Accepted Manuscript

Submission Date: 2024-06-19
Accepted Date: 2025-02-14
Accepted Manuscript online: 2025-02-18

Methods of Information in Medicine

Harnessing Advanced Machine Learning Techniques for Microscopic Vessel Segmentation in Pulmonary Fibrosis Using Novel Hierarchical Phase-Contrast Tomography (HiP-CT) Images

Pardeep Vasudev, Moucheng Xu, Mehran Azimbagarad, Shahab Aslani, Yufei Wang, Robert Chapman, Hannah Coleman, Christopher Werlein, Claire Walsh, Peter Lee, Paul Tafforeau, Joseph Jacob.

Affiliations below.

DOI: 10.1055/a-2540-8166

Please cite this article as: Vasudev P, Xu M, Azimbagarad M et al. Harnessing Advanced Machine Learning Techniques for Microscopic Vessel Segmentation in Pulmonary Fibrosis Using Novel Hierarchical Phase-Contrast Tomography (HiP-CT) Images. Methods of Information in Medicine 2025. doi: 10.1055/a-2540-8166

Conflict of Interest: The authors declare that they have no conflict of interest.

Abstract:
Background: Fibrotic lung disease is a progressive illness that causes scarring and ultimately respiratory failure, with irreversible damage by the time its diagnosed on computed tomography imaging. Recent research postulates the role of the lung vasculature on the pathogenesis of the disease, and with the recent development of high-resolution hierarchical phase contrast tomography (HiP-CT), we have the potential to understand and detect changes in the lungs long before conventional imaging. However, to gain quantitative insight into vascular changes you first need to be able to segment the vessels before further downstream analysis can be conducted. Aside from this, HiP-CT generates large volume, high resolution data which is time consuming and expensive to label. Objectives: This project aims to qualitatively assess the latest machine learning methods for vessel segmentation in HiP-CT data to enable label propagation as the first step for imaging biomarker discovery, with the goal to identify early-stage interstitial lung disease amenable to treatment, before fibrosis begins. Methods: Semi-supervised learning has become a growing method to tackle sparsely labelled datasets due to its leveraging of unlabelled data. In this study we will compare 2 semi-supervised learning methods; Seg PL, based on pseudo labelling and MisMatch, using consistency regularisation against state of the art supervised learning method, in nnU-Net, on vessel segmentation in sparsely labelled lung HiP-CT data. Results: On initial experimentation, both MisMatch and SegPL showed promising performance on qualitative review. In comparison with supervised learning, both MisMatch and SegPL showed better on out of distribution performance within the same sample (different vessel morphology and texture vessels), though supervised learning provided more consistent segmentations for well represented labels in the limited annotations. Conclusion: Further quantitative research is required to better assess the generalisability of these findings, though they show promising first steps towards leveraging this novel data to tackle fibrotic lung disease.

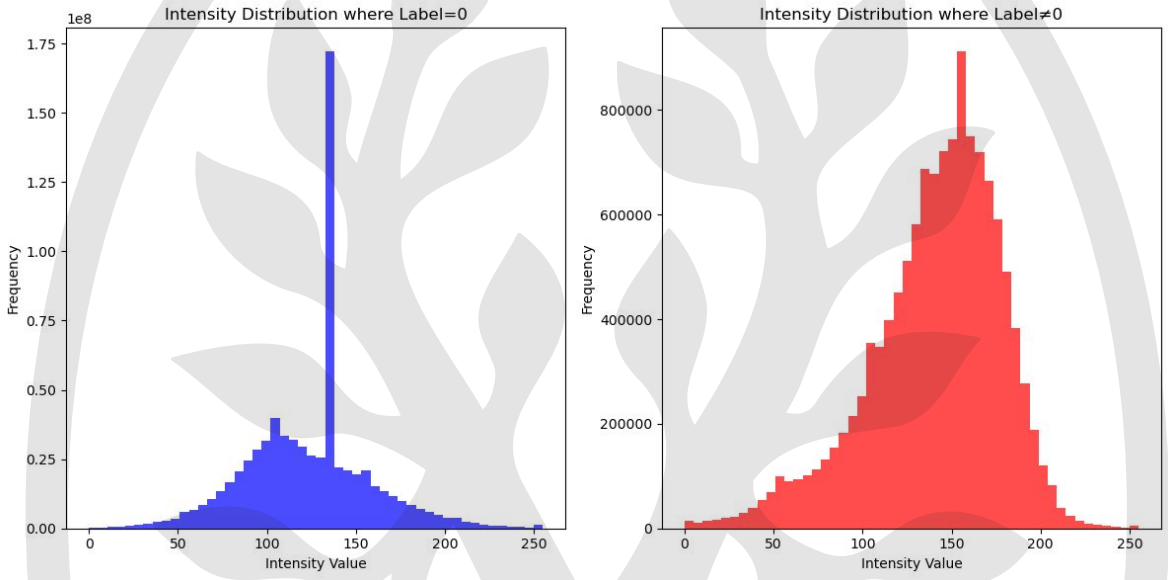
Corresponding Author:
Dr. Pardeep Vasudev, University College London Institute of Health Informatics, 222 Euston Road, NW1 2DA London, United Kingdom of Great Britain and Northern Ireland, pardeep.vasudev.19@ucl.ac.uk

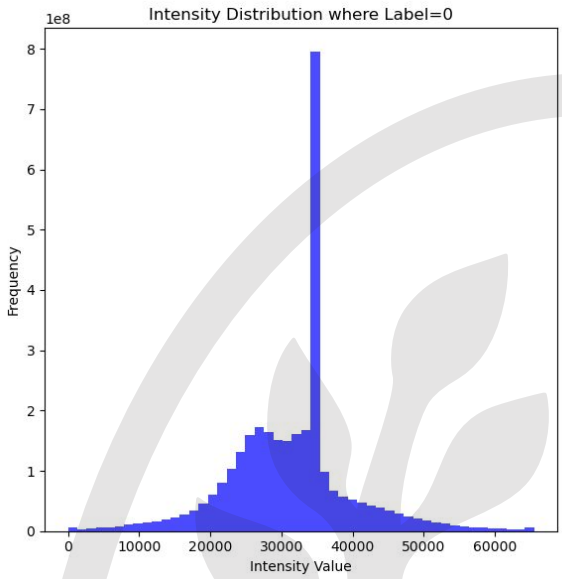
Affiliations:
Pardeep Vasudev, University College London Institute of Health Informatics, London, United Kingdom of Great Britain and Northern Ireland
Pardeep Vasudev, University College London Centre for Medical Image Computing, London, United Kingdom of Great Britain and

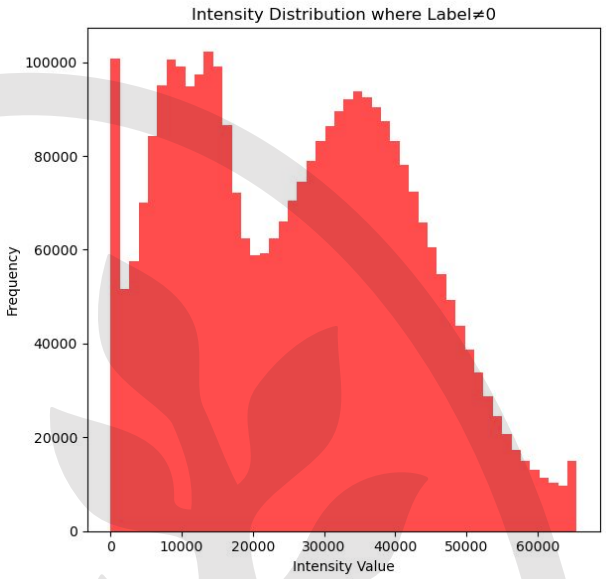
Northern Ireland
Moucheng Xu, University College London Centre for Medical Image Computing, London, United Kingdom of Great Britain and Northern Ireland
[...]
Joseph Jacob, University College London Centre for Medical Image Computing, London, United Kingdom of Great Britain and Northern Ireland

