-labelling, entropy minimization, consistency regularization, and a FixMatch-based approach, described in the following subsections.

5.1.2.1 Dynamic Pseudo-Labelling

In segmentation tasks, pseudo-labelling can be applied at the pixel level by assigning class labels to unlabelled pixels. In the dynamic pseudo-label approach, pseudo-labels were generated during training from the current model predictions and incorporated directly into the loss computation whenever the prediction confidence exceeded predefined thresholds. This provided additional supervision for pixels assigned to the ignore label, which corresponded to unlabelled regions inside the lung parenchyma, and for this reason the background class was excluded from pseudo-label generation.

Let Ω_{ign} denote the set of pixels assigned to the ignore label, and let C_L denote the set of lung-related classes, including the major pathological findings, the healthy class, and the "others" class. For each ignored pixel i of image b in the current training batch, the network outputs a probability $p_{b,i,c}$ for each class c after the softmax operation. The predicted probabilities were first restricted to the lung-related classes and renormalised as:

$$\tilde{p}_{b,i,c} = \frac{p_{b,i,c}}{\sum_{k \in C_L} p_{b,i,k}}, \quad c \in C_L \quad (5.2)$$

where $\tilde{p}_{b,i,c}$ represents the probability assigned to class c after excluding the background class and renormalising the lung-related classes. For each ignored pixel, the pseudo-label was assigned as the lung-related class with the highest renormalised probability. Formally, the pseudo-label $\hat{y}_{b,i}$ and its associated confidence score $q_{b,i}$ were defined as:

$$\hat{y}_{b,i} = \arg \max_{c \in C_L} \tilde{p}_{b,i,c}, \quad (5.3)$$

$$q_{b,i} = \max_{c \in C_L} \tilde{p}_{b,i,c} \quad (5.4)$$

To determine whether a pseudo-label was sufficiently reliable, different confidence thresholds were used for each class instead of a single global threshold. This was motivated by the strong class imbalance of the dataset, since different classes may produce different confidence distributions. Using the same threshold for all classes could favour more frequent or easier classes while discarding useful pseudo-labels from rarer findings.

For each class c , two thresholds were considered: a lower threshold τ_c^{low} and a higher threshold τ_c^{high} . These thresholds were used to define a confidence weight $\alpha_{b,i}$ for each pseudo-labelled pixel:

$$\alpha_{b,i} = \begin{cases} 1, & q_{b,i} \geq \tau_{\hat{y}_{b,i}}^{high}, \\ q_{b,i}, & \tau_{\hat{y}_{b,i}}^{low} \leq q_{b,i} < \tau_{\hat{y}_{b,i}}^{high}, \\ 0, & q_{b,i} < \tau_{\hat{y}_{b,i}}^{low}. \end{cases} \quad (5.5)$$

where $\alpha_{b,i}$ represents the contribution of the pseudo-labelled pixel to the training objective. Pseudo-labels with confidence above the higher threshold contributed with full weight, pseudo-labels with confidence between the lower and higher thresholds contributed proportionally to their confidence, and pseudo-labels below the lower threshold were ignored.

Let $y_{b,i}$ denote the original training target and $y_{b,i}^*$ the target after pseudo-label insertion. Ignored pixels with $\alpha_{b,i} > 0$ were temporarily replaced by their pseudo-label during the current training iteration, while ignored pixels with $\alpha_{b,i} = 0$ remained assigned to the ignore label.

The standard nnU-Net compound loss was then calculated using the predicted probabilities, the combined target, and the confidence weights. Manual annotations were assigned weight 1, pseudo-labelled pixels contributed according to $\alpha_{b,i}$, and ignored pixels without an accepted pseudo-label did not contribute to the loss. Therefore, this method did not require a separate unsupervised loss term, as pseudo-labelled pixels were incorporated directly into the supervised loss computation:

$$\mathcal{L}_{total} = \mathcal{L}(p, y^*; \alpha), \quad (5.6)$$

where p denotes the model predictions, y^* denotes the combined target after pseudo-label insertion, and α represents the pixel-wise confidence weights.

Two variants were evaluated for estimating the thresholds associated with each class. In the first variant, thresholds were estimated from the confidence distribution obtained on labelled validation pixels. Since confidence values in this setting were generally high, the 10th and 30th percentiles were used as the lower and higher thresholds, respectively. In the second variant, thresholds were computed from the confidence distribution of pseudo-labels generated on ignored pixels during training, using the 50th and 80th percentiles as the lower and higher thresholds, respectively. In both variants, confidence values were accumulated during each epoch and used to update the class-specific thresholds at the end of that epoch. The updated thresholds were then applied in the following epoch, allowing the pseudo-label selection criterion to adapt progressively as the model evolved.

5.1.2.2 Entropy Minimization

Entropy minimization is an SSL method that encourages the model to produce confident predictions for unlabelled data. Originally proposed by Grandvalet and Bengio [104], this approach penalizes high-entropy class probability distributions, encouraging decision boundaries to pass

through low-density regions of the data distribution. In the context of segmentation, entropy minimization can be applied at the pixel level by reducing prediction uncertainty in unlabelled regions. This strategy has also been explored for ILD segmentation by Anthimopoulos et al. [55].

The method was applied only to pixels assigned to the ignore label. Since these pixels were known to belong to the lung region, the entropy term was computed only over the lung-related classes, excluding the background class. The probabilities were renormalised over these classes as defined in Equation 5.2.

The entropy for each ignored pixel was then computed as:

$$H_{b,i} = - \sum_{c \in C_L} \tilde{p}_{b,i,c} \log(\tilde{p}_{b,i,c}) \quad (5.7)$$

Let Ω_{ign} denote the set of ignored pixels in the current training batch. The mean entropy over these pixels was first computed as:

$$\mathcal{L}_{ent} = \frac{1}{|\Omega_{ign}|} \sum_{(b,i) \in \Omega_{ign}} H_{b,i} \quad (5.8)$$

To account for the proportion of unlabelled and labelled foreground pixels in each batch, this entropy term was weighted by a batch-dependent factor. Let Ω_{lab}^{fg} denote the set of labelled foreground pixels, excluding both background and ignored pixels. The weighting factor was defined as:

$$w_u = \frac{|\Omega_{ign}|}{|\Omega_{lab}^{fg}|} \quad (5.9)$$

The unsupervised component for entropy minimization was therefore given by:

$$\mathcal{L}_{unsup} = w_u \mathcal{L}_{ent} \quad (5.10)$$

This term was then used as $\mathcal{L}_{unsup}$ defined in Equation 5.1. In this strategy, λ_u was set to 0.15.

Two experimental settings were evaluated for this method. In the first setting, entropy minimization was applied using only the MedGIFT database [58], exploiting the unlabelled pixels already present within the sparsely annotated training images. In the second setting, the same method was extended by incorporating the additional unlabelled HRCT studies from HSJ, with the objective of assessing whether a larger amount of unlabelled data could improve the semi-supervised training process. In this setting, each training batch was composed of images from both datasets: eight images were sampled from MedGIFT, while the remaining four images were sampled from the unlabelled HSJ dataset.

Since the HSJ dataset did not include lung masks, lung parenchyma segmentation was first performed using the automated method proposed by Hofmanninger et al. [105], which generated binary lung masks for each HRCT volume. These masks were then used to define the anatomical regions used during training: pixels outside the lungs were assigned to the background class, while

pixels inside the lung parenchyma were treated as unlabelled pixels and therefore contributed only through the entropy-minimization term.

5.1.2.3 Consistency Regularization

Consistency regularization was implemented by comparing the predictions obtained from two differently augmented versions of the same training image. The first corresponded to a weakly augmented image, obtained using the standard nnU-Net data augmentation pipeline, while the second corresponded to a strongly augmented version of the same image. The strong branch applied substantially more aggressive appearance transformations, including Gaussian noise, brightness and contrast adjustments, gamma transformations, Gaussian blurring, simulated low-resolution effects through pixelation, intensity quantisation, image sharpening, and block-wise compression artefacts. Figure 5.1 presents an overview of the consistency regularization framework.

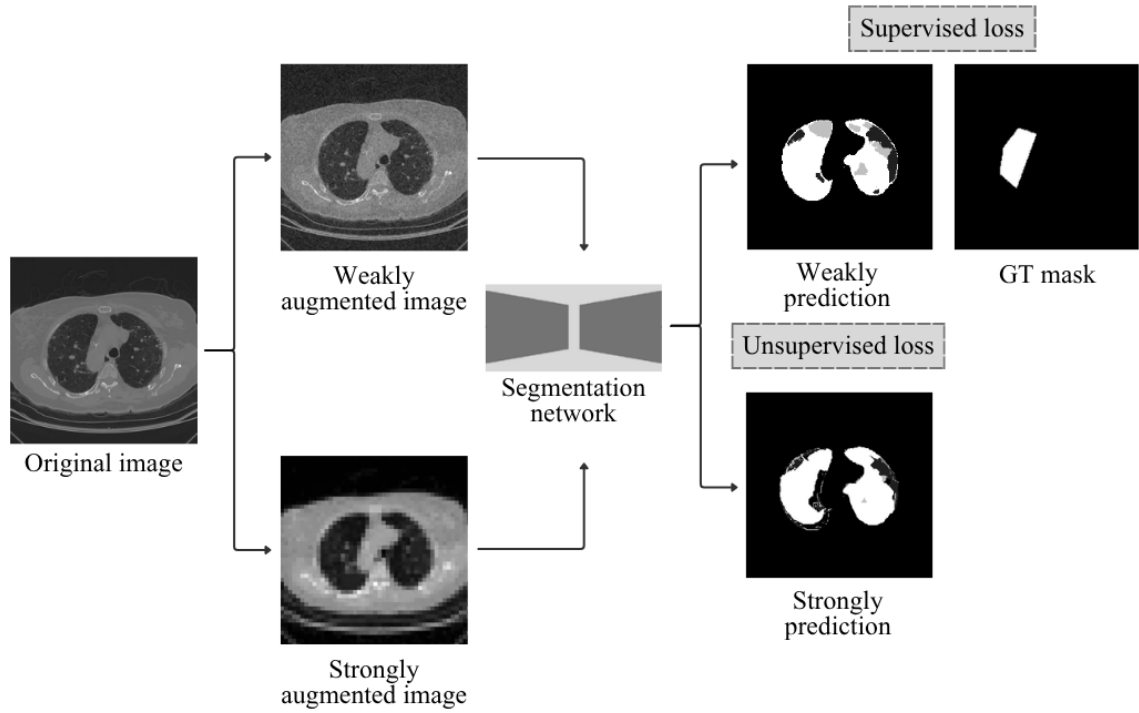


Figure 5.1: Overview of the consistency regularization framework adopted.

Let $p_{b,i,c}^w$ and $p_{b,i,c}^s$ denote the predicted probabilities for class c at pixel i of image b , obtained from the weakly and strongly augmented inputs, respectively. The prediction from the weak branch was treated as a fixed target.

The consistency term was computed only over ignored pixels for which the weak prediction did not correspond to the background class. This avoided enforcing consistency in ignored regions where the model predicted non-lung background, focusing the unsupervised signal on pixels more likely to belong to classes within the lung parenchyma. Let Ω_{cons} denote this set of pixels:

$$\Omega_{cons} = \{(b, i) \in \Omega_{ign} \mid \arg \max_c p_{b,i,c}^w \neq 0\}, \quad (5.11)$$

where class 0 corresponds to the background class.

The unsupervised consistency loss was defined as the mean squared difference between the weak and strong predicted probability distributions over the selected pixels:

$$\mathcal{L}_{unsup} = \frac{1}{|\Omega_{cons}|} \sum_{(b,i) \in \Omega_{cons}} \sum_{c \in C} (p_{b,i,c}^w - p_{b,i,c}^s)^2 \quad (5.12)$$

This term was then used as $\mathcal{L}_{unsup}$ in Equation 5.1, with $\lambda_u = 0.1$ to impose consistency as a regularization term without allowing it to dominate the supervised loss. This encouraged the model to produce similar predictions for weakly and strongly augmented versions of the same image, improving prediction stability in unlabelled lung regions.

5.1.2.4 FixMatch

FixMatch was implemented as a hybrid SSL method that combined the pseudo-labelling and consistency regularization principles described in the previous sections. In this method, pseudo-labels were generated from the weakly augmented branch and then used as targets for the predictions obtained from the strongly augmented branch. Thus, the model was encouraged to produce consistent predictions under stronger perturbations, but only for unlabelled pixels whose pseudo-labels were considered reliable.

For each training batch, the weak branch was first processed by the network and used to compute the supervised loss on labelled pixels. The same weak prediction was also used to generate pseudo-labels for pixels assigned to the ignore label. As in the dynamic pseudo-labelling method, pseudo-label generation was restricted to lung-related classes using the renormalised probabilities, and background predictions inside ignored lung regions were discarded.

The accepted pseudo-labels were then used as targets for the strong branch. The strong branch received the same input images after stronger non-geometric perturbations, following the augmentation setting described in previous subsection. Since no geometric transformations were applied, the weak pseudo-labels and the strong predictions remained spatially aligned, allowing the unsupervised loss to be computed pixel-wise. The overall workflow of the FixMatch framework adopted is illustrated in Figure 5.2.