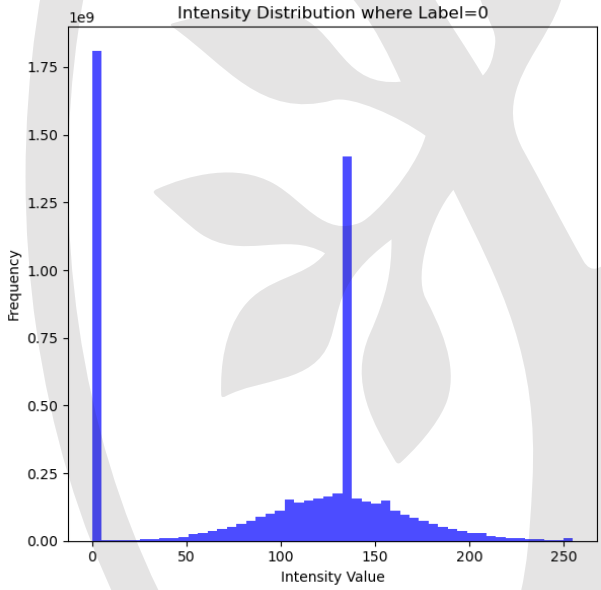
Appendix

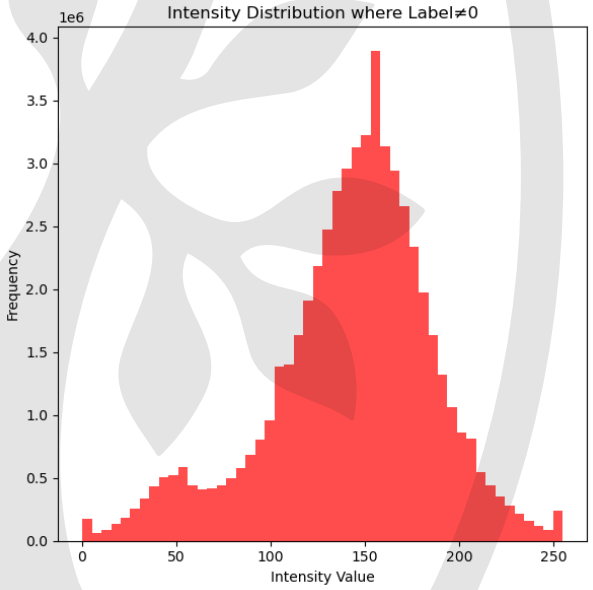
Appendix 1: Separate patches where background intensities and vessel intensities defined by labelled areas in patch have been compared. Label ≠ 0 is vessel and Label = 0 is everything else including background. There is subtle difference in peaks though significant overlap.

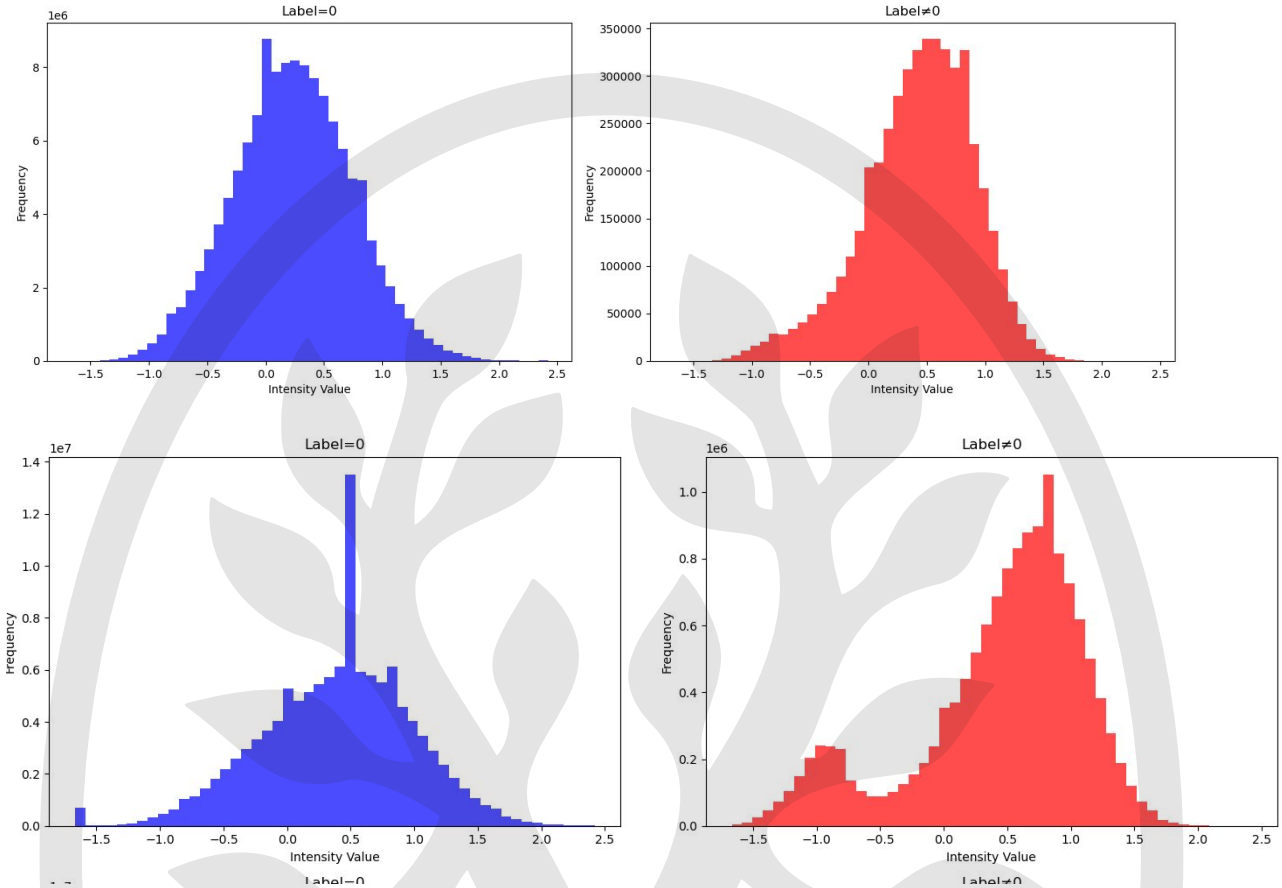




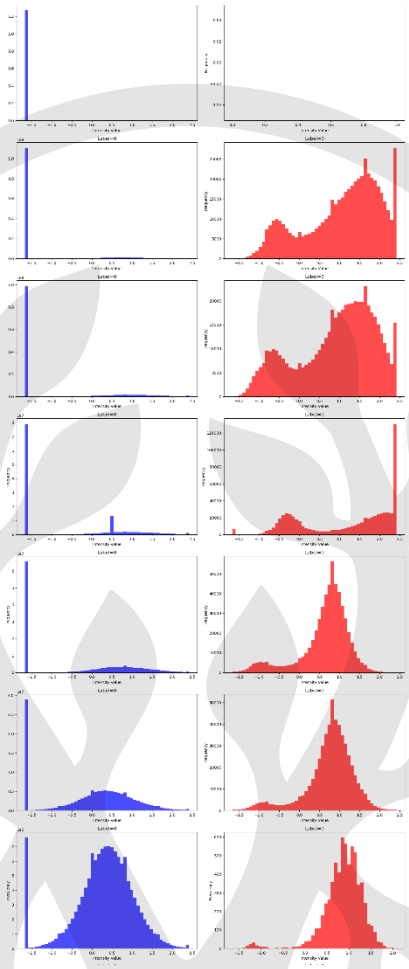








Appendix 2: Separate patches where background intensities and vessel intensities defined by labelled areas in smaller patches (size 128 x 128) have been compared. Label ≠ 0 is vessel and Label = 0 is everything else including background. In the 10 smaller patches, the differences in the distribution of the vessels becomes more apparent.



Appendix 3: Example Hyperparameters to tune for MisMatch Model

Hyperparameter	Explanation	Value in Paper
Batch size	Batch size of labelled volumes.	1
Optimiser	This specifies the algorithm used for weight optimisation across nodes	Adam

Learning Rate		2e-5
alpha	Consistency regularization weight	0.002
epochs	Number of epochs	50

Appendix 1: Example hyperparameters to tune for SegPL(BPL)

Hyperparameter	Explanation	Value in Paper
Batch Size	Batch size of labelled volumes.	2
Optimiser	This specifies the algorithm used for weight optimisation across nodes	Adam
Learning Rate		0.001
temp	Temperature scaling on output	1
Batch_u	Set to 0 for supervised setting.	2
Pri_mu	Mean of prior	**0.7**
Pri_std	Standard deviation of Prior	0.15
alpha	Weight on the unsupervised part if semi-supervised learning is used	**1**
beta	Weight for pseudo supervision loss	**1**
warmup	Ratio between warm up iterations and total	**0.1**

	iterations	
Warmup_start	Ratio between warm up starting iteration and total iterations.	**0.4**

Appendix 2: nn-UNet performance on Dataset B for differently prepared datasets.

Training Datasets (A + C)	IoU	DICE	FN pixels	FP pixels	TN pixels	TP pixels	Predicted label pixels	Total label pixels
Full Volumes	0.6057	0.7544	2846910	3431312	715209124	9645034	13076346	12491944
128 patches	6.40E-07	1.28E-06	12491936	35	718640401	8	43	12491944
256 patches	0	0	12491944	0	718640436	0	0	12491944
512 patches	0	0	12491944	0	718640436	0	0	12491944
128 patches with 1 percent thereshold	0.0020	0.0041	1246076	162162	718478274	25868	188030	12491944
256 patches with 1 percent threshold	0.0147	0.0290	6821466	37240682 0	346233616	567047 8	378077298	12491944
512 patched with 1 percent threshold	0.0048	0.0097	12428839	434954	718205482	63105	498059	12491944

Appendix 6: Table showing experiments carried out on SegPL, without validation sets on Datasets A+C. The left-hand column are the datasets by patch size (128, 256, 512) and further divided by the percentage of thresholding (t). Alpha, batch , primu refer to different hyperparameters.