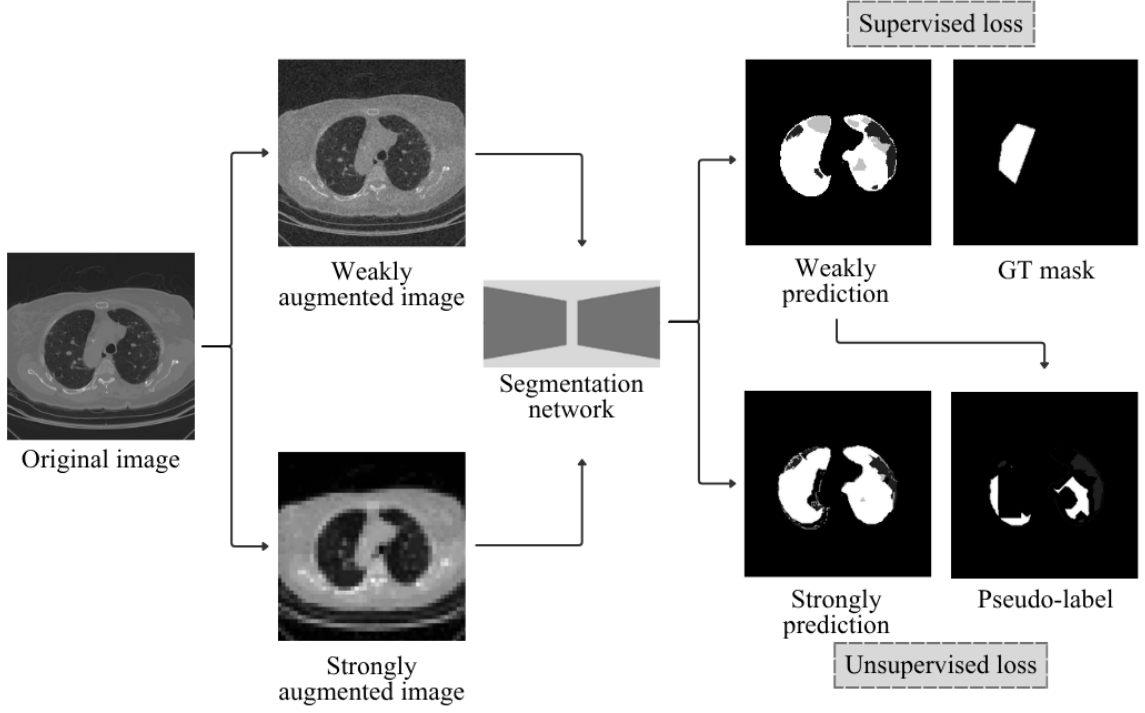


Figure 5.2: Overview of the FixMatch framework adopted.

Let $\hat{y}_{b,i}^w$ denote the pseudo-label generated from the weak prediction for pixel i of image b , and let $\alpha_{b,i}^w$ denote its confidence weight. The unsupervised loss was computed by comparing the strong prediction with the weak pseudo-labels. This term combined weighted cross-entropy and weighted soft dice loss:

$$\mathcal{L}_{unsup} = \mathcal{L}_{CE} + \mathcal{L}_D, \quad (5.13)$$

where both components were computed only over ignored pixels with accepted pseudo-labels. The confidence weight $\alpha_{b,i}^w$ controlled the contribution of each pseudo-labelled pixel, allowing greater influence to more confident pseudo-labels.

Three variants of FixMatch were evaluated, differing in how pseudo-label confidence was converted into acceptance masks or weights. In the first variant, class-specific percentile thresholds were used, following the same principle as in the dynamic pseudo-labelling strategy. Thresholds were estimated from the confidence distribution of pseudo-labels generated on ignored pixels, using the 50th and 80th percentiles as the lower and higher thresholds, respectively. Pseudo-labels above the higher threshold contributed with full weight, pseudo-labels between both thresholds contributed proportionally to their confidence, and pseudo-labels below the lower threshold were ignored. Unlike the original FixMatch formulation [94], which relies on a single fixed confidence threshold applied uniformly across all classes, this variant adapted the selection criterion to the per-class confidence distribution observed during training.

The second variant followed the FreeMatch principle [106], replacing fixed thresholds with self-adaptive thresholds updated during training. In this variant, two quantities were estimated

using an EMA decay factor set to $m = 0.999$: a global threshold τ_t , reflecting the general confidence of the model on unlabelled pixels, and a class-specific estimate $r_t(c)$, reflecting the average predicted probability of each class. For a confidence score $q_{b,i} = \max_c p_{b,i,c}$, these quantities were updated as:

$$\tau_t = m\tau_{t-1} + (1-m) \frac{1}{|\Omega_{ign}|} \sum_{(b,i) \in \Omega_{ign}} q_{b,i}, \quad (5.14)$$

$$r_t(c) = mr_{t-1}(c) + (1-m) \frac{1}{|\Omega_{ign}|} \sum_{(b,i) \in \Omega_{ign}} p_{b,i,c}. \quad (5.15)$$

The final threshold for class c was obtained by modulating the global threshold with the normalised local estimate:

$$\tau_t(c) = \frac{r_t(c)}{\max_{k \in C} r_t(k)} \tau_t. \quad (5.16)$$

Thus, classes predicted with consistently high confidence were subject to stricter acceptance criteria, while rarer or harder classes received lower thresholds and could be accepted more easily into the pseudo-label set.

The third variant was based on SoftMatch [107]. While FreeMatch adapts the acceptance threshold, it still makes a hard decision by either accepting or discarding each pseudo-label. SoftMatch instead assigns continuous weights according to the confidence distribution observed during training. The running mean μ_t and variance σ_t^2 of the confidence values were estimated over ignored pixels using an EMA. For each pseudo-label, the confidence weight was then defined as:

$$\alpha_{b,i} = \begin{cases} 1, & q_{b,i} \geq \mu_t, \\ \exp\left(-\frac{(q_{b,i} - \mu_t)^2}{2\sigma_t^2}\right), & q_{b,i} < \mu_t. \end{cases} \quad (5.17)$$

Pseudo-labels with confidence above the running mean contributed with full weight, whereas lower-confidence pseudo-labels contributed according to the gaussian weighting function. This provided a softer alternative to threshold-based selection, allowing uncertain pseudo-labels to contribute proportionally to their relative confidence instead of being completely discarded.

The three variants were evaluated to analyse how different pseudo-label acceptance mechanisms affect semi-supervised training under sparse and highly imbalanced annotations. Since the labelled set covers only a small fraction of the available pixels and class frequencies are heavily skewed, the threshold or weighting strategy used to select pseudo-labels is expected to have a non-trivial impact on training stability and class coverage. Comparing these three strategies under the same experimental conditions allows for a more informed understanding of the trade-off between pseudo-label quality and quantity in this specific ILD findings annotation setting.

The resulting unsupervised term was used as $\mathcal{L}_{unsup}$ in Equation 5.1. In this strategy, λ_u was set to 0.15. Therefore, the supervised component optimised the annotated pixels from the weak branch, while the FixMatch component encouraged the model to produce predictions on

strongly augmented images that were consistent with reliable pseudo-labels generated from the weak branch.

5.1.3 SSL via Ensemble Pseudo-Label Generation

In contrast to the dynamic approach, the pseudo-labels generated in this strategy remained fixed throughout retraining and were not updated as the model parameters evolved. For the ensemble-based pseudo-labelling strategy, pseudo-labels were generated prior to retraining using the predictions of the three previously trained network configurations described in subsection 4.1.5. These configurations corresponded to the baseline nnU-Net model, the reduced-patch and the residual encoder nnU-Net variants. The objective was to exploit the complementary behaviour of models trained with different architectural and input-size configurations, generating pseudo-labels only when there was sufficient confidence and inter-model agreement.

For each network configuration, the five checkpoints obtained from the 5-fold cross-validation performed within the training/validation pool were applied to each case, and the resulting class-probability maps were averaged pixel-wise and class-wise to produce one aggregated probability map per configuration. Pseudo-label generation was restricted to pixels assigned to the ignore label in the original masks and background predictions were not promoted to pseudo-labels. To account for differences in confidence distributions across classes and model configurations, class-specific thresholds were estimated separately for each model. For each class, the 80th percentile of the confidence values obtained on ignored pixels was used as the confidence threshold. Therefore, only the 20% most confident predictions of each class and configuration were considered eligible for pseudo-labelling.

After applying the thresholds, the final pseudo-label assignment was determined through majority voting across the three network configurations. A pixel was promoted to a class only if at least two models produced a valid prediction for that same class. Pixels that did not satisfy this majority criterion remained assigned to the ignore label. This consensus-based selection was used to reduce the likelihood of propagating incorrect pseudo-labels, since agreement across multiple models increases the reliability of the generated annotations.

The ensemble-based pseudo-labelling procedure was performed in two stages, with the overall workflow illustrated in Figure 5.3. In the first stage, the three fully supervised models were used to generate pseudo-labels only for the MedGIFT training data. These pseudo-labelled masks were then incorporated into the training dataset together with the original GT annotations, and the three network configurations were retrained using this expanded supervision.

In the second stage, the retrained models were used as a new ensemble to generate a second set of pseudo-labels. This stage was applied not only to the MedGIFT training data but also to the additional unlabelled HRCT studies from HSJ. For MedGIFT, the pseudo-labels generated in the first stage were retained, and the second stage was used to promote additional ignored pixels when the ensemble predictions satisfied the confidence and majority voting criteria. For the HSJ dataset, pseudo-labels were generated inside the automatically segmented lung regions. Finally, a subsequent retraining stage was performed using the original GT annotations, the first and second

sets of MedGIFT pseudo-labels, and the pseudo-labels generated for the HSJ dataset. This final retraining was carried out using only the reduced-patch nnU-Net configuration, corresponding to configuration 2, allowing the model to learn from both the manually annotated data and the expanded set of pseudo-labelled regions.

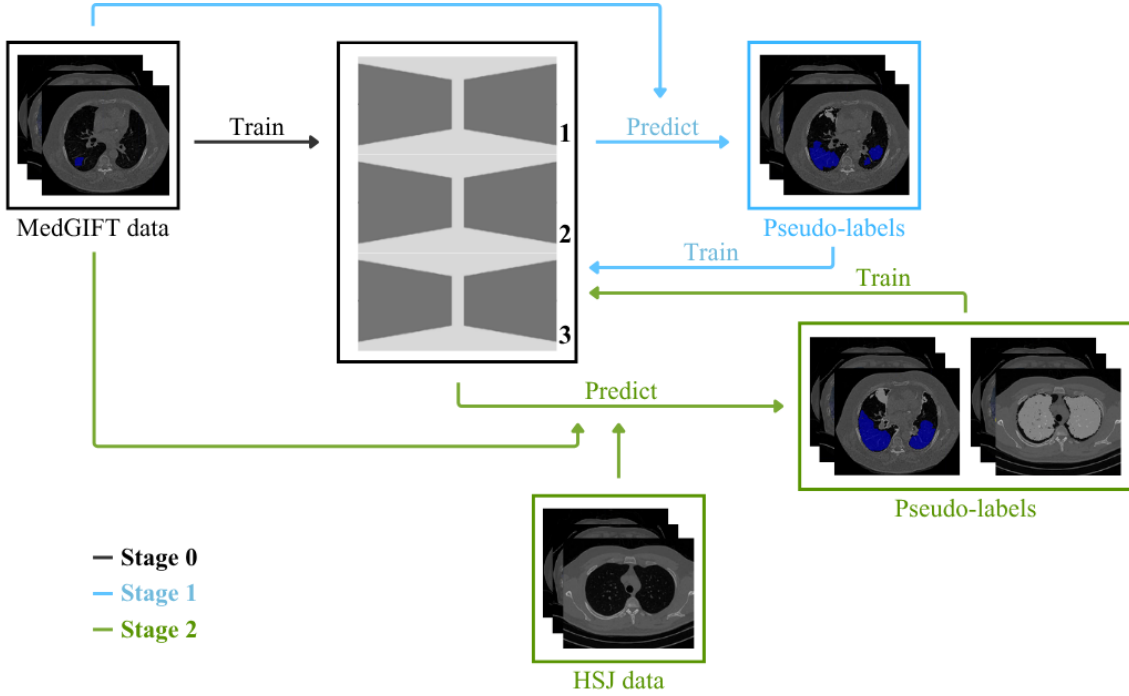


Figure 5.3: Overview of the ensemble-based pseudo-label generation pipeline. Stage 0 corresponds to the original supervised models, while stages 1 and 2 represent the first and second rounds of pseudo-label generation and retraining, respectively, applied to the MedGIFT and HSJ datasets.

To further refine the pseudo-labels generated for the HSJ dataset, a multi-label text classification approach previously developed within the laboratory was employed. This approach automatically extracts structured clinical information from Portuguese radiology reports by identifying the ILD findings mentioned in each examination and determining their negation status. Accurate extraction of this information is inherently challenging, as clinicians frequently employ varied terminology and introduce typographical errors when describing identical findings. Furthermore, contextual relationships between terms must be considered to avoid extracting clinically redundant information. The framework is based on the BERT model MediAlbertina 1.5B [108]. Contextualised token-level representations are generated for both the radiology report and a predefined set of 55 ILD findings and are subsequently refined through an adaptation network to improve the separability of finding embeddings. Finding detection is then performed by computing the cosine similarity between report-token embeddings and finding embeddings, producing a similarity map from which reported findings are identified through noise suppression and threshold-based classification. The negation detection component was evaluated on a validation subset, achieving an accuracy of 90.50%, a precision of 92.45%, a recall of 74.24%, and an F1-score of 82.35%.

For each examination, applying this approach to HSJ radiology reports results in one of three possible outcomes for each finding. The finding may be explicitly detected with the negation set to false, indicating that it was reported as present. Alternatively, it may be explicitly detected with the negation set to true, indicating that it was reported as absent. Finally, it may not be detected at all, indicating that the finding was simply not mentioned in the report, which does not necessarily imply that it is absent from the examination.

Based on these outcomes, the predictions were used to filter the pseudo-labels generated for the HSJ dataset. Filtering was applied only to the pathological classes with direct correspondence to the detected report findings, namely fibrosis, nodules, ground-glass opacities, consolidation, cysts, and emphysema. The "others" and healthy classes were not filtered and were therefore kept as generated by the ensemble models. Two filtering strategies were evaluated. The first strategy retained pseudo-labels only for the pathological classes detected with negation set to false, keeping only findings reported as present. The second strategy removed pseudo-labels corresponding to findings detected with negation set to true, suppressing findings explicitly reported as absent while preserving the remaining pseudo-labels. After filtering, the resulting HSJ pseudo-labels were combined with the MedGIFT GT annotations and pseudo-labels, following the same retraining procedure used in the second stage.

5.2 Results and Discussion

This section presents and discusses the results obtained with the semi-supervised segmentation experiments described in the previous section. As in the supervised chapter, the evaluation includes quantitative performance metrics, as well as qualitative visual examples of the predicted segmentation masks.

The section is organised into two parts. Subsection 5.2.1 evaluates SSL methods that incorporate unlabelled signal directly during model optimisation, including dynamic pseudo-labelling, entropy minimization, consistency regularization, and FixMatch variants. Subsection 5.2.2 presents the ensemble-based pseudo-label generation pipeline. In both cases, results are compared against the respective supervised baseline to assess the benefit of incorporating unlabelled data.

5.2.1 SSL via Training-Time Signal Integration

The first set of experiments benchmarked the SSL methods that incorporate unlabelled signal directly during model optimisation. These experiments aimed to assess how different strategies for exploiting unlabelled pixels during training affect segmentation performance.

Table 5.1 reports the DSC obtained for each method and its variants, with values reported separately for each class. The supervised baseline is included for reference, with shaded cells highlighting SSL methods that surpass it. The remaining evaluation metrics, namely recall, precision and specificity, are provided in Appendix A.2 for the best SSL method.

Table 5.1: Segmentation performance of the supervised baseline and the training-time SSL methods and their variants, measured by the DSC. Values are reported as mean $\pm$ standard deviation (%) across five data splits. Bold values indicate the best result per class and shaded grey cells indicate improvement over the supervised baseline. The cysts class is omitted, as it achieved $0.00 \pm 0.00\%$ across all experiments.

Method	Variant	Fibrosis	Nodules	Ground glass	Consolidation	Emphysema	Others	Healthy	Average
Supervised		81.35 $\pm$ 4.72	53.64 $\pm$ 8.69	44.84 $\pm$ 10.28	22.95 $\pm$ 15.66	52.89 $\pm$ 14.95	0.01 $\pm$ 0.02	30.66 $\pm$ 16.20	35.79 $\pm$ 26.27
Dynamic Pseudo-labelling	Validation Threshold	79.18 $\pm$ 4.45	51.19 $\pm$ 13.00	44.43 $\pm$ 5.57	15.85 $\pm$ 9.74	60.30 $\pm$ 15.74	0.04 $\pm$ 0.07	21.26 $\pm$ 16.67	34.03 $\pm$ 27.25
	Unlabelled Threshold	80.87 $\pm$ 5.99	51.27 $\pm$ 9.59	41.20 $\pm$ 13.56	23.38$\pm$15.29	54.42 $\pm$ 18.85	0.03 $\pm$ 0.07	35.08 $\pm$ 17.95	35.78 $\pm$ 25.89
Entropy Minimization	MedGIFT	83.05$\pm$3.56	55.82 $\pm$ 17.10	46.12$\pm$13.51	17.01 $\pm$ 9.38	50.61 $\pm$ 23.12	0.03 $\pm$ 0.05	26.79 $\pm$ 14.29	34.93 $\pm$ 27.27
	MedGIFT + HSJ	80.16 $\pm$ 3.29	55.92 $\pm$ 2.33	43.01 $\pm$ 12.15	18.15 $\pm$ 9.40	73.49$\pm$7.29	0.22$\pm$0.30	35.06 $\pm$ 8.87	38.25$\pm$28.85
Consistency Regularization		77.90 $\pm$ 4.01	50.71 $\pm$ 9.70	41.61 $\pm$ 8.73	16.53 $\pm$ 7.46	57.38 $\pm$ 16.25	0.04 $\pm$ 0.05	22.53 $\pm$ 7.66	33.34 $\pm$ 26.35
FixMatch	Unlabelled Threshold	78.30 $\pm$ 4.96	56.54$\pm$2.80	35.64 $\pm$ 8.96	13.54 $\pm$ 11.75	52.53 $\pm$ 25.49	0.01 $\pm$ 0.00	27.75 $\pm$ 2.37	33.04 $\pm$ 26.41
	FreeMatch	69.94 $\pm$ 8.61	41.64 $\pm$ 15.03	43.60 $\pm$ 7.74	23.35 $\pm$ 19.81	42.32 $\pm$ 25.62	0.18 $\pm$ 0.33	42.95$\pm$17.77	33.00 $\pm$ 22.35
	SoftMatch	73.93 $\pm$ 4.66	45.37 $\pm$ 7.15	32.00 $\pm$ 8.25	13.45 $\pm$ 8.08	49.77 $\pm$ 25.11	0.01 $\pm$ 0.02	28.22 $\pm$ 13.64	30.34 $\pm$ 24.07

The results presented in Table 5.1 show that the effect of incorporating unlabelled information during training varied considerably across methods and classes. Among all experiments, entropy minimization using both MedGIFT and HSJ data achieved the highest average DSC.

For dynamic pseudo-labelling, the variant using thresholds estimated from unlabelled pixels achieved a higher average DSC than the validation threshold variant. This improvement was mainly associated with better results for fibrosis, consolidation, and healthy tissue. In particular, the healthy class increased from 21.26% to 35.08%, suggesting that adapting the confidence thresholds to the unlabelled data distribution helped the model retain more useful pseudo-labelled pixels for this class. An inspection of the thresholds during training showed that the higher thresholds were similarly elevated across both variants for most classes, while the main difference was observed in the lower thresholds, which were considerably higher when estimated from labelled validation pixels. This indicates that validation-based thresholds may have imposed an overly strict selection criterion, rejecting many pixels with intermediate confidence that could still provide a useful training signal. In contrast, thresholds estimated from unlabelled pixels were more adapted to the confidence distribution encountered during training, allowing a larger set of pseudo-labelled pixels to contribute to the loss. However, this effect was not uniformly beneficial across classes, as the unlabelled threshold variant reduced the DSC for ground-glass opacities and emphysema. This suggests that lower thresholds can improve learning for some classes, but may also introduce noisier supervision when the model is less reliable.

Entropy minimization showed the best overall performance among the evaluated methods, with the variant incorporating additional data from HSJ achieving the highest average DSC of

38.25%. When only MedGIFT images were used, this method achieved the highest DSC for fibrosis and ground-glass opacities. Extending the unlabelled data with additional studies from HSJ increased the average DSC by 3.32 percentage points, mainly due to the substantial improvement in emphysema and the improvement in healthy tissue. This suggests that the additional unlabelled data may have provided useful variability for some visual patterns. Nevertheless, performance decreased for fibrosis and ground-glass opacities, which may indicate that the added data encouraged more confident predictions that were not equally beneficial for all classes.

Consistency regularization obtained an average DSC of 33.34%, performing below both dynamic pseudo-labelling variants and entropy minimization. In this setting, where annotations are sparse and several classes are under-represented, this signal may have been insufficient to guide the model towards better class separation.

Among the FixMatch variants, the unlabelled-threshold configuration achieved the highest average DSC, with a value of 33.04%, closely followed by FreeMatch with 33.00%. Although the average performance of these two variants was similar, their class-wise behaviour differed. The unlabelled-threshold variant obtained the best result for nodules among all training-time SSL methods, suggesting that this pseudo-label selection strategy was beneficial for this localised finding. In contrast, FreeMatch achieved better results for consolidation and healthy tissue within the FixMatch group. SoftMatch obtained the lowest average DSC, with 30.34%, and did not improve over the other variants in any class. Overall, the differences between FixMatch variants suggest that the pseudo-label selection mechanism has a meaningful impact on class-wise performance, yet the combined pseudo-labelling and consistency signal did not translate into a consistent advantage over simpler SSL approaches such as entropy minimization or dynamic pseudo-labelling.

Across all methods, fibrosis remained one of the best segmented findings, while cysts and the "others" class continued to show very poor performance. Cysts obtained a DSC of 0.00% in all experiments, confirming that the SSL signal was not sufficient to compensate for the extremely limited representation of this class. Overall, no training-time SSL method consistently outperformed the others across all classes. Instead, the results showed a strong class-dependent behaviour, with different methods benefiting different findings.

Regarding the comparison with the supervised baseline, the shaded cells in Table 5.1 show that at least one SSL method exceeded the supervised baseline for every class except cysts, although no single method achieved this consistently across all classes simultaneously. Entropy minimization with HSJ data achieved the best average DSC, representing an improvement of 2.46% points over the supervised baseline.

To complement the quantitative results, qualitative examples were also examined to assess the spatial behaviour of the predicted masks, with a particular focus on the unlabelled regions within the lung parenchyma. Since these regions do not have manual annotations, it is not possible to confirm whether the predictions are correct, but the examples allow a visual inspection of how each SSL method influences the model's behaviour in unannotated areas. Figure 5.4 presents representative examples obtained with the best-performing variant of each SSL method, with the supervised predictions also included as a reference.

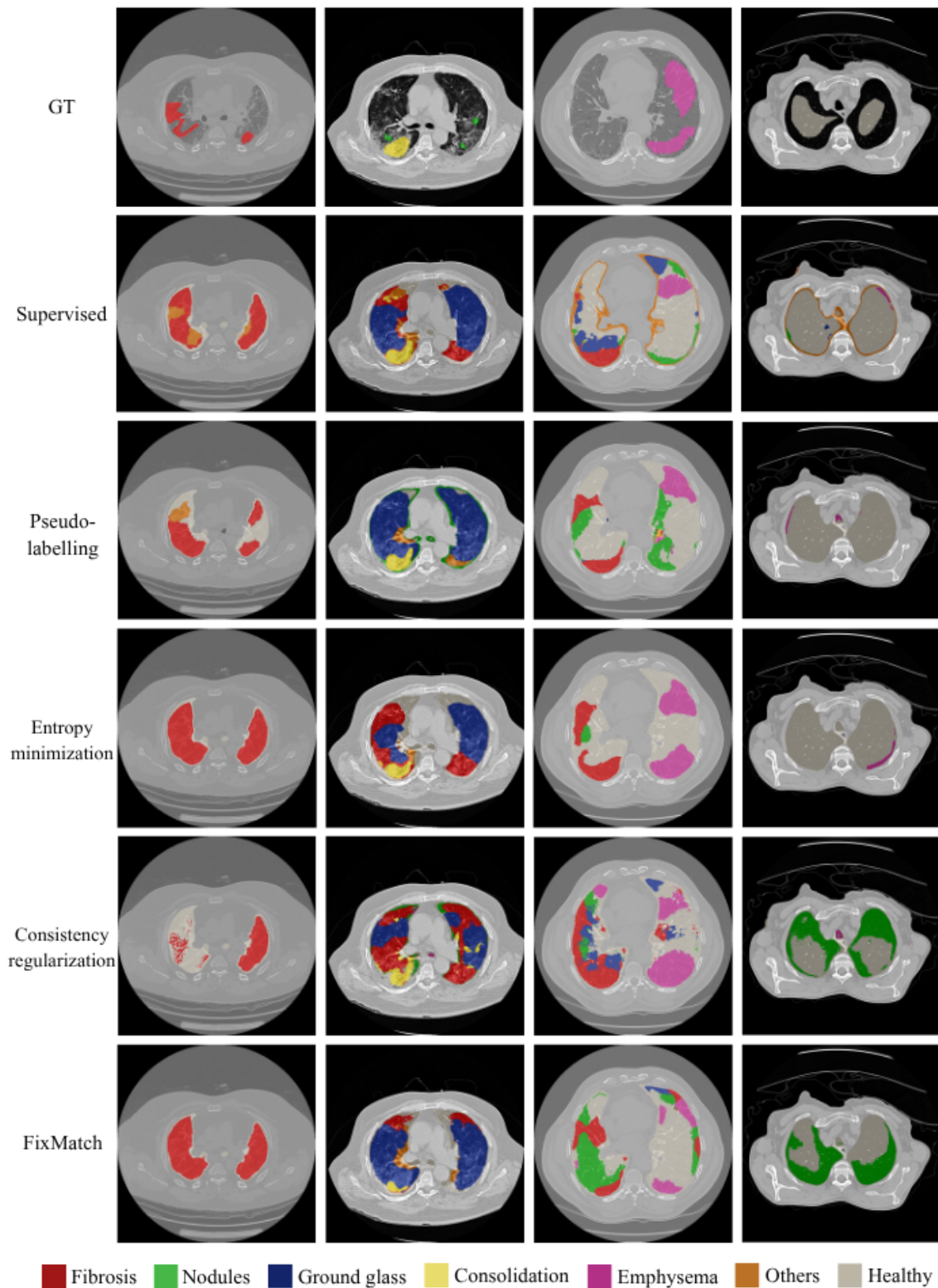


Figure 5.4: Qualitative comparison of the segmentation masks obtained with the supervised baseline and the best-performing variant of each training-time SSL method, including representative HRCT slices, the corresponding GT annotations, and the predicted masks generated by each strategy.

The qualitative examples illustrate that the SSL methods produced different behaviours predominantly in the unlabelled areas of the lung parenchyma. In the first example, the GT annotation corresponds to a fibrosis region, which was correctly identified by most methods with reasonable spatial overlap. Consistency regularization was the main exception, producing a more fragmented mask in one of the lung regions. In the unlabelled areas, entropy minimization and FixMatch showed particularly strong agreement, generating nearly identical predictions for fibrosis and healthy tissue. Dynamic pseudo-labelling stood out by additionally predicting a small region as the "others" class, a behaviour not observed in the remaining SSL methods.

In the second example, the consolidation region annotated in the GT was detected by all methods. Dynamic pseudo-labelling, entropy minimization, and consistency regularization produced similar consolidation masks, while FixMatch detected the same region with a smaller spatial extent. In contrast, the small nodular annotations present in the GT were not recovered by any method. In the unlabelled regions, the methods showed partial agreement, with fibrosis and ground-glass opacities being the most prominent additional predictions. Although the extent of the segmented regions varied across methods, the overall pattern suggests that the models identified similar abnormal tissue in some unannotated areas.

In the third example, corresponding to an emphysema region in the GT, entropy minimization produced the smoothest and most spatially coherent prediction, with the closest agreement to the annotated boundary. The remaining methods also detected emphysema, but with less precise lesion extent. In the unlabelled regions, all methods predicted additional fibrosis and nodular patterns, with differences in the size and spatial distribution of the segmented areas. A notably similar behaviour was observed between dynamic pseudo-labelling and entropy minimization in one of the lung regions, suggesting that these two methods may have captured comparable image patterns in areas without manual annotation.

The final example focuses on healthy tissue, where all methods assigned at least part of the annotated region to the correct class. Dynamic pseudo-labelling and entropy minimization provided the most complete overlap with the GT, although both methods also assigned a small number of peripheral pixels to emphysema, possibly due to the low-attenuation appearance of regions near the lung boundary, which can visually resemble emphysematous patterns. Consistency regularization and FixMatch detected healthy tissue but produced less complete masks, with part of the annotated region assigned to nodules instead. This behaviour is consistent with the confusion between healthy tissue and nodules observed both quantitatively and qualitatively in the predicted masks.

Overall, the qualitative analysis shows that the SSL methods often agreed in the annotated regions, particularly for larger findings. However, differences became more evident in unlabelled lung regions, where the absence of GT prevents definitive assessment of correctness. In several cases, multiple methods predicted similar classes in the same unannotated areas, suggesting that these regions may contain visually meaningful patterns not captured by the sparse annotations. At the same time, the variability in lesion extent and class assignment across methods reflects the inherent ambiguity of ILD findings, where visually similar textures can correspond to different

pathological classes, making consistent segmentation difficult.