	train iou ema (smoothing 0.60)	train seg loss (smoothing 0.60)	epoch	train time (days)
Patch size 128				
t0	0.8612	0.2032	37036	2.309
t1	0.939	0.0824	31409	2.311
t2.5	0.936	0.06175	35095	2.27
t5	0.9637	0.04149	27016	2.305
alpha 2.0 batch 8	0.9742	0.03204	31806	5.625
alpha 0.9 batch 8	0.9698	0.03368	29376	5
t2.5				
alpha 0.1	0.954	0.05073	49380	2.315
Alpha 0.5	0.9554	0.0686	33150	2.312
Alpha 2.0	0.9499	0.0683	33483	2.312
Batch 4	0.9641	0.0694	22891	2.778
Batch 8	0.9666	0.0633	13409	2.775
primu 0.5	0.9656	0.0871	31174	2.312
primu 0.7	0.943	0.0598	32104	2.312
Patch size 256				
t0	0.6801	0.4355	10079	2.298
t1	0.8305	0.2578	26820	4.617
t2.5	0.8714	0.2187	22100	4.617
t5	0.8798	0.2295	14882	4.617
Patch size 512				
t0	0.5251	0.6273	1297	2.29
t1	0.6266	0.5432	1539	2.304
t2.5	0.6658	0.4908	2819	4.617
t5	0.8049	0.397	4364	4.617

Appendix 7: Table showing experiments carried out on SegPL, with validation sets on Datasets A. The left hand column are the datasets by patch size (128, 256) and further divided by the percentage of thresholding (t). Alpha, batch , primu refer to different hyperparameters.

	best train iou	best validation iou	train seg loss (smoothing	epoch	train time (days)

			0.60)		
128					
t0	0.8755	0.6099	0.2032	26596	2.315
t1	0.9394	0.709	0.061	50000	4.2
t2.5	0.9751	0.7167	0.0488	50000	4.6
t5	0.9608	0.702	0.0569	18514	1.197
t2.5					
Alpha 0.5	0.9599	0.7045	0.0463	50000	2.312
Alpha 2.0	0.9499	0.7061	0.0571	33750	2.312
Batch 4	0.9601	0.7115	0.0618	23744	2.775
Batch 8	0.9636	0.7117	0.0556	25195	4.6
primu 0.5	0.9656	0.7095	0.0592	33000	2.312
primu 0.7	0.9457	0.7003	0.0598	27230	2.312
256					
t0	0.6688	0.4362	0.4842	6742	2.3
t1	0.8919	0.5788	0.301	10827	5.2
t2.5	0.8215	0.6411	0.275	10500	5.2
t5	0.8773	0.634	0.2208	11085	5.2

Appendix 8: Table showing experiments conducted on MIsMatch, without validation sets on Datasets A+C. The left hand column are the datasets by patch size (128, 256) and further divided by the percentage of thresholding (t). Con (consistency), batch , lbl (the number of labelled) data in the dataset, refer to different hyperparameters.

	loss	loss_seg	loss_seg_dice	epoch	train time (hours)
Patch size128					
t0	0.4705	0.04732	0.6182	51720	2.302
t1	0.3805	0.05934	0.5355	9863	10.12
t2.5	0.3176	0.06803	0.4779	80272	6.5
t5	0.3039	0.08395	0.4032	29614	22.13
t2.5					
con 0.5	0.4121	0.0907	0.4358	20668	12.65

con 5.0					
Batch 8	0.2617	0.04731	0.398	48197	2.308
Batch 16	0.3368	0.07866	0.5027	28268	2.308
lbl 4	0.2683	0.02773	0.5051	56139	2.312
lbl 8	0.2923	0.0378	0.4966	33539	2.687
Patch size 256					
t0	0.4812	0.01531	0.7267	77640	4.6
t1	0.4049	0.02524	0.4709	117356	5.8
t2.5	0.3166	0.03799	0.5539	80192	7
t5	0.3605	0.03999	0.5457	45747	2.3

Appendix 9: Table showing experiments conducted on MisMatch, with validation sets on Datasets A. The left-hand column are the datasets by patch size (128) and further divided by the percentage of thresholding (t). Con (consistency), lbl (the number of labelled) data in the dataset, refer to different hyperparameters. Further experiments were not possible with larger patch sizes due to lack of memory.

	Train loss	Validation loss	Train loss_seg	Validation loss_seg	Train loss_seg_dice	epoch	train time hr
Patch 128							
t5	0.217	0.2904	0.2655	0.4443	0.3139	120000	3.44
t2.5							
con 0.5	0.3094	0.2787	0.2148	0.4504	0.296	na	na
con 5.0	0.26	0.285	0.2496	0.4574	0.02297	80000	na
lbl 4	0.2903	0.2753	0.2489	0.5417	0.3016	70000	na
lbl 8	0.2922	0.3001	0.2762	0.5325	0.3132	22500	2.164

Harnessing Advanced Machine Learning Techniques for Microscopic Vessel Segmentation in Pulmonary Fibrosis Using Novel Hierarchical Phase-Contrast Tomography (HiP-CT) Images

Pardeep Vasudev, Institute of Health Informatics, University College London, London, UK; Centre of Medical Image Computing, University College London. London, UK

Mehran Azimbagarad, Centre of Medical Image Computing, University College London, London, UK

Shahab Aslani, Centre of Medical Image Computing, University College London, London, UK

Moucheng Xu, Centre of Medical Image Computing, University College London, London, UK

Yufei Wang, Department of Mechanical Engineering, University College London, London, UK

Robert Chapman, Division of Medicine, University College London, London, UK.

Hannah Coleman, Centre for Advanced Biomedical Imaging, University College London, London, UK

Christopher Werlein, Institute of Pathology, Hannover Medical School, Hannover, Germany

Claire Walsh, Department of Mechanical Engineering, University College London, London, UK; Centre for Advanced Biomedical Imaging, University College London, London, UK

Peter Lee, Department of Mechanical Engineering, University College London, London, UK

Paul Tafforeau, European Synchrotron Radiation Facility, Grenoble, France

Joseph Jacob*, Centre for Medical Image Computing, University College London, London, UK, UCL Respiratory, University College London, London, UK

*Corresponding Author:

Joseph Jacob, MD, Centre for Medical Image Computing, University College London, London, UK, UCL Respiratory, University College London, London, UK; The UCL Centre for Medical Image Computing (CMIC), 90 High Holborn, Floor 1, London, WC1V 6LJ, United Kingdom; j.jacob@ucl.ac.uk

Structured Abstract

Background: Fibrotic lung disease is a progressive illness that causes scarring and ultimately respiratory failure, with irreversible damage by the time its diagnosed on computed tomography imaging. Recent research postulates the role of the lung vasculature on the pathogenesis of the disease, and with the recent development of high-resolution hierarchical phase contrast tomography (HiP-CT), we have the potential to understand and detect changes in the lungs long before conventional imaging. However, to gain quantitative insight into vascular changes you first need to be able to segment the vessels before further downstream analysis can be conducted. Aside from this, HiP-CT generates large volume, high resolution data which is time consuming and expensive to label. Objectives: This project aims to qualitatively assess the latest machine learning methods for vessel segmentation in HiP-CT data to enable label propagation as the first step for imaging biomarker discovery, with the goal to identify early-stage interstitial lung disease amenable to treatment, before fibrosis begins. Methods: Semi-supervised learning has

become a growing method to tackle sparsely labelled datasets due to its leveraging of unlabelled data. In this study we will compare 2 semi-supervised learning methods; Seg PL, based on pseudo labelling and MisMatch, using consistency regularisation against state of the art supervised learning method, nnU-Net, on vessel segmentation in sparsely labelled lung HiP-CT data. Results: On initial experimentation, both MisMatch and SegPL showed promising performance on qualitative review. In comparison with supervised learning, both MisMatch and SegPL showed better on out of distribution performance within the same sample (different vessel morphology and texture vessels), though supervised learning provided more consistent segmentations for well represented labels in the limited annotations. Conclusion: Further quantitative research is required to better assess the generalisability of these findings, though they show promising first steps towards leveraging this novel data to tackle fibrotic lung disease.

Keywords

Vessel Segmentation, Hierarchical Phase-Contrast Tomography, Semi-Supervised Learning, Pulmonary Fibrosis, Interstitial Lung Disease

Introduction

The NHS aims to improve lung cancer detection, by expanding the Lung Health Check program for individuals aged 55-74 with a GP record and a history of smoking.[1] High-risk individuals will receive a CT scan every two years, the standard method for lung cancer detection in the UK.[2] Currently available in limited UK centres, the NHS aims for full coverage by 2029, increasing

annual CT scans from 200,000 to 1 million.[3] Approximately 2% of cases will have lung cancer, and about 1.5% will have pulmonary fibrosis.[4]

Pulmonary fibrosis describes presence of lung thickening and scarring, resulting in symptoms such as shortness of breath and a cough.[5] It can be the result of a variety of heterogeneous conditions that can have varying prognosis, including interstitial lung diseases (ILD) such as Idiopathic Pulmonary Fibrosis (IPF) and Pleuro-parenchymal Fibroelastosis (PPFE), which have median survival times of 2.5 to 5 years.[6,7] IPF is characterized by progressive scarring of the lungs with alveolar destruction, leaving the lungs stiff with decreasing ability for gaseous exchange, causing significant morbidity and eventually respiratory failure, with a median survival of 3-4 years without treatment.[8] By contrast, PPFE is a relatively rare ILD which involves scarring of the upper lobes involving the pleura and subpleural lung.[9] The prognosis varies depending on phenotype, though is worse for those with pre-existing IPF and may be worse in the late stages of idiopathic PPFE than IPF.[6,10,11] Early-stage disease is often asymptomatic or presents with non-specific symptoms, such as progressive shortness of breath, cough, and lethargy.[12] Currently pulmonary fibrosis is best assessed on CT, though visual analysis methods were designed for the description of imaging patterns constituting established and irreversible disease. Hence sensitive and specific descriptors of early are not well known due to a lack of corresponding histopathological-scale ground truth with which to base clinical CT descriptors, limiting treatment options to supportive care or, more recently, antifibrotic agents that target patients in the late stages of the disease.[13]

One of the main challenges in identifying early-stage imaging biomarkers is the lack of understanding of the exact pathophysiological mechanisms of both diseases. Some studies suggest that microvascular changes in fibrotic lung tissues may not only be a result but also a

cause of these lung conditions with abnormal anastomoses (connections between the pleural and parenchymal blood circulation), vascular remodelling and capillary dilatation all seen.[8,15,16] Hence, understanding the nature of the vasculature in both healthy individuals and those with established PPFE or IPF may enable the discovery of specific imaging biomarkers to identify those at risk of progressive disease. This information could also then be used to not only to identify patients for possible drugs trials and allow for a method of monitoring their progress, but also as targets for therapies aimed at vascular remodelling and angiogenesis.

Hierarchical Phase-Contrast Tomography

Hierarchical Phase-Contrast Tomography (HiP-CT), a novel three-dimensional imaging method using X-ray propagation technique , offers greater resolution and precision in ex vivo imaging,.[17] HiP-CT advances tissue differentiation by utilizing phase contrast imaging, relying on the phase shift of X-rays passing through different tissues, achieving greater resolution at the microscopic scale compared with attenuation-based X-ray imaging.[18] Its hierarchical approach uses phase-contrast imaging scans at varying scales, providing anatomical structure visualization from organ-level overviews to microscopic details, achieving down to 2.5-micron resolution.[17] Despite its promise, HiP-CT faces challenges, primarily the vast data volumes it generates with a single volume-of-interest through lung depth captured at 6 μm per voxel amassing ~600 GB of data. .[17]

The challenge of sparsely labelled data:

Given the potential ability to assess microvascular changes on HiP-CT, an initial step would be to identify a method to segment the vasculature on HiP-CT for further quantitative insight. A recent review on blood vessel segmentation concluded that factors beyond model choice, such as

varied metric choices, lack of definite ground truth, and insufficient study of pathological vessels, make it difficult to define a 'gold standard algorithm'.[19] Supervised learning techniques requires large volumes of high-quality labelled data to accurately represent the data distribution for optimal model performance.[20] Without this, issues such as overfitting and lack of generalizability can arise. In this study, and generally in medical imaging, obtaining sufficient high-quality labelled data is challenging due to several factors, primarily the high time and monetary costs of expert labelling, resulting in sparsely labelled or small datasets. These datasets present several challenges, including bias due to data imbalance, as sparse labels may lead to certain features being labelled more frequently.[21] For example, in vessel segmentation, larger and well-defined vessels may be annotated more often than smaller ones, leading to better quality segmentation for larger vessels. Additionally, less skilled or experienced labellers may be used, resulting in decreased annotation quality and less reliable data.[22] Finally, the same feature might be labelled inconsistently by the same labeller (intra-observer variability) or different labellers (inter-observer variability), hindering the consistency of labels.[20] While supervised learning with labelled data remains the 'gold standard,' semi-supervised learning (SSL) aims to tackle label scarcity by leveraging unlabelled data.

Semi-Supervised Learning:

Semi-supervised learning harnesses both supervised and unsupervised learning by using both labelled and unlabelled data to make predictions.[23] It is particularly useful for small datasets or sparse labels as it leverages large amounts of unlabelled data, reducing the need for extensive labelled data.[24] In this study, we use semi-supervised methods based on consistency regularization, derived from entropy minimization to reduce prediction uncertainty as a strong

regularization technique on unlabelled data to find a decision boundary.[26,27] Consistency regularisation can be thought of as having 2 different types, soft and hard.[27] Soft regularization applies a distance-based loss function to predicted probabilities, allowing for similar but not identical predictions. Hard regularization uses pseudo-labels with strict boundaries to train model predictions. Of note both methods can be used at the input level and feature level.

The soft concept of consistency regularisation was first proposed by Bachman, and subsequently popularised through the introduction of the Pseudo-Ensemble Agreement regularisation, a term that is used to minimise the difference between the output of an original datapoint (so called parent) and its perturbed versions (so called children).[28-30] In essence, when a parent data point is perturbed, it creates several children datapoints. Applying the regularization term to these children datapoints ensures they align on the same lower-dimensional manifold or 'surface' within a higher-dimensional space, thus producing similar outputs. This approach effectively utilizes unlabelled data, which, although not explicitly labelled, now carries useful information that can be leveraged. Consequently, it helps a model produce consistent outputs when given the same input subjected to different semantic-preserving perturbations. The sensitivity to perturbations causing differences in predictions on the same input is penalized by a regularization term, typically based on mean square error or K-L divergence.

Another methodology to leverage unlabelled data in the chosen models is the use of pseudo-labels, a long standing concept popularized by Lee[33]. The idea is for the network to initially train on the available labelled data, which is then used to make predictions on the unlabelled data. If a certain confidence threshold is met, the predicted label is treated as a ‘pseudo-label’ for the unlabelled data, which is then incorporated into the labelled dataset for subsequent training iterations. The threshold is crucial to ensure that the predicted pseudo-labels are accurate, as one

significant limitation of this technique is the potential propagation of errors if the predicted labels are incorrect.[34] A subsequent state of the art technique, FixMatch, proposed a streamlined approach that combines both consistency regularization at the input level and pseudo-labelling.[35] In this method, a weakly augmented image is first fed into the model, and pseudo-labels are generated based on the model's predictions. If the model produces a high-confidence prediction above a certain threshold for a given image, that pseudo-label is retained. Then, when the model is presented with a strongly augmented version of the same image, it is trained to predict the pseudo-label using cross-entropy loss. These models form the foundation of the current ' models being used in the project.

This study employs two state-of-the-art semi-supervised learning algorithms. The first, MisMatch uses morphological perturbations at the feature level with consistency regularization, learning optimal perturbations from data via attention mechanisms.[36] The second, SegPL, is a purely pseudo-label-based model set as an expectation maximization problem, offering robust performance against noise and adversarial attacks with less computational cost.[27,37]

Both models outperform existing state-of-the-art models, including FixMatch, when applied to pulmonary vessel segmentation, although this is on traditional CT imaging.

Objectives:

To achieve this the project aims to evaluate the chosen SSL methods for segmenting vasculature in HiP-CT datasets. The goal is to identify histopathology correlating imaging biomarkers of early fibrosis that could be amenable to potential pharmaceutical intervention. Additionally, the performance of these semi-supervised models will be compared against the state-of-the-art

supervised learning technique nnU-Net to assess improvements in handling label-scarce environments. This project is the initial step into understanding challenges in producing optimal segmentations of the vasculature in novel HiP-CT data to ultimately discover new biomarkers in early fibrosis.

Research Question

How effective are the latest machine learning techniques (e.g., semi-supervised learning and nnU-Net) in performing vessel segmentation on novel Hierarchical Phase-Contrast Tomography (HiP-CT) images with sparse annotations to delineate vascular anatomy in cases of early pulmonary fibrosis?

Methodology

Dataset selection and curation:

The data for this study comprised of HiP-CT images of the lungs from the ESRF-EBS, with lung tissues scanned at 25-micron voxel resolution, with regions of interest further zoomed to 6- and 1-2.5-micron resolution—over 100 times the resolution of current clinical CT. The datasets used in this study comprised ‘sub-stacks’ of 2.5-micron resolution from lung biopsy samples, representing a small portion of the total imaged volume.

The original sub-stacks came from two different sources: an area with PPFE (Figure 1A), an area with PPFE in a patient with IPF (Figure 1B). The PPFE/IPF sample consisted of 960 slices with dimensions of 1823 x 1823 pixels and 1 mm spacing between slices (Dataset C). The PPFE

sample delivered two sub-stacks: one with 1838 x 1838 pixels, 1863 slices, and 1 mm spacing (Dataset A), and another with 1838 x 1838 pixels, 220 slices, and 1 mm spacing (Dataset B), which only became available later in the study. The largest sub-stack was over 6GB when stored in NIfTI format and zipped.

Annotations were manually performed to label the vessels, including both the vessel wall and the lumen. The labelling process was a group consensus effort involving consultation with a Consultant Thoracic Radiologist and arbitration by a Pathologist for uncertain cases. Non-experts performed the actual labelling individually for each of the three available datasets. Given the large volume of data, a recurrent cadence was used, with a growing algorithm in 3D-Slicer to interpolate the vessel between slices. Vessels were identified by drawing around the vessel and filling the outlined region (Figure 1C).

An additional unlabelled dataset from the PPFE-only source was used for training for the semi-supervised models, consisting of 1300 slices of 1823 x 1823 pixels (Dataset D).