5.2.2 SSL via Ensemble Pseudo-Label Generation

The second set of experiments evaluated the ensemble-based pseudo-label generation strategy. Unlike the training-time SSL methods presented in the previous subsection, this approach aimed to improve segmentation performance by expanding the available annotations through consensus-based pseudo-labels generated by an ensemble of models.

Table 5.2 summarises the DSC obtained by the different models involved in the pseudo-labelling generation pipeline, with stage 0 corresponding to the supervised configurations included for reference. Additionally, Table 5.2 compares the DSC achieved with the unfiltered HSJ pseudo-labels against two filtering strategies, positive and negative, identified as configurations 2, 2.1, and 2.2, respectively, where positive filtering retained only findings classified as present and negative filtering removed findings classified as absent. The remaining evaluation metrics, namely recall, precision and specificity, are provided in Appendix A.3 for the best model.

Table 5.2: Segmentation performance of the models obtained at different stages of the ensemble-based pseudo-labelling pipeline, measured by the DSC. Values are reported as mean $\pm$ standard deviation (%) across five data splits. The term "Config" refers to the experimental configuration used in each setup. Bold values indicate the best result per class, and shaded cells indicate improvement over the supervised baseline of the corresponding configuration, with grey, cyan, and orange representing configurations 1, 2, and 3, respectively. The cysts class is omitted, as it achieved $0.00 \pm 0.00\%$ across all experiments.

Stage	Config	Fibrosis	Nodules	Ground glass	Consolidation	Emphysema	Others	Healthy	Average
0	1	81.35 $\pm$ 4.72	53.64 $\pm$ 8.69	44.84 $\pm$ 10.28	22.95 $\pm$ 15.66	52.89 $\pm$ 14.95	0.01 $\pm$ 0.02	30.66 $\pm$ 16.20	35.79 $\pm$ 26.27
	2	84.84 $\pm$ 2.43	56.61 $\pm$ 5.43	51.23 $\pm$ 9.14	24.92 $\pm$ 8.31	64.71 $\pm$ 22.13	0.01 $\pm$ 0.02	62.63 $\pm$ 5.11	43.12 $\pm$ 29.34
	3	83.96 $\pm$ 2.36	55.30 $\pm$ 5.80	60.33 $\pm$ 5.33	10.97 $\pm$ 4.76	61.61 $\pm$ 22.07	0.01 $\pm$ 0.02	62.13 $\pm$ 14.98	41.79 $\pm$ 30.72
1	1	80.58 $\pm$ 5.10	56.79 $\pm$ 3.02	55.61 $\pm$ 8.32	21.83 $\pm$ 6.66	57.31 $\pm$ 26.33	0.00 $\pm$ 0.00	54.27 $\pm$ 6.09	40.80 $\pm$ 27.84
	2	86.62$\pm$1.39	63.87 $\pm$ 3.28	61.30 $\pm$ 5.96	28.92$\pm$8.82	60.04 $\pm$ 22.50	0.00 $\pm$ 0.00	66.20 $\pm$ 7.99	45.87 $\pm$ 30.27
	3	85.54 $\pm$ 1.18	59.36 $\pm$ 6.76	63.75 $\pm$ 6.41	16.69 $\pm$ 3.41	61.56 $\pm$ 27.23	0.01 $\pm$ 0.01	64.56 $\pm$ 9.21	43.93 $\pm$ 31.02
2	2	85.34 $\pm$ 0.73	65.22$\pm$4.92	65.36$\pm$4.26	24.67 $\pm$ 2.86	79.25$\pm$3.40	0.00 $\pm$ 0.00	68.02$\pm$5.39	48.48$\pm$32.63
	2.1	79.33 $\pm$ 1.91	58.54 $\pm$ 3.17	58.69 $\pm$ 5.96	20.64 $\pm$ 7.11	75.88 $\pm$ 2.06	0.00 $\pm$ 0.00	60.27 $\pm$ 8.53	44.17 $\pm$ 30.36
	2.2	78.72 $\pm$ 4.55	58.31 $\pm$ 4.42	62.34 $\pm$ 7.04	15.93 $\pm$ 6.72	76.48 $\pm$ 4.08	0.00 $\pm$ 0.00	60.08 $\pm$ 5.66	43.98 $\pm$ 31.07

The results of the first stage show that retraining the three network configurations with MedGIFT pseudo-labels improved the performance of all models compared to their supervised counterparts. This indicates that expanding the labelled training set with consensus-based pseudo-labels provided additional useful supervision for all configurations. Despite the improvement observed across all models, configuration 2 remained the best-performing setup, as observed in the supervised experiments. This reduced-patch configuration achieved the highest average DSC in the first stage and obtained the best results for fibrosis and consolidation. Configuration 3 achieved the best result for ground-glass opacities among the first stage models, while configuration 1 showed the

lowest average performance. These results suggest that the advantages observed for the reduced-patch configuration in the supervised setting were preserved after pseudo-label retraining.

The improvement obtained in the first stage can be related to the increase in the amount of supervised signal available during training. The pseudo-labelling process substantially expanded the number of labelled pixels for most classes, with increases of approximately 171% for fibrosis, 118% for nodules, 158% for ground-glass opacities, 81% for consolidation, 96% for emphysema, and 402% for healthy tissue. This additional supervision likely helped the models learn more representative class patterns beyond the sparse manual annotations, which explains the overall improvement observed across the three configurations. However, the benefit was not uniform across all classes. Cysts increased by only 8%, while the "others" class increased by approximately 24%, remaining much less represented than the other classes. This indicates that the ensemble pseudo-labelling strategy was effective in expanding the supervision for classes that already had some degree of representation, but was not sufficient to recover classes with extremely limited annotations. In this sense, the method improved the available training signal, but may also have reinforced the existing class imbalance by generating substantially more pseudo-labels for the more recognisable and already better-represented classes.

The second stage result should be compared primarily with configuration 2 from the first stage and with stage 0, corresponding to the supervised baseline, since all three models used the same reduced-patch network configuration. In this comparison, the second stage further improved the average DSC by 2.61 percentage points over the first stage, achieving the best overall result of the ensemble-based pseudo-labelling pipeline. When compared with the supervised baseline, this corresponds to an overall improvement of 5.36 percentage points. This improvement was primarily driven by higher performance for nodules, ground-glass opacities, emphysema, and healthy tissue. The result for emphysema was particularly notable, since it was also accompanied by a marked reduction in standard deviation compared with the first stage configurations, suggesting that the additional pseudo-labelled data may have provided more consistent supervision for this class across folds. These gains compensated for the decrease observed in fibrosis and consolidation, which performed better in the first stage configuration 2 model.

The improvement in the second stage can again be linked to the expansion of the pseudo-labelled training set. Within MedGIFT, the second pseudo-labelling stage promoted additional pixels for most classes, increasing the amount of labelled supervision by approximately 60% for fibrosis, 36% for nodules, 82% for ground-glass opacities, 48% for consolidation, 32% for emphysema, and 55% for healthy tissue compared to the pseudo-labelled set available after the first stage. In addition, the HSJ dataset introduced a large amount of new pseudo-labelled lung regions, particularly for healthy tissue. This broader pseudo-labelled set may have exposed the model to a wider range of pathological appearances, contributing to the improvements and generalisation observed across classes. However, the second stage also reinforced some of the limitations observed in the first stage. No additional cyst pixels were promoted in MedGIFT, and no cyst pseudo-labels were generated from the HSJ dataset, which is consistent with the persistent DSC of 0.00% for this class. The "others" class also remained poorly represented, despite a small increase in MedGIFT

and very limited pseudo-label generation in HSJ. Therefore, although the second stage improved the overall average DSC, its benefit was concentrated in classes for which the ensemble could generate a sufficient number of confident and consistent pseudo-labels.

Building on the second stage ensemble model, the results obtained with the filtered HSJ pseudo-labels, identified in Table 5.2 as configurations 2.1 and 2.2, show that filtering did not improve segmentation performance compared to the unfiltered baseline. This degradation was observed not only for the pathological classes directly affected by filtering, but also for the healthy class, which was not explicitly filtered. This suggests that reducing the amount of pathological pseudo-label supervision may have indirectly disrupted the balance between healthy and pathological tissue during retraining, making it harder for the model to distinguish normal parenchyma from disease-related patterns.

A relevant aspect to consider when interpreting these results is the inherent asymmetry between the two filtering strategies. In the classification output, findings with negation set to false occurred substantially more often than findings with negation set to true, with 1,668 and 779 occurrences respectively. This asymmetry was also reflected at the case level. Under the first filtering strategy, the number of studies with pseudo-labels decreased from 389 to 201 for fibrosis, from 118 to 40 for nodules, from 215 to 33 for ground-glass opacities, from 112 to 23 for consolidation, and from 55 to 17 for emphysema. Under the second filtering strategy, the corresponding reductions were comparatively smaller, with cases decreasing from 389 to 377 for fibrosis, from 118 to 97 for nodules, from 215 to 202 for ground-glass opacities, from 112 to 102 for consolidation, and from 55 to 54 for emphysema. It is worth noting that no pseudo-labels were generated for cysts, despite this finding being reported in approximately 40 examinations according to the classification output.

Despite this difference in the amount of data removed, both strategies led to comparable performance drops, which suggests that the quantity of pseudo-labels may be a more critical factor than their correctness in this setting. However, this interpretation should be treated with caution, as the evaluation was performed exclusively on the MedGIFT test set, while the filtering was applied to the HSJ pseudo-labels. This mismatch prevents a direct observation of the filtering effect on the data it was actually applied to, and it is possible that the filtered pseudo-labels were of higher quality, but that this benefit was not reflected in the MedGIFT evaluation.

Furthermore, a fundamental limitation of the current filtering approach lies in the mismatch between the level at which textual and image-based information operate. Radiology reports describe findings at the examination level, whereas segmentation requires decisions at the pixel level. As a result, a pseudo-label may be correctly retained according to the report-level information while still being inconsistent with the true location of the finding in the image. Extending the classification framework to incorporate anatomical localisation could enable a more precise and spatially aware filtering of pseudo-labels, reducing the risk of retaining incorrectly placed annotations. A possible future direction would be to explore Vision Language Models (VLMs), which are designed to jointly process text and images and establish correspondences between language descriptions and image regions. In the context of ILD segmentation, such models could help

align findings described in radiology reports with their anatomical locations in HRCT scans. This would represent a meaningful step towards filtering pseudo-labels not only by what is reported, but also by where it belongs, bringing report-derived supervision closer to the precision required for pixel-level segmentation.

To better illustrate the effect of the ensemble-based pseudo-labelling pipeline, Figure 5.5 shows examples of pseudo-labels generated for slices that originally contained no manual annotations. In several cases, the generated pseudo-labels appeared in slices adjacent to annotated slices and represented a spatial continuation of pathological findings already present in the manual annotations. This suggests that the ensemble was able to exploit contextual consistency across neighbouring HRCT slices, promoting additional labels in regions where the original dataset provided no explicit supervision.

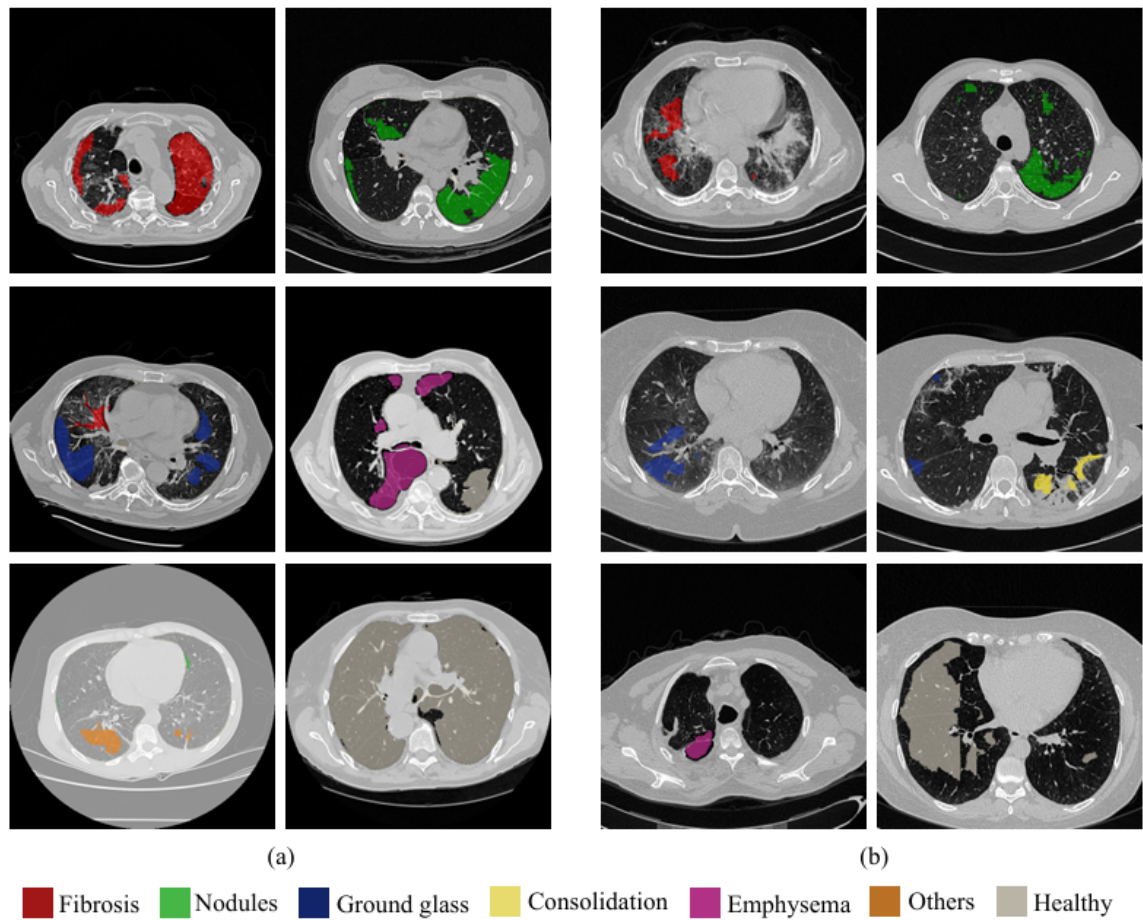


Figure 5.5: Examples of pseudo-labels generated by the ensemble-based pipeline for slices without original manual annotations. Examples are shown for (a) MedGIFT data and (b) HSJ data.