Models:

MisMatch[36]:

MisMatch is a method that improves semi-supervised segmentation by perturbing morphological features of unlabelled images with consistency regularisation. The model leverages different attention mechanisms to respectively dilate and erode foreground features which are combined in a consistency driven framework. Any encoder-decoder architecture can accommodate the MisMatch framework.

MisMatch is a semi-supervised segmentation method designed to leverage unlabelled data through morphological perturbations of feature maps. The approach combines dilation and erosion operations with consistency regularization, effectively manipulating the effective receptive field (ERF) of the network's predictions to enhance segmentation performance. The MisMatch architecture consists of an encoder-decoder framework with two parallel decoder branches that apply distinct attention-shifting mechanisms (Figure 2).

At the heart of MisMatch is the concept of the ERF, which measures the region of an image contributing most significantly to the prediction of a central pixel. By controlling the ERF, the framework enhances the model’s ability to differentiate between foreground and background features. Larger ERFs, achieved through dilated convolutions, allow for high-confidence predictions over broader regions, simulating dilation. Conversely, smaller ERFs, facilitated by skip connections, restrict the model's focus to a narrower context, mimicking erosion.

The architecture comprises a single encoder, f_e , which extracts high-dimensional feature maps from the input image. These feature maps are then fed into two parallel decoders:

Positive Attention Shifting Block (PASB): This decoder focuses on expanding the ERF using dilated convolutions with a dilation rate of 5. The outputs from the main branch and the dilated side branch are combined using element-wise multiplication to enhance foreground predictions.

Negative Attention Shifting Block (NASB): This decoder reduces the ERF through skip connections, which ensemble shorter effective paths. The outputs from its main and side branches are also combined element-wise to simulate erosion of the feature map.

The outputs of the PASB and NASB are subsequently averaged to produce the final segmentation prediction. This design ensures a balanced perspective, capturing both dilated and eroded features.

The training process for MisMatch employs distinct loss functions depending on the data type:

For labeled data, supervised loss is computed using the Dice coefficient:

$$L_{\text{supervised}} = \text{DiceLoss}\left(f_{d1}(x), \text{GroundTruth}\right)$$

For unlabeled data, consistency regularization loss is applied between the outputs of the two decoders:

$$L_{\text{consistency}} = \text{MSE}\left(f_{d1}(x) - f_{d2}(x)\right)$$

The total loss function combines these two components, weighted by a hyperparameter α\alphaα to balance supervised and unsupervised learning:

$$L_{\text{total}} = \alpha L_{\text{consistency}} + L_{\text{supervised}}\}$$

The diagram provides a visual representation of this workflow, showing the parallel paths from the encoder to the PASB and NASB, the merging of their outputs, and the delineation of supervised and unsupervised loss calculations. As depicted, labeled data follows a path to supervised Dice loss computation, while unlabeled data proceeds to the consistency regularization stage, reflecting the dual objectives of the framework.

By integrating these components, MisMatch effectively captures the complementary benefits of dilation and erosion, ensuring robust segmentation even in scenarios with limited labeled data. The framework's flexibility allows it to be applied to various encoder-decoder architectures, making it a versatile choice for semi-supervised segmentation tasks.

SegPL (Bayesian Pseudo Labels)[37]:

Bayesian Pseudo Labels (SegPL) is a semi-supervised learning method particularly effective for small or noisy datasets. The method frames pseudo-labelling as an Expectation-Maximization (EM) algorithm, iteratively refining pseudo-labels and model parameters to improve segmentation accuracy (Figure 3).

E-Step:

In the E-step, pseudo-labels ($y_u^{'}$) for unlabelled data (x_U) are generated by estimating their posterior probabilities using the model's predictions (θ):

$$y_u^{'} = 1\left(\theta\left(x_u\right) > T\right)$$

Here, T represents the threshold for determining pseudo-labels. In standard BPL, T is fixed (commonly 0.5). However, BPL can also employ variational inference to dynamically learn T, allowing for more adaptive thresholding in noisy datasets.

M-Step:

In the M-step, the pseudo-labels ($y_u^{'}$) generated in the E-step are used to update the model parameters (θ) by optimizing a combined loss function over both labeled (X_L) and unlabeled (X_U) data:

$$L_{\text{total}} = \alpha\, L_U + L_L$$

where:

- $L_U = \text{DiceLoss}\left(\theta\left(x_u\right), y_u^{'}\right)$ is the unsupervised loss on pseudo-labeled data.

- $L_L = \text{DiceLoss}\left(\theta\left(x_l\right), y_l\right)$ is the supervised loss on labeled data.
- α is a hyperparameter controlling the weight of the unsupervised loss.

nnU-Net

nnU-Net (no new U-Net) builds on the original U-Net architecture by automating the configuration process, including preprocessing, network architecture, training, and post-processing.[38,39] This automation simplifies the setup, making it easier to train and deploy U-Net in new environments.

For this study, nnU-Net version 2 was used, which requires minimal metadata for training, thereby streamlining the setup process. Key metadata fields include the type of imaging modality, labels, and the number of training samples. Virtual environments were set up according to the model requirements before using the nnU-Net models.

Study Design

Several factors were considered when designing this study for the 2 semi-supervised methods: the small size and uniqueness of the datasets (two initial datasets), the large size of individual datasets (several GBs when zipped), very sparse labels, lack of a well-labelled region for validation/testing, and limited computing resources. Initially, only the two larger datasets (A and C) and an unlabelled dataset (Dataset D) were available; the smaller dataset (B) became available later.

The study design was informed by lessons from Oliver et al. on evaluating semi-supervised learning algorithms.[40] It involved comparing the semi-supervised model architecture with a fully trained supervised version, as well as a transfer learning model, though this was not possible. Specifically, MisMatch, based on a U-Net architecture, was compared with nnU-Net, which achieves state-of-the-art performance across different segmentation tasks[39] (Isensee et al., 2021). Additionally, the study varied the ratio of labelled to unlabelled data, as algorithms can be sensitive to this ratio. It is also important to report if the unlabelled data for training comes from outside the distribution used for training, as this can worsen performance. In this study, the unlabelled dataset came from the same distribution as one of the two training samples (IPF and PPFE).

Two study designs were devised to achieve the clinical utility of propagating labels for downstream analysis while working within these constraints:

Design 1: Use all initial available data for training (Dataset A + C) while performing qualitative assessment on the unlabelled dataset. This approach aimed to provide sufficient data for meaningful insights and assess the generalizability of the model when trained on sparse labels, particularly in the PPFE/IPF labelled samples. Once dataset B became available, it was used to test and compare the performance of an optimised supervised method. The downside was the lack of a validation set for testing overfitting or hyperparameter tuning, which was considered less critical given the sparse labelling across all datasets.

Design 2: This approach aimed to optimize performance on a single distribution (PPFE) by focusing on the dataset with the highest label density. Dataset B was used for both validation and testing due to its similar distribution to Dataset A and the limited availability of labelled data. While this dual use of Dataset B introduces potential bias in reported performance metrics, it

aligns with the study's objective to develop a model that performs well on a consistent data distribution. This focus on overfitting to a single distribution was deliberate to facilitate a **human-in-the-loop framework**, where human reviewers could iteratively correct predictions and retrain the model efficiently. Dataset C, containing both PPFE and IPF, was excluded from this design to avoid confounding effects from a mixed distribution and sparse annotations, which could reduce the model's ability to learn effectively. The limitations of this design, including potential overestimation of performance on Dataset B, are acknowledged, and qualitative assessments were prioritized over quantitative metrics.

These designs aimed to balance the need for data sufficiency, model generalizability, and the constraints of label sparsity and computational resources. Of note for the supervised method for comparison, as nnU-Net does its own preprocessing, Datasets A and C were given to train and Dataset B was used for test. All 3 study designs are summarised in Table 1.

Data preprocessing.

Semi-Supervised Models:

The two larger datasets initially available consisted of 2D images, while the semi-supervised models required 3D volumes. To convert the tiff files into 3D volumes, images and labels were loaded into lists and then converted into NumPy arrays, significantly reducing processing time. Images were cropped to remove non-tissue areas, and 3D patch sizes of 128, 256 and 512 pixels were used to manage memory constraints. Overlapping patches were employed to maintain spatial relationships and ensure all data was utilized.

Unlabelled data underwent equivalent preprocessing, including normalization, standardization, and conversion to NIfTI files. Data augmentation was performed on the fly, with MisMatch using random cropping and SegPL using random contrast, zoom, orthogonal slicing, and cropping.

nnU-Net:

For nnU-Net, .tiff files were collated into 3D volumes in the NIfTI format, named, and filed according to nnU-Net requirements. The data was processed using the plan and preprocess function for preprocessing, fingerprint extraction, and experiment planning.

Models are summarised in Table 2.

Hyper-Parameter Optimization

Hyperparameter tuning was limited due to scarce validation data. For MisMatch (appendix 3), alpha (the weight for the unsupervised loss function) and batch size were optimised for regularization and training stability. For SegPL (appendix 4), the ratio of unlabelled to labelled data was optimized, as it is a critical parameter.[40]

Data Analysis:

Analysing the results presents several challenges. Firstly, without fully annotated samples, a reliable test set for quantitative analysis is unavailable. Predictions might be incorrectly classified as false positives due to the absence of ground truth labels for certain vessels.

Secondly, the scarcity of labelled data means that using any for testing would further reduce the training set, potentially degrading model performance.

The ideal testing method would involve reconstructing the entire volume to assess structural connectivity. However, given the limited data, only one volume at most could be used for testing, making statistical comparisons impractical due to insufficient observations. Comparing individual chunks before reconstruction might not be meaningful, as sparse labelling in some chunks would distort the metrics.

Filtered test set images ensuring a minimum percentage of labelled data per chunk could facilitate quantitative comparisons. However, this would alter the original image and may not accurately reflect model performance. Such filtering could lead to overestimation of performance, particularly in terms of positive predictive value, by excluding complex structures that mimic vessels. Consequently, while some basic metrics were observed, statistical tests were deemed inappropriate, and a qualitative review was preferred in these initial stages.

Finally, as outlined in Study Design 2, Dataset B served as both validation and test data. This choice was driven by the limited availability of labelled data and the need to focus on a single distribution (PPFE). While this approach emphasizes model performance on a consistent dataset, it also highlights the limitations of quantitative metrics due to sparse annotations and potential overlap in validation and test data usage.

Ethical Considerations:

For the use of novel hierarchical phase contrast tomography in this study, original ethics approval of the data was obtained at Hannover Medical School, Germany for the use of Human

tissue culture as ex vivo models for the analysis of end stage lung disease (ESLD) and lung tumours on 04/02/2022 under the following ethics approval number: 10194_BO_K_2022 (Ethics Review Board Chair Prof. Dr. Bernhard Schmidt). Approval for this retrospective study was obtained from the local research ethics committees and Leeds East Research Ethics Committee: 20/YH/0120.

Results:

Dataset Evaluation

Three datasets and their corresponding label maps were evaluated, consisting of two classes: vessels (label) and non-label (everything else, including biopsy tissue contents and background). The exact number of vessels labelled in each sample is unknown due to variations in vessel quantity and labelling levels. The total volume of labelled data and visual inspection served as surrogate markers, as shown in Table 3. Dataset C was sparsely labelled with only 0.1% of all available voxels being labelled (compared with 0.9% for Dataset A and 1.7% for Dataset B) and on visual inspection these were all small vessels. Dataset B , had the greatest percentage of available voxels labelled, though it was the smallest datatset with only 220 slices and therefore had almost 5 times fewer labelled voxels than Dataset A (12.5 million vs 56.6 million labelled voxels), though these were mostly of larger vessels. Dataset A had the most labelled vessels due to its larger volume, with just under 1% of the volume labelled.

Label Quality and Image Review

Labels were inspected for quality, with images labelled approximately every five slices and interpolated in between. Vessels were defined as everything inside the outer walls, including the lumen. Visual inspection revealed significant variation in label quality, from partial to full inclusion of vessels, complicating the network's task of identifying label class features. The images themselves differed significantly. Normally, a lung image shows a thin pleural layer at the edge, with the lung appearing as a black background with a web-like overlay of terminal acini and airways interspersed with vessels. However, the PPFE sample showed collapsed and` fibrosed tissue with little 'black' lung, while the PPFE and IPF sample showed some airways. Pixel intensities of vessel walls were similar to the background, though often with a visual boundary. Histograms of image intensities for labelled and non-labelled areas showed significant overlap (Appendices 1 and 2). In the PPFE datasets, two overlapping histograms were observed, with peaks representing vessel walls and lumens. The PPFE and IPF dataset showed only one clear peak, likely due to obscured secondary peaks from the background pixel distribution. Dividing main volumes into smaller volumes showed varying degrees of overlap, reflecting differences in tissue, vessel lumen, vessel wall, and background.

Training Challenges and Improvements

Initially, the MisMatch algorithm was trained on the 3D 512-pixel patches. Larger patches typically yield better performance by reducing the loss of contextual information. The initial experiment, using baseline hyperparameters and a batch size of 2, produced noisy, non-converging training due to the random 3D 96-pixel crops from the data generator often lacking labels. To address this, images were filtered for labels at different levels (>0, 1, 2.5, and 5% of the image volume) to ensure consistent labelled data. Experiments focused on smaller patch sizes

(128 and 256 pixels), closer to the cropped patch size, improving training speed and model performance, reducing training time to days. Validation sets were used to monitor overfitting, which was not a significant issue given the sparse labelling. The same strategy was applied to the SegPL method. Training time per model was substantial, with GPU and memory requirements as limiting factors. Training curves were monitored, and models were cut when the curve began to flatten to balance performance and time efficiency. Appendices 6-9 detail the training results. SegPL generally achieved more stable and better metrics compared to MisMatch. Smaller patches trained faster, likely because there was less chance of empty or near-empty labels. Increasing labelled samples, batch size, and decreasing alpha (reducing regularization) improved SegPL's performance. Validation curves showed models often began to 'overfit' within the first 10-20,000 iterations, though this wasn't necessarily when the best segmentation maps were produced visually due to incomplete labelling. Larger patch sizes during validation were not possible due to memory constraints. Appendix 5 shows nnU-Net results, which performed well on unpatched data with a DICE score of 0.75 on the full volume. However, performance declined once the data was patched and normalized. Post-processing for nnU-Net, still under development, showed that simple thresholding was ineffective as it removed tiny vessels.

Segmentation Comparison

Example segmentations from the five different strategies (Figures 2-6) showed similar results, with false positives needing removal. nnU-Net segmented larger and medium-sized vessels more effectively, while MisMatch and SegPL also segmented smaller vessels. MisMatch had slightly fewer false positives. Single distribution training with a validation set produced slightly better segmentations for MisMatch (which may be expected as Dataset B was used for validation),

though there was no significant difference between the two design methods when tested with SegPL. This also highlights the limitation of relying solely on metrics in the setting of incompletely labelled data, as they may not accurately reflect segmentation quality.

Discussion

This study has shown promise for semi-supervised learning models for vascular segmentation in HiP-CT datasets, particularly when compared with the states of the art supervised method for smaller out of distribution vessels. However supervised learning provided more consistent segmentations of the majority label phenotype of vessels. Several limitations were noted at the onset, not least the lack of labelled data which was discussed as a motivation for this piece of work in the introduction section. However, unlike on traditional imaging techniques such as CT where vessels typically appear as tubular structures, HiP-CT at the microscopc level exhibits a wider variation in structural appearances. This, combined with the large volume of data in each stack, made potential noise in the annotated data a significant limitation. Limited labels also risked selection bias, in which the predominantly labelled vessels are preferentially segmented, as seen with nnU-Net, and hence for a supervised approach a more representative labelled sample may help as well as highlighting the need for 'human in the loop' validation. Model bias must also be considered, as pseudo-labelling methods often assume balanced class distributions, which was not the case in this sparsely labelled data. Proper model validation and implementation strategies in addressing class imbalance in pseudo-labelling are essential to mitigate these biases and prevent downstream inequalities.[41]

To further improve the study, comparisons with transfer learning methods, as suggested by Oliver et al. could be undertaken.[40] However, finding a model trained on a similar 3D task is challenging. Improvements to preprocessing pipeline could include focusing on training with a single distribution of data to explore the potential for an active learning approach, though no significant advantage was found here. Also favouring a model with a tendency for false positives could be preferable, as it is quicker to remove false positives than to add new labels. Other improvements noted from the results would be to completely remove background areas especially as small patch sizes produced significant artefacts and even increasing the number of non-expert labelled samples, as those have been shown to lead to accurate segmentations.[42] To address over-labelling (false positives) in the pseudo-labelling samples, a histogram-based attention mechanism, could be beneficial, especially since histogram overlap is less pronounced in patches.[43] Finally in post processing, dealing with connectivity remains a challenge. A single threshold is ineffective as some areas are larger than vessels.

Vessel segmentation on sparsely labelled data presents unique challenges, particularly when comparing various SSL methods across different imaging modalities. Most literature focuses on 2D retinal imaging, which complicates direct comparisons with HiP-CT data, not least because the 3D data creates computational and algorithmic challenges. Additionally, retinal imaging benefits from abundant data, even if unlabelled, whereas HiP-CT is an emerging technology with fewer samples from diverse sources. The morphological and structural differences between vascular structures in conventional retinal, cerebral etc imaging which generally resemble tubular structures are very different to those observed in HiP-CT images, which more closely resemble histopathological images but in 3D volumes.

Comparing to the literature on SSL vessel segmentation, two notable studies, Hou, Ding, and Deng (2021) and Lin, Xia et al. (2023) focus on the Mean Teacher Method, which employs a student and teacher network paradigm.[43,44] Both networks start with the same random or pretrained weights. The student network is trained on labelled data, calculating a supervised loss, while unlabelled data is passed through both networks, and a consistency loss is calculated between them. The student's weights are updated based on these losses, while the teacher's weights are updated through an exponential moving average of the student's weights, providing regularisation, and reducing the likelihood of overfitting. Hou integrates adversarial and consistency regularisation within a GAN-based framework, improving generalisation by making the discriminator's task more challenging and forcing the network to improve. Results on retinal datasets show slightly better sensitivity, but slightly reduced specificity compared to other state-of-the-art models. The clinical utility of this increased performance, the computational cost, and generalisability to other tasks remain unexplored.