To complement the quantitative results, qualitative examples were examined to assess the visual impact of the successive pseudo-labelling stages. Rather than comparing all configurations, the examples focus on configuration 2 from the first stage and the second stage model, since these represent the best result of each stage and allow a more focused comparison of the effect

of additional pseudo-label expansion and the incorporation of HSJ data. The supervised baseline, corresponding to stage 0, was also included as a reference. Figure 5.6 presents representative examples obtained with these three models.

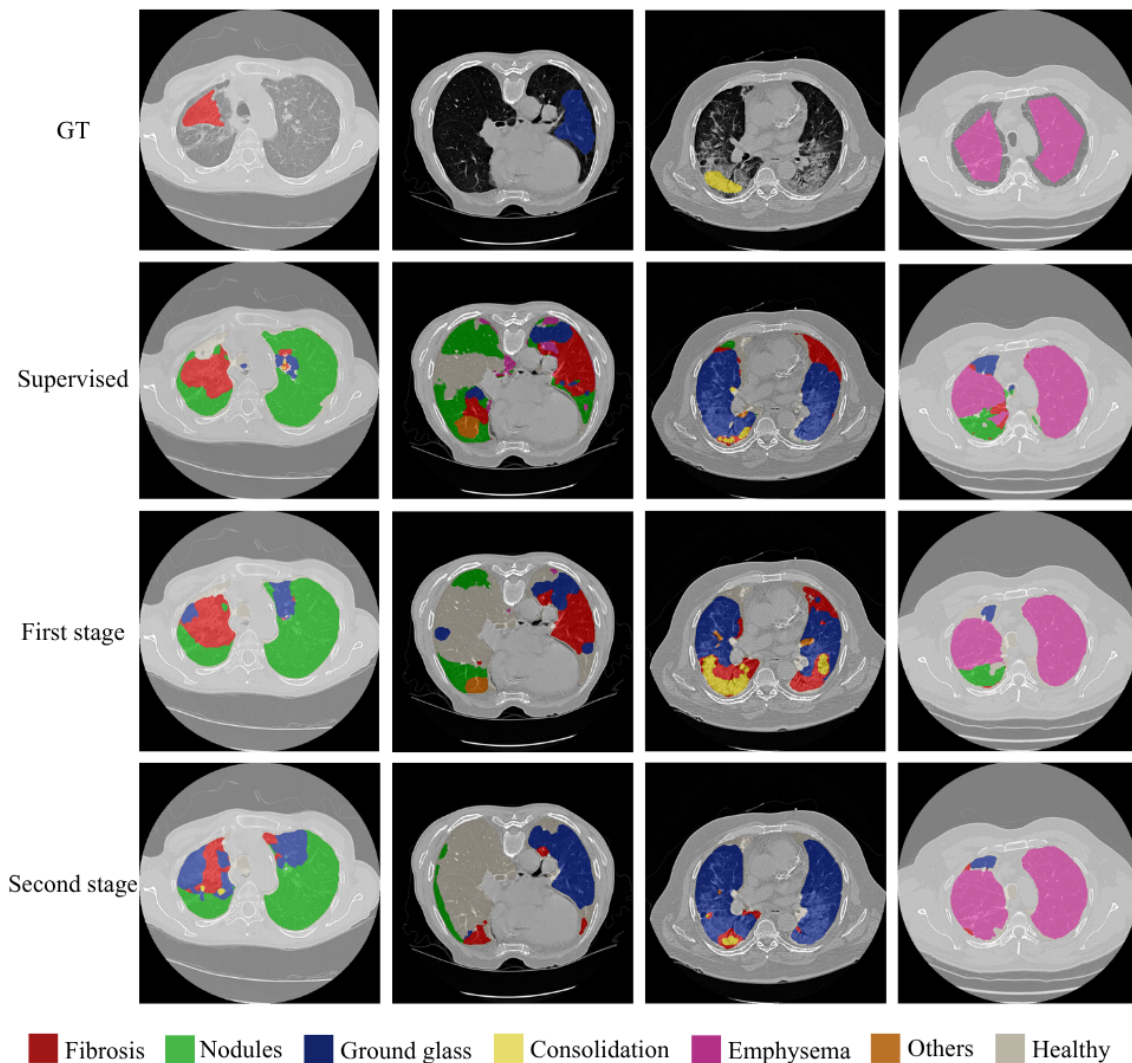


Figure 5.6: Qualitative comparison of the segmentation masks obtained with configuration 2 from the supervised setting, the first stage and the second stage model of the ensemble-based pseudo-labelling pipeline, including representative HRCT slices, the corresponding GT annotations, and the predicted masks generated by each model.

The qualitative examples were selected to illustrate classes where the quantitative results differed across the supervised, first stage, and second stage models, namely fibrosis, ground-glass opacities, consolidation and emphysema. Healthy tissue was not included as a separate example, despite its quantitative improvement, since all models produced visually similar segmentations for this class.

In general, the predictions in the labelled GT regions visually support the trends observed in Table 5.2. In the fibrosis example, the first stage model produced a more complete fibrosis

mask than the supervised baseline, whereas the second stage model showed some confusion with ground-glass opacities. Conversely, in the ground-glass example, the second stage model better covered the annotated region, while the first stage and supervised models assigned part of the finding to fibrosis. For consolidation, the first stage model showed better agreement with the GT than the second stage model, although both predictions reflected confusion between consolidation, ground-glass opacities, and fibrosis. This confusion is expected, as these three classes share overlapping radiological appearances in HRCT, often manifesting as regions of increased parenchymal attenuation that can be difficult to distinguish visually, even in manual annotations. Furthermore, incomplete GT annotations for consolidation may have contributed to this confusion during training, which is also consistent with the generally low segmentation performance observed for this class in all experiments. Finally, in the emphysema example, the second stage model produced a more complete emphysema segmentation, whereas the first stage and supervised models confused part of the region with nodules.

In the unlabelled regions, the first and second stage models produced broadly similar predictions in several areas, which is expected since both stages used the same original GT annotations and shared the first set of MedGIFT pseudo-labels. However, the second stage model introduced some refinements in the extent and class assignment of the predicted regions, reflecting the additional pseudo-labels generated in the second stage and the incorporation of HSJ data. As in the previous qualitative analyses, these predictions cannot be confirmed without complete annotations, but their agreement across stages suggests that the ensemble-based pseudo-labelling pipeline captured consistent image patterns in some unlabelled lung regions.

To further inspect the behaviour of the best ensemble-based model, Figure 5.7 presents the pixel-wise confusion matrix obtained for the second stage configuration 2 model. The second stage ensemble model increased the proportion of correctly classified pixels for several pathological findings. This was particularly evident for ground-glass opacities and emphysema, with nodules also showing improvement. These changes are consistent with the DSC results and indicate that the expanded pseudo-labelled supervision helped the model recover a larger proportion of pixels from these classes. However, the improvement in ground-glass opacities was accompanied by increased confusion from other classes into this label.

Another relevant change concerns the healthy class. Although the percentage of correctly classified healthy pixels remained similar, the second stage model reduced the tendency to classify diseased regions as healthy for several classes, including nodules, ground-glass opacities, emphysema, and "others". This is clinically relevant, since assigning abnormal parenchyma to the healthy class may lead to under-estimation of disease extent.

Overall, the confusion matrix supports the trends observed in the quantitative results. The second stage ensemble model improved the detection of several relevant findings, while persistent limitations remained for cysts, "others", and consolidation.



Figure 5.7: Pixel-wise confusion matrix normalized by true label for the second stage configuration 2 model of the ensemble-based pseudo-labelling pipeline.

5.3 Summary

This chapter presented the SSL experiments developed to exploit unlabelled lung regions for ILD findings segmentation. Building on the supervised baseline established in the previous chapter, two main semi-supervised directions were investigated: methods that incorporate unlabelled signal directly during training, and an ensemble-based pseudo-label generation pipeline designed to expand the available annotations before retraining.

The training-time SSL experiments showed that the effect of unlabelled data was strongly dependent on the method and the target class. Entropy minimization achieved the best average performance among these methods, mainly due to improvements in emphysema, healthy tissue, and nodules. However, no training-time SSL method consistently improved all classes, suggesting that direct use of unlabelled signal during optimisation can introduce class-specific benefits but also unstable behaviour.

The ensemble-based pseudo-labelling pipeline produced the strongest semi-supervised results. By combining predictions from multiple models and promoting only high-confidence consensus predictions, this approach expanded the available supervision while reducing the risk of propagating unreliable pseudo-labels. The second stage of this pipeline, which incorporated additional pseudo-labels from both MedGIFT and HSJ data, achieved the best overall performance of the chapter. This indicates that, under the sparse annotation conditions of the MedGIFT database, generating pseudo-labels before retraining was more effective than relying only on training-time SSL methods.

Comparison with supervised baselines showed that SSL improved performance when applied in comparable experimental settings. Nevertheless, the benefits of SSL remained strongly dependent on the initial labelled data. Classes with some initial representation, such as fibrosis, nodules, ground-glass opacities, emphysema, and healthy tissue, were the ones that benefited the most from pseudo-labelling. However, the improvement was not determined only by the number of annotated pixels. Emphysema, for instance, had limited representation in the original dataset, similar to the "others" class in terms of annotated slices, but showed one of the largest improvements with SSL. This suggests that its visual appearance may have been more distinguishable, allowing the model to generate more reliable pseudo-labels once additional unlabelled data were introduced. In contrast, cysts and "others" remained poorly segmented, showing that SSL was not sufficient to recover classes for which the supervised model had little or no reliable signal from the beginning. This reinforces the idea that pseudo-label quality is highly influenced by the performance of the initial model, since classes that were not reliably learnt in the supervised stage are unlikely to produce consistent pseudo-labels during the generation process.

It is also important to note that the evaluation was restricted to pixels with available GT annotations, as in the supervised experiments. Since the MedGIFT database is sparsely annotated, only a limited portion of the lung parenchyma contributed to the quantitative metrics. As a result, the reported DSC values may not fully capture the effect of SSL, especially if the model improved its predictions in unlabelled regions that were excluded from evaluation. This is particularly relevant because the qualitative examples showed additional predictions in unannotated areas, which in some cases appeared anatomically plausible. A fully annotated test set would therefore be required to assess more completely the benefit of the SSL models across the entire lung region and across datasets. A related limitation concerns the HSJ data used during SSL training. Since these studies were unlabelled, it was not possible to quantitatively assess whether incorporating HSJ data improved performance on the HSJ domain itself. Therefore, potential gains in generalization to this external dataset may not be reflected in the MedGIFT-based evaluation.

Overall, this chapter demonstrates the value of SSL for ILD findings segmentation using a public dataset with sparse and incomplete annotations. To the best of our knowledge, no previous work has systematically evaluated multiple semi-supervised strategies for ILD findings segmentation under these conditions. This is particularly relevant because the MedGIFT database presents several practical challenges, including incomplete delineation of pathological regions, coarse annotations in some findings and strong class imbalance. These factors directly affect both SL and

SSL, since the semi-supervised models depend on the reliability of the initial supervised signal. Therefore, while SSL improved performance for classes with more consistent or visually distinguishable patterns, its benefit remained limited for classes with very scarce or heterogeneous annotations. The results provide a relevant benchmark for future work using partially labelled medical imaging datasets, while also suggesting that the impact of SSL could be greater if more robust initial annotations were available for the under-represented classes.

Chapter 6

Conclusions and Future Work

This dissertation investigated SL and SSL strategies for the segmentation of ILD findings, motivated by the increasing relevance of automated and quantitative image analysis tools, as well as by the practical limitations associated with the availability of reliable pixel-level annotations in public ILD datasets. Although DL methods have shown strong potential for medical image segmentation, their application to ILD findings remains challenging, as current state-of-the-art approaches remain predominantly supervised and do not leverage the large amount of unlabelled data that is typically available.

The first phase of this dissertation focused on SL using a 2D nnU-Net framework applied to the MedGIFT ILD database. Since the available annotations were sparse and incomplete, different supervision strategies were evaluated to determine how unlabelled pixels, healthy tissue, background, and non-target findings should be represented during training. The results showed that treating all non-annotated pixels as background introduced noisy supervision, whereas ignoring unlabelled regions led to more reliable segmentation behaviour. Different network configurations were also investigated, with the reduced-patch U-Net achieving the best overall performance, suggesting that more effective sampling of annotated regions was more beneficial than using the full axial slice as input. In general, these experiments established a supervised baseline and highlighted the main limitations imposed by sparse annotations and class imbalance, motivating the subsequent exploration of SSL strategies to exploit unlabelled lung regions.

Building on the supervised baseline, the second phase of this dissertation investigated SSL strategies to determine whether unlabelled lung regions could improve ILD findings segmentation under realistic annotation constraints. Two main directions were explored: training-time SSL methods, including dynamic pseudo-labelling, entropy minimization, consistency regularization, and FixMatch-based approaches, and an ensemble-based pseudo-label generation pipeline, where pseudo-labels were generated before retraining using agreement between multiple trained models. The training-time SSL experiments showed class-specific benefits, but no method consistently improved all findings. In contrast, the ensemble-based pipeline produced the strongest semi-supervised results, particularly in its second stage, where pseudo-labels from both MedGIFT and HSJ data were incorporated.

As a continuation of the second stage of the SSL ensemble-based pseudo-labelling pipeline, an additional experiment explored whether information extracted from radiology reports could be used to filter the pseudo-labels generated for the HSJ dataset. A multi-label text classification approach, previously developed within the laboratory, was used to identify findings described as present or absent in each examination, and this information was applied as a positive or negative filtering constraint. However, the filtered pseudo-label strategies did not improve performance compared with the unfiltered ensemble pseudo-labels. A finding may be correctly mentioned in a report while still being incorrectly localized by the segmentation model, since the report provides no spatial information to validate the pseudo-labels at the pixel or slice level. Furthermore, this interpretation should be treated with caution, as the evaluation was performed exclusively on the MedGIFT test set, while the filtering was applied to the HSJ pseudo-labels.

The comparison between supervised and semi-supervised models showed that SSL can improve segmentation performance when applied under comparable experimental settings. The improvements were most evident for classes that had some reliable initial signal or visually distinguishable appearance. Emphysema was a particularly relevant example, since it benefited substantially from SSL, suggesting that its visual pattern was sufficiently distinctive for the model to generate useful pseudo-labels once additional unlabelled data were introduced. In contrast, classes such as cysts and "others" remained poorly segmented. This reinforces an important conclusion of the dissertation, namely that SSL can amplify useful signals already present in the labelled data, but it cannot fully compensate for classes that are almost absent, poorly defined, or inconsistently annotated from the beginning.