Lin applies semi-supervised learning via a teacher-student network using Swim-U-Net as the backbone. The teacher network is trained on labelled data to minimise cross-entropy, dice similarity, and boundary loss, producing predictions that serve as pseudo-labels for the student network. Pseudo-labelling often leads to over-segmentation, so they employ 'adaptive histogram attention' to minimise this, focusing the model on vessel areas. Tested in brain vessel images, the network demonstrates lower surface error compared to nnU-Net and Cross Pseudo Supervision, indicating better performance in labelling unlabelled data. However, high memory usage suggests significant computational cost, which may not be feasible for large HiP-CT datasets.

Another study proposed a ‘hierarchical segmentation network’, using a pseudo-label approach to leverage unlabelled data.[45] It initially uses labelled data to train a posterior network, where the

posterior distribution of the image is learnt from its label. This in turn is used to train a prior network, with K-L divergence loss ensuring minimal difference between the prior and posterior networks. This prior network then obtains pseudo-labels on unlabelled data, with a confidence threshold above which a pseudo-label is retained. The new pseudo-labelled data is then used iteratively to train the segmentation network. The performance of the model is evaluated on both 2D retinal images and 3D liver CT images, showing improved accuracy on 3D images compared to other methods, though with reduced sensitivity (61.5% vs. 70.0%) indicating lower effectiveness in detecting vessels, which are the minority class.[46,47] Notably, the sensitivity for 3D vessel segmentation (61.5%) is lower than that for 2D retinal imaging (79%), highlighting the challenges posed by 3D datasets.

These studies indicate various techniques applied to relatively limited data types, highlighting that vessel segmentation in sparsely labelled datasets is still a valid area of research.

Clinical and Public Health Implications

The clinical implications of this work are significant, demonstrating the potential to achieve vessel segmentation in HiP-CT data. This advancement potentially paves the way for the study of disease processes that are limited by existing imaging techniques. Specifically, if vascular imaging biomarkers could be identified using HiP-CT before the onset of fibrosis, it could potentially enable the mapping of these biomarkers onto current clinical CT scans. This would provide surrogate biomarkers for pharmaceutical interventions, targeting disease processes before irreversible damage occurs.

This study offers initial insights into the challenges of vessel segmentation in HiP-CT and presents methodologies to address them. By addressing the current limitations and leveraging advanced semi-supervised learning models, this work paves the way for improved diagnostic tools and early detection methods. These advancements can significantly impact patient outcomes by facilitating early treatment and potentially slowing the progression of chronic lung diseases like IPF and PPFE. This research also contributes to a broader understanding and application of vessel segmentation in HiP-CT, which can extend to other diseases where microvascular changes are critical.

Conclusion

Overall, this study has shown promise in using semi-supervised learning models for vascular segmentation in HiP-CT datasets, particularly when compared with SOTA supervised method for smaller out of distribution vessels, though supervised learning provided more consistent segmentations of the majority label phenotype of vessels (large). However, several major obstacles remain, including the need for improved annotation processes, better model optimization, and effective post-processing strategies, not forgetting the challenges of dealing with vast quantities of data. Addressing these challenges will be crucial for advancing the application of semi-supervised learning in medical imaging.

Conflict of Interest: none declared.

References

1. NHS. Lung Health Checks. Accessed July 15, 2023. https://www.nhs.uk/conditions/lung-health-checks/
2. National Institute for Health and Care Excellence (NICE). Lung Cancer: Diagnosis and Management. NICE Guideline [NG122]. Updated 08 March 2024. Accessed May 30, 2024. https://www.nice.org.uk/guidance/ng122/chapter/Recommendations-for-research#diagnosis-and-staging
3. UK Government. New lung cancer screening roll-out to detect cancer sooner. Accessed July 17, 2023. https://www.gov.uk/government/news/new-lung-cancer-screening-roll-out-to-detect-cancer-sooner
4. Hewitt R, Bartlett E, Ganatra R, et al. Lung cancer screening provides an opportunity for early diagnosis and treatment of interstitial lung disease. *Thorax*. 2022 2022;77doi:10.1136/thorax-2022-219068
5. Mayo Clinic. Pulmonary Fibrosis: Symptoms and Causes. Updated 2018. Accessed June 13, 2023. https://www.mayoclinic.org/diseases-conditions/pulmonary-fibrosis/symptoms-causes/syc-20353690
6. Chua F, Desai SR, Nicholson AG, et al. Pleuroparenchymal Fibroelastosis. A Review of Clinical, Radiological, and Pathological Characteristics. Ann Am *Thorac Soc*. 2019 2019;16(11):1351-1359. doi:10.1513/AnnalsATS.201902-181CME
7. Fujimoto H, Kobayashi T, Azuma A. Idiopathic Pulmonary Fibrosis: Treatment and Prognosis. *Clinical Medicine Insights: Circulatory, Respiratory and Pulmonary Medicine*. 2015 2015;9s1doi:10.4137/CCRPM.S23321
8. May J, Mitchell JA, Jenkins RG. Beyond epithelial damage: vascular and endothelial contributions to idiopathic pulmonary fibrosis. *J Clin Investig*. 2023/09// 2023;133(18)doi:10.1172/JCI172058
9. Gudmundsson E, Zhao A, Mogulkoc N, et al. Delineating associations of progressive pleuroparenchymal fibroelastosis in patients with pulmonary fibrosis. *ERJ Open Res*. 2023 2023;9(2)doi:10.1183/23120541.00637-2022
10. Gudmundsson E, Zhao A, Mogulkoc N, et al. Pleuroparenchymal fibroelastosis in idiopathic pulmonary fibrosis: Survival analysis using visual and computer-based computed tomography assessment. *eClinicalMedicine*. 2021 2021;38:101009. doi:10.1016/j.eclinm.2021.101009
11. Shioya M, Otsuka M, Yamada G, et al. Poorer Prognosis of Idiopathic Pleuroparenchymal Fibroelastosis Compared with Idiopathic Pulmonary Fibrosis in Advanced Stage. *Canadian Respiratory Journal*. 2018 2018;2018:1-7. doi:10.1155/2018/6043053
12. Ishii H, Kinoshita Y, Kushima H, Nagata N, Watanabe K. The similarities and differences between pleuroparenchymal fibroelastosis and idiopathic pulmonary fibrosis. *Chronic Respiratory Disease*. 2019 2019;16:1479973119867945. doi:10.1177/1479973119867945
13. Maher TM, Strek ME. Antifibrotic therapy for idiopathic pulmonary fibrosis: time to treat. *Respiratory Research*. 2019 2019;20(1):205. doi:10.1186/s12931-019-1161-4
14. Sgalla G, Iovene B, Calvello M, Ori M, Varone F, Richeldi L. Idiopathic pulmonary fibrosis: pathogenesis and management. *Respiratory Research*. 2018 2018;19(1):32. doi:10.1186/s12931-018-0730-2
15. Gaikwad AV, Lu W, Dey S, et al. Vascular remodelling in idiopathic pulmonary fibrosis patients and its detrimental effect on lung physiology: potential role of endothelial-to-mesenchymal transition. *ERJ Open Res*. 2022 2022;8(1)doi:10.1183/23120541.00571-2021
16. Barratt S, Millar A. Vascular remodelling in the pathogenesis of idiopathic pulmonary fibrosis. *QJM*. 2014/02// 2014;107(7):515-519. doi:10.1093/qjmed/hcu012
17. Walsh CL, Tafforeau P, Wagner WL, et al. Imaging intact human organs with local resolution of cellular structures using hierarchical phase-contrast tomography. *Nat Methods*. 2021 2021;18(12):1532-1541. doi:10.1038/s41592-021-01317-x