These results must be interpreted in light of the main limitations of the available data. The MedGIFT database is valuable because it is public and contains pixel-level annotations for several ILD findings, but its annotations are incomplete, and highly imbalanced. Many lung regions remain unlabelled, and qualitative analysis showed that some GT masks were coarse, ambiguous, or restricted to selected ROIs. In particular, small or poorly delineated findings, such as nodules and cysts, often appeared to include mixtures of pathological and normal tissue, which may have contributed to the confusion between these classes and healthy parenchyma. Because the quantitative metrics were computed only on labelled pixels, this incomplete annotation in the test set also limits the extent to which the reported values capture the effect of SSL in unlabelled regions. Although some additional predictions in these regions appeared anatomically plausible, this could not be confirmed without complete annotations. A further limitation concerns the HSJ data used during SSL training, which did not contain manual annotations. Therefore, although incorporating HSJ data improved some results on MedGIFT, it was not possible to directly assess whether the model also improved on the HSJ domain itself. A fully annotated test set would be necessary to evaluate generalization across datasets and to better quantify the impact of incorporating external unlabelled data.

Following from the limitations discussed, several future directions can be considered. The first priority should be the refinement of the available annotations, particularly in the MedGIFT database. Even a limited extension or correction of annotations for under-represented classes could

improve the initial supervised signal, with particular attention given to cysts and "others", as these classes remained difficult throughout the experiments. Refining coarse ROIs and reducing the inclusion of normal tissue inside pathological regions could also help reduce confusion between healthy parenchyma and disease patterns. This step is especially important because the results of this dissertation showed that SSL methods remain strongly dependent on the quality of the initial labelled data.

Future work should also include the creation or acquisition of more complete evaluation sets. A fully annotated subset of MedGIFT would allow a more accurate evaluation of model performance across the entire lung parenchyma, including regions that are currently excluded from metric computation. Similarly, an annotated subset of HSJ data would make it possible to evaluate external performance and determine whether incorporating unlabelled HSJ scans improves generalization to that dataset.

In terms of methodology, more advanced SSL frameworks could be explored. Contrastive learning, in particular, could be investigated as a way to learn more discriminative feature representations, encouraging regions with similar pathological patterns to be closer in the feature space while separating visually similar but clinically distinct findings. This may be particularly relevant for classes that showed frequent confusion, such as nodules and healthy tissue, or fibrosis and ground-glass opacities. However, these approaches should ideally be investigated after improving the initial annotation quality, since more complex SSL methods are unlikely to fully overcome the absence of reliable signal for poorly represented classes.

Finally, future approaches could investigate methods that link textual descriptions to anatomical regions or image slices. VLMs may be particularly relevant in this context, as they are designed to jointly process textual and visual information and could potentially help associate reported findings with corresponding regions in HRCT images.

In conclusion, this dissertation demonstrated that SSL is a promising direction for ILD findings segmentation, while also showing that its success is fundamentally constrained by the quality and structure of the available annotations. By systematically evaluating multiple SL and SSL strategies on a public ILD dataset, this work provides a relevant benchmark for a range of approaches that has not been previously explored under this level of annotation incompleteness. Ultimately, this investigation suggests that progress in this domain depends less on increasingly complex modelling and more on the careful alignment between method design and the underlying data, a balance that future research will need to address to fully realise the potential of SSL for ILD findings segmentation.

References

- [1] Marlies Wijsenbeek, Atsushi Suzuki, and Toby M Maher. “Interstitial lung diseases”. In: *The Lancet* 400.10354 (Sept. 2022), pp. 769–786. ISSN: 01406736. DOI: [10.1016/S0140-6736\(22\)01052-2](https://doi.org/10.1016/S0140-6736(22)01052-2).
- [2] Claudia Valenzuela and Vincent Cottin. “Epidemiology and real-life experience in progressive pulmonary fibrosis”. In: *Current Opinion in Pulmonary Medicine* 28.5 (Sept. 2022), pp. 407–413. ISSN: 1070-5287, 1531-6971. DOI: [10.1097/MCP.0000000000000908](https://doi.org/10.1097/MCP.0000000000000908).
- [3] Shelly Soffer et al. “Artificial Intelligence for Interstitial Lung Disease Analysis on Chest Computed Tomography: A Systematic Review”. In: *Academic Radiology* 29 (Feb. 2022), S226–S235. ISSN: 10766332. DOI: [10.1016/j.acra.2021.05.014](https://doi.org/10.1016/j.acra.2021.05.014).
- [4] Jonas Widell and Mats Lidén. “Interobserver variability in high-resolution CT of the lungs”. In: *European Journal of Radiology Open* 7 (2020), p. 100228. ISSN: 2352-0477. DOI: [10.1016/j.ejro.2020.100228](https://doi.org/10.1016/j.ejro.2020.100228).
- [5] S. Agarwala et al. “Deep learning for screening of interstitial lung disease patterns in high-resolution CT images”. In: *Clinical Radiology* 75.6 (June 2020), 481.e1–481.e8. ISSN: 00099260. DOI: [10.1016/j.crad.2020.01.010](https://doi.org/10.1016/j.crad.2020.01.010).
- [6] Jooae Choe et al. “CT Quantification of Interstitial Lung Abnormality and Interstitial Lung Disease: From Technical Challenges to Future Directions”. In: *Investigative Radiology* 60.1 (Jan. 2025), pp. 43–52. ISSN: 1536-0210, 0020-9996. DOI: [10.1097/RLI.0000000000001103](https://doi.org/10.1097/RLI.0000000000001103).
- [7] Anna Russo et al. “AI-Based HRCT Quantification in Connective Tissue Disease-Associated Interstitial Lung Disease”. In: *Diagnostics* 15.17 (Aug. 2025), p. 2179. ISSN: 2075-4418. DOI: [10.3390/diagnostics15172179](https://doi.org/10.3390/diagnostics15172179).
- [8] Xiangde Luo et al. “Semi-supervised Medical Image Segmentation through Dual-task Consistency”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.10 (May 2021), pp. 8801–8809. ISSN: 2374-3468, 2159-5399. DOI: [10.1609/aaai.v35i10.17066](https://doi.org/10.1609/aaai.v35i10.17066).
- [9] Kai Han et al. “Deep semi-supervised learning for medical image segmentation: A review”. In: *Expert Systems with Applications* 245 (July 2024), p. 123052. ISSN: 09574174. DOI: [10.1016/j.eswa.2023.123052](https://doi.org/10.1016/j.eswa.2023.123052).
- [10] Moshe Haddad and Sandeep Sharma. “Physiology, Lung”. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2025. URL: <http://www.ncbi.nlm.nih.gov/books/NBK545177/>.

- [11] G. AmarTej and R. Nakkeeran. “Advancements in Interstitial Lung Disease Classification and Segmentation: A Comprehensive Review”. In: *2023 Second International Conference on Advances in Computational Intelligence and Communication (ICACIC)*. Puducherry, India: IEEE, Dec. 2023, pp. 1–6. ISBN: 979-8-3503-1845-6. DOI: [10.1109/ICACIC59454.2023.10435032](https://doi.org/10.1109/ICACIC59454.2023.10435032).
- [12] Carlotta Bertolina et al. “A review on Interstitial Lung Diseases: an overview from diagnosis to innovative therapies”. In: *Working Paper of Public Health 11.1* (Sept. 2023). ISSN: 2279-9761. DOI: [10.4081/wpph.2023.9785](https://doi.org/10.4081/wpph.2023.9785).
- [13] Guy’s and St Thomas’ Specialist Care. *Interstitial lung disease (ILD)*. URL: <https://guysandstthomasspecialistcare.co.uk/conditions/interstitial-lung-disease-ild/> (visited on 10/30/2025).
- [14] Sergio Harari et al. “Recent advances in the pathobiology and clinical management of lymphangioleiomyomatosis”. In: *Current Opinion in Pulmonary Medicine* 24.5 (Sept. 2018), pp. 469–476. ISSN: 1070-5287. DOI: [10.1097/MCP.0000000000000502](https://doi.org/10.1097/MCP.0000000000000502).
- [15] Dinesh Khanna et al. “Etiology, Risk Factors, and Biomarkers in Systemic Sclerosis with Interstitial Lung Disease”. In: *American Journal of Respiratory and Critical Care Medicine* 201.6 (Mar. 2020), pp. 650–660. ISSN: 1073-449X, 1535-4970. DOI: [10.1164/rccm.201903-0563CI](https://doi.org/10.1164/rccm.201903-0563CI).
- [16] Gregory C. McDermott, Tracy J. Doyle, and Jeffrey A. Sparks. “Interstitial lung disease throughout the rheumatoid arthritis disease course”. In: *Current Opinion in Rheumatology* 33.3 (May 2021), pp. 284–291. ISSN: 1040-8711, 1531-6963. DOI: [10.1097/BOR.0000000000000787](https://doi.org/10.1097/BOR.0000000000000787).
- [17] David J. Lederer and Fernando J. Martinez. “Idiopathic Pulmonary Fibrosis”. In: *New England Journal of Medicine* 378.19 (May 2018). Ed. by Dan L. Longo, pp. 1811–1823. ISSN: 0028-4793, 1533-4406. DOI: [10.1056/NEJMr1705751](https://doi.org/10.1056/NEJMr1705751).
- [18] Joan B Soriano et al. “Prevalence and attributable health burden of chronic respiratory diseases, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017”. In: *The Lancet Respiratory Medicine* 8.6 (June 2020), pp. 585–596. ISSN: 2213-2600. DOI: [10.1016/S2213-2600\(20\)30105-3](https://doi.org/10.1016/S2213-2600(20)30105-3).
- [19] Bhavika Kaul et al. “Variability in Global Prevalence of Interstitial Lung Disease”. In: *Frontiers in Medicine* 8 (Nov. 2021), p. 751181. ISSN: 2296-858X. DOI: [10.3389/fmed.2021.751181](https://doi.org/10.3389/fmed.2021.751181).
- [20] Boris Duchemann et al. “Prevalence and incidence of interstitial lung diseases in a multi-ethnic county of Greater Paris”. In: *European Respiratory Journal* 50.2 (Aug. 2017), p. 1602419. ISSN: 0903-1936, 1399-3003. DOI: [10.1183/13993003.02419-2016](https://doi.org/10.1183/13993003.02419-2016).
- [21] B. T. Society and S. O. C. Committee. “The Diagnosis, Assessment and Treatment of Diffuse Parenchymal Lung Disease in Adults”. In: *Thorax* 54.Supplement 1 (Apr. 1999), S1–S28. ISSN: 0040-6376. DOI: [10.1136/thx.54.suppl_1.S1](https://doi.org/10.1136/thx.54.suppl_1.S1).
- [22] Jay H. Ryu et al. “Diagnosis of Interstitial Lung Diseases”. In: *Mayo Clinic Proceedings* 82.8 (Aug. 2007), pp. 976–986. ISSN: 00256196. DOI: [10.4065/82.8.976](https://doi.org/10.4065/82.8.976).
- [23] Jürgen Behr et al. “Management of patients with idiopathic pulmonary fibrosis in clinical practice: the INSIGHTS-IPF registry”. In: *European Respiratory Journal* 46.1 (July 2015), pp. 186–196. ISSN: 0903-1936, 1399-3003. DOI: [10.1183/09031936.00217614](https://doi.org/10.1183/09031936.00217614).

- [24] Ganesh Raghu et al. “Comorbidities in idiopathic pulmonary fibrosis patients: a systematic literature review”. In: *The European Respiratory Journal* 46.4 (Oct. 2015), pp. 1113–1130. ISSN: 1399-3003. DOI: [10.1183/13993003.02316-2014](https://doi.org/10.1183/13993003.02316-2014).
- [25] A. Giménez Palleiro, S.P. Mazzini, and T. Franquet. “Basic HRCT patterns in diffuse interstitial lung disease”. In: *Radiología (English Edition)* 64 (Dec. 2022), pp. 215–226. ISSN: 21735107. DOI: [10.1016/j.rxeng.2022.12.001](https://doi.org/10.1016/j.rxeng.2022.12.001).
- [26] Ella A. Kazerooni. “High-Resolution CT of the Lungs”. In: *American Journal of Roentgenology* 177.3 (Sept. 2001), pp. 501–519. ISSN: 0361-803X, 1546-3141. DOI: [10.2214/ajr.177.3.1770501](https://doi.org/10.2214/ajr.177.3.1770501).
- [27] Jay H. Ryu et al. “Diagnostic Approach to the Patient With Diffuse Lung Disease”. In: *Mayo Clinic Proceedings* 77.11 (Nov. 2002), pp. 1221–1227. ISSN: 00256196. DOI: [10.4065/77.11.1221](https://doi.org/10.4065/77.11.1221).
- [28] Tami D. DenOtter and Johanna Schubert. “Hounsfield Unit”. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2025. URL: <http://www.ncbi.nlm.nih.gov/books/NBK547721/>.
- [29] Diedre Carmo et al. *Automated computed tomography and magnetic resonance imaging segmentation using deep learning: a beginner’s guide*. arXiv:2304.05901 [eess]. Apr. 2023. DOI: [10.48550/arXiv.2304.05901](https://doi.org/10.48550/arXiv.2304.05901).
- [30] Marieke Hovinga et al. “CT Imaging of Interstitial Lung Diseases”. In: *Multidetector-Row CT of the Thorax*. Ed. by U. Joseph Schoepf and Felix G. Meinel. Series Title: Medical Radiology. Cham: Springer International Publishing, 2016, pp. 105–130. ISBN: 978-3-319-30353-6 978-3-319-30355-0. DOI: [10.1007/978-3-319-30355-0_7](https://doi.org/10.1007/978-3-319-30355-0_7).
- [31] Wayne Richard Webb. *High-resolution CT of the lung*. 5th ed. Philadelphia: Wolters Kluwer Heath, 2015. ISBN: 978-1-4511-7601-8.
- [32] Karen Rodriguez et al. “High-Resolution Computed Tomography of Fibrotic Interstitial Lung Disease”. In: *Seminars in Respiratory and Critical Care Medicine* 43.06 (Dec. 2022), pp. 764–779. ISSN: 1069-3424, 1098-9048. DOI: [10.1055/s-0042-1755563](https://doi.org/10.1055/s-0042-1755563).
- [33] Joseph Jacob and David M. Hansell. “HRCT of fibrosing lung disease”. In: *Respirology* 20.6 (Aug. 2015), pp. 859–872. ISSN: 1323-7799, 1440-1843. DOI: [10.1111/resp.12531](https://doi.org/10.1111/resp.12531).
- [34] Anastasia Oikonomou and Panos Prassopoulos. “Mimics in chest disease: interstitial opacities”. In: *Insights into Imaging* 4.1 (Feb. 2013), pp. 9–27. ISSN: 1869-4101. DOI: [10.1007/s13244-012-0207-7](https://doi.org/10.1007/s13244-012-0207-7).
- [35] David A Lynch et al. “Diagnostic criteria for idiopathic pulmonary fibrosis: a Fleischner Society White Paper”. In: *The Lancet Respiratory Medicine* 6.2 (Feb. 2018), pp. 138–153. ISSN: 22132600. DOI: [10.1016/S2213-2600\(17\)30433-2](https://doi.org/10.1016/S2213-2600(17)30433-2).
- [36] Anna Rita Larici et al. “ESR Essentials: imaging in fibrotic lung diseases—practice recommendations by the European Society of Thoracic Imaging”. In: *European Radiology* 35.4 (Sept. 2024), pp. 2245–2255. ISSN: 1432-1084. DOI: [10.1007/s00330-024-11054-2](https://doi.org/10.1007/s00330-024-11054-2).
- [37] Yigal Abramowitz et al. “Pleural Effusion: Characterization with CT Attenuation Values and CT Appearance”. In: *American Journal of Roentgenology* 192.3 (Mar. 2009), pp. 618–623. ISSN: 0361-803X, 1546-3141. DOI: [10.2214/AJR.08.1286](https://doi.org/10.2214/AJR.08.1286).

- [38] Akihiko Tamura et al. “Potential asbestos exposure among patients with primary lung cancer in Japan”. In: *Journal of Occupational Health* 60.3 (May 2018), pp. 236–245. ISSN: 1341-9145, 1348-9585. DOI: [10.1539/joh.2017-0199-OA](https://doi.org/10.1539/joh.2017-0199-OA).
- [39] Gerald F. Abbott et al. “Imaging of Small Airways Disease”. In: *Journal of Thoracic Imaging* 24.4 (Nov. 2009), pp. 285–298. ISSN: 0883-5993. DOI: [10.1097/RTI.0b013e3181c1ab83](https://doi.org/10.1097/RTI.0b013e3181c1ab83).
- [40] David M. Hansell et al. “Fleischner Society: Glossary of Terms for Thoracic Imaging”. In: *Radiology* 246.3 (Mar. 2008), pp. 697–722. ISSN: 0033-8419, 1527-1315. DOI: [10.1148/radiol.2462070712](https://doi.org/10.1148/radiol.2462070712).
- [41] Maged Hassan et al. “Imaging of pleural disease”. In: *Breathe* 20.1 (Mar. 2024), p. 230172. ISSN: 1810-6838, 2073-4735. DOI: [10.1183/20734735.0172-2023](https://doi.org/10.1183/20734735.0172-2023).
- [42] Takeshi Johkoh et al. “Crazy-paving Appearance at Thin-Section CT: Spectrum of Disease and Pathologic Findings”. In: *Radiology* 211.1 (Apr. 1999), pp. 155–160. ISSN: 0033-8419, 1527-1315. DOI: [10.1148/radiology.211.1.r99ap10155](https://doi.org/10.1148/radiology.211.1.r99ap10155).
- [43] Rupali Jain et al. “Spontaneous Pneumothorax: A Review of Underlying Etiologies and Diagnostic Imaging Modalities”. In: *Tomography* 11.11 (Nov. 2025), p. 125. ISSN: 2379-139X. DOI: [10.3390/tomography11110125](https://doi.org/10.3390/tomography11110125).
- [44] Natanael Hernández-Vázquez et al. “Fibrous Tissue Semantic Segmentation in CT Images of Diffuse Interstitial Lung Disease”. In: *Journal of Imaging Informatics in Medicine* (Feb. 2025). ISSN: 2948-2933. DOI: [10.1007/s10278-025-01420-x](https://doi.org/10.1007/s10278-025-01420-x).
- [45] Guillaume Chassagnon et al. “Deep Learning–based Approach for Automated Assessment of Interstitial Lung Disease in Systemic Sclerosis on CT Images”. In: *Radiology: Artificial Intelligence* 2.4 (July 2020), e190006. ISSN: 2638-6100. DOI: [10.1148/ryai.2020190006](https://doi.org/10.1148/ryai.2020190006).
- [46] Xiaohong Liu, Kai Wang, and Ting Chen. “Deep Active Learning For Fibrosis Segmentation Of Chest CT Scans From Covid-19 Patients”. In: *2021 IEEE International Conference on Image Processing (ICIP)*. Anchorage, AK, USA: IEEE, Sept. 2021, pp. 175–179. ISBN: 978-1-6654-4115-5. DOI: [10.1109/ICIP42928.2021.9506413](https://doi.org/10.1109/ICIP42928.2021.9506413).
- [47] Yexin Lai et al. “Deep Reinforcement Learning Driven Weakly Supervised Lesion Segmentation in Lung CT Sequence Images”. In: *IEEE Transactions on Instrumentation and Measurement* 74 (2025), pp. 1–10. ISSN: 0018-9456, 1557-9662. DOI: [10.1109/TIM.2025.3571146](https://doi.org/10.1109/TIM.2025.3571146).
- [48] Muhunthan Thillai et al. “Deep Learning–based Segmentation of Computed Tomography Scans Predicts Disease Progression and Mortality in Idiopathic Pulmonary Fibrosis”. In: *American Journal of Respiratory and Critical Care Medicine* 210.4 (Aug. 2024), pp. 465–472. ISSN: 1073-449X, 1535-4970. DOI: [10.1164/rccm.202311-2185OC](https://doi.org/10.1164/rccm.202311-2185OC).
- [49] M. Vakalopoulou et al. “Deep patch-based priors under a fully convolutional encoder-decoder architecture for interstitial lung disease segmentation”. In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. Washington, DC: IEEE, Apr. 2018, pp. 855–858. ISBN: 978-1-5386-3636-7. DOI: [10.1109/ISBI.2018.8363706](https://doi.org/10.1109/ISBI.2018.8363706).
- [50] Fan Hou et al. “Quantitative assessment of interstitial lung disease based on RDNet convolutional network”. In: *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Las Vegas, NV, USA: IEEE, Dec. 2022, pp. 1550–1553. ISBN: 978-1-6654-6819-0. DOI: [10.1109/BIBM55620.2022.9995328](https://doi.org/10.1109/BIBM55620.2022.9995328).

- [51] Yexin Lai et al. “Severity-stratification of interstitial lung disease by deep learning enabled assessment and quantification of lesion indicators from HRCT images”. In: *Journal of X-Ray Science and Technology: Clinical Applications of Diagnosis and Therapeutics* 32.2 (Mar. 2024), pp. 323–338. ISSN: 0895-3996, 1095-9114. DOI: [10.3233/XST-230218](https://doi.org/10.3233/XST-230218).
- [52] Sungwon Ham et al. “Enhancement of Semantic Segmentation by Image-Level Fine-Tuning to Overcome Image Pattern Imbalance in HRCT of Diffuse Infiltrative Lung Diseases”. In: *International Journal of Imaging Systems and Technology* 34.5 (Sept. 2024), e23188. ISSN: 0899-9457, 1098-1098. DOI: [10.1002/ima.23188](https://doi.org/10.1002/ima.23188).
- [53] M. Fontanellaz et al. “Computer-Aided Diagnosis System for Lung Fibrosis: From the Effect of Radiomic Features and Multi-Layer-Perceptron Mixers to Pre-Clinical Evaluation”. In: *IEEE Access* 12 (2024), pp. 25642–25656. ISSN: 2169-3536. DOI: [10.1109/ACCESS.2024.3350430](https://doi.org/10.1109/ACCESS.2024.3350430).
- [54] Mingchen Gao, Ziyue Xu, and Daniel J. Mollura. “Interstitial Lung Diseases via Deep Convolutional Neural Networks: Segmentation Label Propagation, Unordered Pooling and Cross-Dataset Learning”. In: *Deep Learning and Convolutional Neural Networks for Medical Image Computing*. Ed. by Le Lu et al. Series Title: Advances in Computer Vision and Pattern Recognition. Cham: Springer International Publishing, 2017, pp. 97–111. ISBN: 978-3-319-42998-4 978-3-319-42999-1. DOI: [10.1007/978-3-319-42999-1_7](https://doi.org/10.1007/978-3-319-42999-1_7).
- [55] Marios Anthimopoulos et al. “Semantic Segmentation of Pathological Lung Tissue With Dilated Fully Convolutional Networks”. In: *IEEE Journal of Biomedical and Health Informatics* 23.2 (Mar. 2019), pp. 714–722. ISSN: 2168-2194, 2168-2208. DOI: [10.1109/JBHI.2018.2818620](https://doi.org/10.1109/JBHI.2018.2818620).
- [56] Andreas Christe et al. “Computer-Aided Diagnosis of Pulmonary Fibrosis Using Deep Learning and CT Images”. In: *Investigative Radiology* 54.10 (Oct. 2019), pp. 627–632. ISSN: 1536-0210, 0020-9996. DOI: [10.1097/RLI.0000000000000574](https://doi.org/10.1097/RLI.0000000000000574).
- [57] Guang-Wei Cai et al. “Semi-Supervised Segmentation of Interstitial Lung Disease Patterns from CT Images via Self-Training with Selective Re-Training”. In: *Bioengineering* 10.7 (July 2023), p. 830. ISSN: 2306-5354. DOI: [10.3390/bioengineering10070830](https://doi.org/10.3390/bioengineering10070830).
- [58] Adrien Depeursinge et al. “Building a reference multimedia database for interstitial lung diseases”. In: *Computerized Medical Imaging and Graphics* 36.3 (Apr. 2012), pp. 227–238. ISSN: 08956111. DOI: [10.1016/j.compmedimag.2011.07.003](https://doi.org/10.1016/j.compmedimag.2011.07.003).
- [59] *BioLINCC: Lung Tissue Research Consortium (LTRC)*. URL: <https://biolincc.nhlbi.nih.gov/studies/ltrc/> (visited on 12/30/2025).
- [60] Joseph Paul Cohen, Paul Morrison, and Lan Dao. *COVID-19 Image Data Collection*. arXiv:2003.11597 [eess]. Mar. 2020. DOI: [10.48550/arXiv.2003.11597](https://doi.org/10.48550/arXiv.2003.11597).
- [61] Samuel G. Armato et al. “The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans”. In: *Medical Physics* 38.2 (Feb. 2011), pp. 915–931. ISSN: 0094-2405, 2473-4209. DOI: [10.1118/1.3528204](https://doi.org/10.1118/1.3528204).
- [62] Yuxiao Gao et al. “Medical Image Segmentation: A Comprehensive Review of Deep Learning-Based Methods”. In: *Tomography* 11.5 (Apr. 2025), p. 52. ISSN: 2379-139X. DOI: [10.3390/tomography11050052](https://doi.org/10.3390/tomography11050052).
- [63] Justin Ker et al. “Deep Learning Applications in Medical Image Analysis”. In: *IEEE Access* 6 (2018), pp. 9375–9389. ISSN: 2169-3536. DOI: [10.1109/ACCESS.2017.2788044](https://doi.org/10.1109/ACCESS.2017.2788044).

- [64] Abeer Aljuaid and Mohd Anwar. “Survey of Supervised Learning for Medical Image Processing”. In: *SN Computer Science* 3.4 (May 2022), p. 292. ISSN: 2661-8907. DOI: [10.1007/s42979-022-01166-1](https://doi.org/10.1007/s42979-022-01166-1).
- [65] Johannes Hofmanninger et al. “Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem”. In: *European Radiology Experimental* 4.1 (Dec. 2020), p. 50. ISSN: 2509-9280. DOI: [10.1186/s41747-020-00173-2](https://doi.org/10.1186/s41747-020-00173-2).
- [66] Alejandro F. Frangi et al. “Multiscale vessel enhancement filtering”. In: *Medical Image Computing and Computer-Assisted Intervention — MICCAI’98*. Ed. by William M. Wells, Alan Colchester, and Scott Delp. Vol. 1496. Series Title: Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 130–137. ISBN: 978-3-540-65136-9 978-3-540-49563-5. DOI: [10.1007/BFb0056195](https://doi.org/10.1007/BFb0056195).
- [67] Guotai Wang et al. “DeepIGeoS: A Deep Interactive Geodesic Framework for Medical Image Segmentation”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.7 (July 2019), pp. 1559–1572. ISSN: 0162-8828, 2160-9292, 1939-3539. DOI: [10.1109/TPAMI.2018.2840695](https://doi.org/10.1109/TPAMI.2018.2840695).
- [68] Risheng Wang et al. “Medical image segmentation using deep learning: A survey”. In: *IET Image Processing* 16.5 (Apr. 2022), pp. 1243–1267. ISSN: 1751-9659, 1751-9667. DOI: [10.1049/ipr2.12419](https://doi.org/10.1049/ipr2.12419).
- [69] Xinyue Zhang et al. *A Review of Non-Fully Supervised Deep Learning for Medical Image Segmentation*. Apr. 2025. DOI: [10.20944/preprints202504.0460.v1](https://doi.org/10.20944/preprints202504.0460.v1).
- [70] Zahra Solatidehkordi and Imran Zualkernan. “Survey on Recent Trends in Medical Image Classification Using Semi-Supervised Learning”. In: *Applied Sciences* 12.23 (Nov. 2022), p. 12094. ISSN: 2076-3417. DOI: [10.3390/app122312094](https://doi.org/10.3390/app122312094).
- [71] Burr Settles. *Active Learning Literature Survey*. Computer Sciences Technical Report 1648. University of Wisconsin–Madison, 2009.
- [72] A. B. M. Kamrul Islam Riad et al. “Reinforcement Learning in Medical Imaging: Taxonomy, LLMs, and Clinical Challenges”. In: *Future Internet* 17.9 (Aug. 2025), p. 396. ISSN: 1999-5903. DOI: [10.3390/fi17090396](https://doi.org/10.3390/fi17090396).
- [73] Mingzhe Hu et al. “Reinforcement learning in medical image analysis: Concepts, applications, challenges, and future directions”. In: *Journal of Applied Clinical Medical Physics* 24.2 (Feb. 2023), e13898. ISSN: 1526-9914, 1526-9914. DOI: [10.1002/acm2.13898](https://doi.org/10.1002/acm2.13898).
- [74] Pengzhen Ren et al. *A Survey of Deep Active Learning*. arXiv:2009.00236. Dec. 2021. DOI: [10.48550/arXiv.2009.00236](https://doi.org/10.48550/arXiv.2009.00236).
- [75] Samuel Budd, Emma C. Robinson, and Bernhard Kainz. “A survey on active learning and human-in-the-loop deep learning for medical image analysis”. In: *Medical Image Analysis* 71 (July 2021), p. 102062. ISSN: 13618415. DOI: [10.1016/j.media.2021.102062](https://doi.org/10.1016/j.media.2021.102062).
- [76] S. Kevin Zhou et al. *Deep reinforcement learning in medical imaging: A literature review*. arXiv:2103.05115 [eess]. Mar. 2021. DOI: [10.48550/arXiv.2103.05115](https://doi.org/10.48550/arXiv.2103.05115).
- [77] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. “Efficient Graph-Based Image Segmentation”. In: *International Journal of Computer Vision* 59.2 (Sept. 2004), pp. 167–181. ISSN: 0920-5691, 1573-1405. DOI: [10.1023/B:VISI.0000022288.19776.77](https://doi.org/10.1023/B:VISI.0000022288.19776.77).
- [78] Xi Wang et al. “Deep semi-supervised multiple instance learning with self-correction for DME classification from OCT images”. In: *Medical Image Analysis* 83 (Jan. 2023), p. 102673. ISSN: 13618415. DOI: [10.1016/j.media.2022.102673](https://doi.org/10.1016/j.media.2022.102673).