18. Viermetz M, Birnbacher L, Willner M, Achterhold K, Pfeiffer F, Herzen J. High resolution laboratory grating-based X-ray phase-contrast CT. *Scientific Reports*. 2018 2018;8(1):15884. doi:10.1038/s41598-018-33997-5
19. Moccia S, Momi ED, Hadji SE, Mattos LS. Blood vessel segmentation algorithms — Review of methods, datasets and evaluation metrics. *Comput Methods Programs Biomed*. 2018 2018;158:71-91. doi:10.1016/j.cmpb.2018.02.001
20. Willemink MJ, Koszek WA, Hardell C, et al. Preparing Medical Imaging Data for Machine Learning. *Radiology*. 2020 2020;295(1):4-15. doi:10.1148/radiol.2020192224
21. Tommasi T, Patricia N, Caputo B, Tuytelaars T. A Deeper Look at Dataset Bias. *arXiv preprint arXiv:150501257*. 2015;doi:10.48550/arXiv.1505.01257
22. Zhou Y, Anderson M, Sadiq S, et al. Biomedical Data Annotation: An OCT Imaging Case Study. *J Ophthalmol*. 2023;2023:5747010. doi:10.1155/2023/5747010
23. Zhu X, Goldberg AB. *Introduction to Semi-Supervised Learning*. Springer; 2009.
24. Chapelle O, Scholkopf B, Zien A. *Semi-Supervised Learning*. MIT Press; 2010.
25. Jiao R, Zhang Y, Ding L, et al. Learning with limited annotations: a survey on deep semi-supervised learning for medical image segmentation. *Comput. in Biol. and Med*. 2023 2023:107840. doi:10.48550/arXiv.2207.14191
26. Grandvalet Y, Bengio Y. Semi-supervised Learning by Entropy Minimization. presented at: Adv. Neural Inform. Process. Syst.; 01 2004; https://proceedings.neurips.cc/paper_files/paper/2004/file/96f2b50b5d3613adf9c27049b2a888c7-Paper.pdf
27. Xu M-C, Zhou Y, Jin C, et al. Expectation maximization pseudo labelling for segmentation with limited annotations. *arXiv preprint arXiv:230501747*. 2023;doi:10.48550/arXiv.2305.01747
28. Laine S, Aila T. Temporal Ensembling for Semi-Supervised Learning. *Clin Orthop Relat Res*. 2016 2016;abs/1610.02242doi:10.48550/arXiv.1610.02242
29. Sajjadi M, Javanmardi M, Tasdizen T. Regularization With Stochastic Transformations and Perturbations for Deep Semi-Supervised Learning. *Clin Orthop Relat Res*. 2016 2016;abs/1606.04586doi:10.48550/arXiv.1606.04586
30. Bachman P, Alsharif O, Precup D. Learning with pseudo-ensembles. *Adv Neural Inf Process Syst*. 2014;27doi:10.48550/arXiv.1412.4864
31. Tarvainen A, Valpola H. Weight-averaged consistency targets improve semi-supervised deep learning results. *Clin Orthop Relat Res*. 2017 2017;abs/1703.01780doi:10.48550/arXiv.1703.01780
32. Miyato T, Maeda SI, Koyama M, Ishii S. Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning. *IEEE Trans Pattern Anal Mach Intell*. 2019;41(8):1979-1993. doi:10.1109/TPAMI.2018.2858821
33. Lee D-H. Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*. 07 2013;
34. Arazo E, Ortego D, Albert P, O'Connor NE, McGuinness K. Pseudo-labeling and confirmation bias in deep semi-supervised learning. *2020 International joint conference on neural networks (IJCNN)*. 2020:1-8. doi:10.1109/IJCNN48605.2020.9207304.
35. Sohn K, Berthelot D, Carlini N, et al. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Adv Neural Inf Process Syst*. 2020;33:596-608. doi:10.48550/arXiv.2001.07685
36. Xu M-C, Zhou Y, Jin C, et al. MisMatch: Calibrated Segmentation via Consistency on Differential Morphological Feature Perturbations with Limited Labels. *arXiv preprint arXiv:211012179*. 2023 2023;doi:10.48550/arXiv.2110.12179

37. Xu M-C, Zhou Y, Jin C, et al. Bayesian Pseudo Labels: Expectation Maximization for Robust and Efficient Semi-supervised Segmentation. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. 2022:580-590. doi:10.48550/arXiv.2208.04435
38. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Clin Orthop Relat Res*. 2015 2015;abs/1505.04597doi:10.48550/arXiv.1505.04597
39. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. Feb 2021;18(2):203-211. doi:10.1038/s41592-020-01008-z
40. Oliver A, Odena A, Raffel CA, Cubuk ED, Goodfellow I. Realistic evaluation of deep semi-supervised learning algorithms. *Adv Neural Inf Process Syst*. 2018;31doi:10.48550/arXiv.1804.09170
41. Wang R, Jia X, Wang Q, Wu Y, Meng D. Imbalanced semi-supervised learning with bias adaptive classifier. *arXiv preprint arXiv:220713856*. 2022;doi:10.48550/arXiv.2207.13856
42. Heim E, Roß T, Seitel A, et al. Large-scale medical image annotation with crowd-powered algorithms. *J Med Imaging (Bellingham)*. Jul 2018;5(3):034002. doi:10.1117/1.Jmi.5.3.034002
43. Lin F, Xia Y, Ravikumar N, Liu Q, MacRaild M, Frangi AF. Adaptive semi-supervised segmentation of brain vessels with ambiguous labels. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2023:106-116. doi:10.1007/978-3-031-58171-7_11
44. Hou J, Ding X, Deng JD. Semi-supervised semantic segmentation of vessel images using leaking perturbations. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022:2625-2634. doi:10.1109/WACV51458.2022.00183.
45. Li C, Ma W, Sun L, et al. Hierarchical deep network with uncertainty-aware semi-supervised learning for vessel segmentation. *Neural Comput Appl*. 2022:1-14. doi:10.48550/arXiv.2105.14732
46. Merveille O, Talbot H, Najman L, Passat N. Curvilinear Structure Analysis by Ranking the Orientation Responses of Path Operators. *IEEE Trans Pattern Anal Mach Intell*. 2018;40(2):304-317. doi:10.1109/TPAMI.2017.2672972
47. Lebre M-A, Vacavant A, Grand-Brochier M, et al. Automatic segmentation methods for liver and hepatic vessels from CT and MRI volumes, applied to the Couinaud scheme. *Computers in Biology and Medicine*. 2019 2019;110:42-51. doi:10.1016/j.compbiomed.2019.04.014

Figure 1: Example of 2.5-micron resolution Synchrotron data of the lung. (A) Biopsy sample with PPFE; (B) Biopsy sample with PPFE and IPF; (C) Sparsely annotated sample with PPFE and IPF.

Figure 2: Workflow of the MisMatch framework for semi-supervised segmentation. The encoder processes the input image, generating features that are passed to two parallel decoders: the Positive Attention Shifting Decoder (f_{d1}) for dilated feature prediction and the Negative Attention Shifting Decoder (f_{d2}) for eroded feature prediction. For labelled data, supervised loss (Dice Loss) is calculated, while for unlabelled data, consistency regularization (MSE Loss) is applied between the outputs of $f_{d1}$and f_{d2}. The final segmentation prediction is obtained by averaging the outputs of the two decoders.

Figure 3: Workflow of the SegPL algorithm with a fixed threshold for pseudo-label generation. Input data is split into labelled and unlabelled datasets. In the E-step, pseudo-labels (y_u❑′) are generated for unlabelled data using the model's predictions (θ) with a fixed threshold (T= 0.5). In the M-step, the model is refined by optimizing a combined loss function (L_{total}) comprising supervised Dice Loss (L_L) for labeled data and unsupervised Dice Loss (L_U) for pseudo-labelled data. The process iterates until convergence, yielding the final segmentation output.

Figure 4: MisMatch segmentation overlay (red) on incomplete ground truth labels (green) from dataset B using Design method 1 (trained on Dataset A+C); Of note a vessel in the top left-hand corner (black arrow) was not labelled in the ground truth, as well as vessel in the middle (blue arrow). There is incomplete labelling of the ground truth. There is partial labelling of the 2 noted vessels not in the ground truth, with minimal additional false positive labels.

Figure 5: MisMatch segmentation overlay (blue) on incomplete ground truth labels (green) from dataset B using Design Method 2 (trained on Dataset A only using validation on dataset C). Incomplete labelling of larger ground truth vessel. More complete labelling of the 2 noted vessels not in the ground truth compared with the other training strategy, as well as other small vessels, however further additional false positive labels.

Figure 6: SegPL segmentation overlay (yellow) on incomplete ground truth labels (green) from dataset B using Design Method 1. There is labelling of the 2 noted vessels not in the ground truth, as well as other small vessels, but with some additional false positive labels, more than seen in MisMatch.

Figure 7: SegPL segmentation overlay (pink) on incomplete ground truth labels (green) from dataset B using Design Method 2. Prediction from SegPL with validation (pink). There is labelling of the 2 vessels not in the ground truth, as well as other small vessels, but with some additional false positive labels.

Figure 8: nnU-Net segmentation overlay (blue) on ground truth labels (green) from dataset B. Good segmentation of the ground truth label and of the small vessel in the top left-hand corner which was not originally labelled. However, smaller vessels are not labelled.

Table 3: Percentage of labelled volume within the total imaged volume for each of the three labelled datasets

Dataset	Pathology	Slices	Total number of labelled pixels within volume	Total number of pixels (labelled and unlabelled) within volume	Percentage of total volume with labelled pixels
A	PPFE	1863	56,665,263	6,188,038,598	0.9157%
B	PPFE	220	12,491,944	731,132,380	1.7086%
C	PPFE + IPF	960	3,204,402	3,146,478,048	0.1018%
D	PPFE	1300	N/A	N/A	N/A

Table 1: Dataset usage in different training designs.

Dataset	SSL Study Design 1	SSL Study Design 2	Supervised Learning
A	Training	Training	Training
B	Test	Validation/ Test	Test
C	Training	Not Used	Training
D	SSL Training	SSL Training	N/A

Table 2 Summary of semi-supervised and supervised model details

Model	Architecture	Key Technique	Loss Function	Training Features
MisMatch	U-Net with Dual Decoders	Consistency Regularization	Training Dice Loss + MSE	Positive/Negative Feature Attention

		via Feature Perturbation		
SegPL	EM-Based Pseudo-Labelling	Dynamic Thresholding for Pseudo-Labels	Dice Loss	Confidence-Based Label Selection
nnU-Net	Self-Configuring U-Net	Automated Pipeline Optimization	Dice Loss	Auto-Generated Preprocessing & Post-Processing