- [79] Isaac Triguero, Salvador García, and Francisco Herrera. “Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study”. In: *Knowledge and Information Systems* 42.2 (Feb. 2015), pp. 245–284. ISSN: 0219-1377, 0219-3116. DOI: [10.1007/s10115-013-0706-y](https://doi.org/10.1007/s10115-013-0706-y).
- [80] Rushi Jiao et al. “Learning with limited annotations: A survey on deep semi-supervised learning for medical image segmentation”. In: *Computers in Biology and Medicine* 169 (Feb. 2024), p. 107840. ISSN: 00104825. DOI: [10.1016/j.compbio.2023.107840](https://doi.org/10.1016/j.compbio.2023.107840).
- [81] Dong-Hyun Lee. “Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks”. In: *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)* (July 2013).
- [82] Xiaojin Zhu and Andrew B. Goldberg. *Introduction to Semi-Supervised Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Cham: Springer International Publishing, 2009. ISBN: 978-3-031-00420-9 978-3-031-01548-9. DOI: [10.1007/978-3-031-01548-9](https://doi.org/10.1007/978-3-031-01548-9).
- [83] Yuyin Zhou et al. “Semi-Supervised 3D Abdominal Multi-Organ Segmentation Via Deep Multi-Planar Co-Training”. In: *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Waikoloa Village, HI, USA: IEEE, Jan. 2019, pp. 121–140. ISBN: 978-1-7281-1975-5. DOI: [10.1109/WACV.2019.00020](https://doi.org/10.1109/WACV.2019.00020).
- [84] Hoel Kervadec et al. “Curriculum Semi-supervised Segmentation”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Ed. by Dinggang Shen et al. Vol. 11765. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 568–576. ISBN: 978-3-030-32244-1 978-3-030-32245-8. DOI: [10.1007/978-3-030-32245-8_63](https://doi.org/10.1007/978-3-030-32245-8_63).
- [85] Juzheng Miao et al. “SC-SSL: Self-Correcting Collaborative and Contrastive Co-Training Model for Semi-Supervised Medical Image Segmentation”. In: *IEEE Transactions on Medical Imaging* 43.4 (2024), pp. 1347–1364. DOI: [10.1109/TMI.2023.3336534](https://doi.org/10.1109/TMI.2023.3336534).
- [86] Shaobo Min et al. *A Two-Stream Mutual Attention Network for Semi-supervised Biomedical Segmentation with Noisy Labels*. arXiv:1807.11719. Dec. 2018. DOI: [10.48550/arXiv.1807.11719](https://doi.org/10.48550/arXiv.1807.11719).
- [87] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. “Regularization With Stochastic Transformations and Perturbations for Deep Semi-Supervised Learning”. In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee et al. Vol. 29. Curran Associates, Inc., 2016. URL: https://proceedings.neurips.cc/paper_files/paper/2016/file/30ef30b64204a3088a26bc2e6ecf7602-Paper.pdf.
- [88] Samuli Laine and Timo Aila. *Temporal Ensembling for Semi-Supervised Learning*. arXiv:1610.02242. Mar. 2017. DOI: [10.48550/arXiv.1610.02242](https://doi.org/10.48550/arXiv.1610.02242).
- [89] Antti Tarvainen and Harri Valpola. *Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results*. arXiv:1703.01780. Apr. 2018. DOI: [10.48550/arXiv.1703.01780](https://doi.org/10.48550/arXiv.1703.01780).
- [90] Lequan Yu et al. “Uncertainty-Aware Self-ensembling Model for Semi-supervised 3D Left Atrium Segmentation”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Ed. by Dinggang Shen et al. Vol. 11765. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 605–613. ISBN: 978-3-030-32244-1 978-3-030-32245-8. DOI: [10.1007/978-3-030-32245-8_67](https://doi.org/10.1007/978-3-030-32245-8_67).

- [91] Sukesh Adiga V., Jose Dolz, and Herve Lombaert. “Anatomically-aware uncertainty for semi-supervised image segmentation”. In: *Medical Image Analysis* 91 (Jan. 2024), p. 103011. ISSN: 13618415. DOI: [10.1016/j.media.2023.103011](https://doi.org/10.1016/j.media.2023.103011).
- [92] Wei Li et al. “Diversity matters: Cross-head mutual mean-teaching for semi-supervised medical image segmentation”. In: *Medical Image Analysis* 97 (Oct. 2024), p. 103302. ISSN: 13618415. DOI: [10.1016/j.media.2024.103302](https://doi.org/10.1016/j.media.2024.103302).
- [93] Yicheng Wu et al. “Exploring Smoothness and Class-Separation for Semi-supervised Medical Image Segmentation”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Ed. by Linwei Wang et al. Vol. 13435. Series Title: Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022, pp. 34–43. ISBN: 978-3-031-16442-2 978-3-031-16443-9. DOI: [10.1007/978-3-031-16443-9_4](https://doi.org/10.1007/978-3-031-16443-9_4).
- [94] Kihyuk Sohn et al. *FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence*. 2020. DOI: [10.48550/ARXIV.2001.07685](https://doi.org/10.48550/ARXIV.2001.07685).
- [95] Yunxin Zhong et al. “Pixelfixmatch: A Semi-Supervised Image Segmentation Method Based on Fixmatch with Pixel Attention”. In: *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. 2024, pp. 1–5. DOI: [10.1109/ISBI56570.2024.10635605](https://doi.org/10.1109/ISBI56570.2024.10635605).
- [96] Jiawei Su et al. “Mutual learning with reliable pseudo label for semi-supervised medical image segmentation”. In: *Medical Image Analysis* 94 (May 2024), p. 103111. ISSN: 13618415. DOI: [10.1016/j.media.2024.103111](https://doi.org/10.1016/j.media.2024.103111).
- [97] Zhaohan Xiong et al. “A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging”. In: *Medical Image Analysis* 67 (2021), p. 101832. ISSN: 1361-8415. DOI: <https://doi.org/10.1016/j.media.2020.101832>.
- [98] Yunhao Bai et al. *Bidirectional Copy-Paste for Semi-Supervised Medical Image Segmentation*. arXiv:2305.00673. May 2023. DOI: [10.48550/arXiv.2305.00673](https://doi.org/10.48550/arXiv.2305.00673).
- [99] Yang Yasng et al. “Semi-supervised medical image segmentation via weak-to-strong perturbation consistency and edge-aware contrastive representation”. In: *Medical Image Analysis* 101 (Apr. 2025), p. 103450. ISSN: 13618415. DOI: [10.1016/j.media.2024.103450](https://doi.org/10.1016/j.media.2024.103450).
- [100] Fabian Isensee et al. “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation”. In: *Nature Methods* 18.2 (Feb. 2021), pp. 203–211. ISSN: 1548-7091, 1548-7105. DOI: [10.1038/s41592-020-01008-z](https://doi.org/10.1038/s41592-020-01008-z).
- [101] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by Nassir Navab et al. Vol. 9351. Cham: Springer International Publishing, 2015, pp. 234–241. ISBN: 978-3-319-24573-7 978-3-319-24574-4. DOI: [10.1007/978-3-319-24574-4_28](https://doi.org/10.1007/978-3-319-24574-4_28).
- [102] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: [10.1109/CVPR.2016.90](https://doi.org/10.1109/CVPR.2016.90).
- [103] Sepp Hochreiter. “The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions”. In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 6 (Apr. 1998), pp. 107–116. DOI: [10.1142/S0218488598000094](https://doi.org/10.1142/S0218488598000094).

- [104] Yves Grandvalet and Yoshua Bengio. “Semi-supervised learning by entropy minimization”. In: *Proceedings of the 18th International Conference on Neural Information Processing Systems*. NIPS’04. Vancouver, British Columbia, Canada: MIT Press, 2004, pp. 529–536.
- [105] Johannes Hofmanninger et al. “Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem”. In: *European Radiology Experimental* 4.1 (Aug. 2020). ISSN: 2509-9280. DOI: [10.1186/s41747-020-00173-2](https://doi.org/10.1186/s41747-020-00173-2).
- [106] Yidong Wang et al. *FreeMatch: Self-adaptive Thresholding for Semi-supervised Learning*. arXiv:2205.07246. Jan. 2023. DOI: [10.48550/arXiv.2205.07246](https://doi.org/10.48550/arXiv.2205.07246).
- [107] Hao Chen et al. *SoftMatch: Addressing the Quantity-Quality Trade-off in Semi-supervised Learning*. arXiv:2301.10921. Mar. 2023. DOI: [10.48550/arXiv.2301.10921](https://doi.org/10.48550/arXiv.2301.10921).
- [108] Miguel Nunes et al. “MediAlbertina: An European Portuguese medical language model”. In: *Computers in Biology and Medicine* 182 (2024), p. 109233. ISSN: 0010-4825. DOI: <https://doi.org/10.1016/j.compbiomed.2024.109233>.

Appendix A

Additional Results

This appendix includes the recall, precision, and specificity metrics for the best-performing models identified in chapters 4 and 5, namely the best supervised configuration, the best training-time SSL method, and the best ensemble-based pseudo-labelling model.

A.1 Best Supervised Configuration

Table A.1: Additional segmentation performance metrics for the best supervised configuration measured by recall, precision, and specificity. Values are reported as mean $\pm$ standard deviation (%) across five data splits.

Metric	Fibrosis	Nodules	Ground glass	Consolidation	Cysts	Emphysema	Others	Healthy	Average
Recall	83.35 $\pm$ 7.82	63.19 $\pm$ 12.21	49.11 $\pm$ 16.11	17.15 $\pm$ 5.89	0.00 $\pm$ 0.00	52.98 $\pm$ 23.10	0.01 $\pm$ 0.01	81.04 $\pm$ 15.60	43.35 $\pm$ 31.59
Precision	87.08 $\pm$ 3.98	54.57 $\pm$ 13.44	56.55 $\pm$ 7.99	48.98 $\pm$ 23.68	0.00 $\pm$ 0.00	89.35 $\pm$ 12.17	0.03 $\pm$ 0.06	52.88 $\pm$ 8.63	48.68 $\pm$ 31.58
Specificity	90.59 $\pm$ 3.94	90.95 $\pm$ 6.31	94.95 $\pm$ 2.26	99.46 $\pm$ 0.43	99.99 $\pm$ 0.01	99.19 $\pm$ 0.29	99.42 $\pm$ 0.79	87.82 $\pm$ 6.46	95.30 $\pm$ 4.59

A.2 Best Training-Time SSL Method

Table A.2: Additional segmentation performance metrics for the best training-time SSL method measured by recall, precision, and specificity. Values are reported as mean $\pm$ standard deviation (%) across five data splits.

Metric	Fibrosis	Nodules	Ground glass	Consolidation	Cysts	Emphysema	Others	Healthy	Average
Recall	73.40 $\pm$ 6.62	63.74 $\pm$ 12.96	37.14 $\pm$ 14.52	11.79 $\pm$ 8.41	0.00 $\pm$ 0.00	61.05 $\pm$ 10.94	0.96 $\pm$ 1.33	50.69 $\pm$ 24.72	37.35 $\pm$ 27.63
Precision	88.94 $\pm$ 5.51	53.39 $\pm$ 13.97	53.17 $\pm$ 9.35	53.21 $\pm$ 21.23	0.00 $\pm$ 0.00	94.64 $\pm$ 5.83	0.13 $\pm$ 0.19	28.67 $\pm$ 8.06	46.52 $\pm$ 33.33
Specificity	92.99 $\pm$ 4.24	90.63 $\pm$ 5.07	95.96 $\pm$ 1.19	99.72 $\pm$ 0.25	100.00 $\pm$ 0.00	99.19 $\pm$ 0.93	95.76 $\pm$ 4.85	79.98 $\pm$ 8.00	94.28 $\pm$ 6.23

A.3 Best Ensemble-Based Pseudo-Labelling Model

Table A.3: Additional segmentation performance metrics for the best ensemble-based pseudo-labelling model measured by recall, precision, and specificity. Values are reported as mean $\pm$ standard deviation (%) across five data splits.

Metric	Fibrosis	Nodules	Ground glass	Consolidation	Cysts	Emphysema	Others	Healthy	Average
Recall	83.73 $\pm$ 0.65	66.60 $\pm$ 4.47	74.94 $\pm$ 8.05	16.01 $\pm$ 3.37	0.00 $\pm$ 0.00	66.09 $\pm$ 5.24	0.00 $\pm$ 0.00	79.77 $\pm$ 9.59	48.39 $\pm$ 34.12
Precision	87.01 $\pm$ 1.16	65.31 $\pm$ 12.57	58.17 $\pm$ 4.36	58.76 $\pm$ 11.45	0.00 $\pm$ 0.00	99.26 $\pm$ 0.62	0.00 $\pm$ 0.00	59.57 $\pm$ 5.30	53.51 $\pm$ 33.85
Specificity	90.83 $\pm$ 0.90	94.35 $\pm$ 3.28	92.95 $\pm$ 1.17	99.70 $\pm$ 0.18	100.00 $\pm$ 0.01	99.90 $\pm$ 0.09	99.84 $\pm$ 0.06	91.60 $\pm$ 1.75	96.15 $\pm$ 3.83