

Quantitative Histopathological Assessment of Liver Fibrosis



Marta Wojciechowska
Lincoln College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Hilary 2023

Acknowledgements

I would like to thank all those who have supported my work on this dissertation, particularly Prof. Jens Rittscher and my colleagues from the Quantitative Biological Imaging group at the University of Oxford. I am grateful for the many fruitful discussions and the moral support which you have all given me throughout this research journey.

I would also like to thank the clinicians who have helped ground my work in its medical context; Dr. Emma Culver for additional supervision and Professors Kenneth Fleming and Robert Goldin for providing many invaluable insights on liver histology. I also want to thank my colleagues at Perspectum; in particular Professor Sir J. Michael Brady and Dr. Caitlin Langford for mentorship and feedback.

Finally, I am so very thankful for everyone who has cheered me on in this DPhil pursuit: my parents, Dorota and Krzysztof, my siblings, and my loving friends from the Pusey House and St Aldate's postgraduate communities. I would not have done this without you.

SDG

This project was supported by the Oxford-Nottingham Biomedical Imaging Centre for Doctoral Training and funded by the Engineering and Physical Sciences Research Council (EPSRC) and Medical Research Council (MRC), grant number EP/L016052/1.

The project has received additional support from Perspectum Ltd., Oxford, UK under the conditions of the Russell studentship agreement.

The computational aspects of this research were supported by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z and the NIHR Oxford BRC. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Abstract

Metabolic-associated fatty liver disease (MAFLD) affects an estimated 25% of the world's population. The feature of MAFLD most strongly associated with adverse outcomes is liver fibrosis. Severe fibrosis is associated with an increased risk of adverse cardiovascular events, the development of malignancy and end-stage liver disease. Pathology staging systems for assessment of the structural changes resulting from liver fibrosis suffer from poor reproducibility. Digital pathology offers ways of quantifying fibrotic changes through simple imaging markers such as collagen proportionate area (CPA). The predictive accuracy of such markers is largely dependent on the quality of the tissue pre-examination processes (i.e histological staining). In this work, we propose a new way of estimating measurement uncertainty of collagen proportionate area through the employment of deep-learning ensembles in a Bayesian inference framework. Next, we design a novel graph-neural network pipeline for automatic staging of liver fibrosis into an established pathological system. Finally, we conduct morphological subtyping of fibrosis patterns in an unsupervised way to develop a novel representation of liver fibrosis at the level of local cellular neighbourhoods. We show how this representation can be integrated with the current imaging markers, such as CPA, to form a succinct visual summary of the current disease state. Finally, we explore the possibility of applying the developed representation to tracking the progression and regression of disease over time.

Contents

List of Figures	viii
1 Introduction	1
1.1 Contributions	2
1.2 Publications	4
2 Background	5
2.1 Metabolic-Associated Liver Disease: Clinical Burden and the Need for Precise Diagnostic Tools	5
2.1.1 Clinical implications in the global context	7
2.1.2 Clinical trials of anti-fibrosis drugs	7
2.2 Liver Function and Histology	8
2.2.1 Liver Under the Microscope	10
2.2.2 Development of fibrosis	14
2.2.3 Pathology staging systems	16
2.2.4 Complementary diagnostic tools	18
2.3 Digital pathology and the liver	18
2.3.1 Quantifying features of MAFLD	19
2.4 Proposal	22
3 Clinical Cohorts and Data Acquisition	25
3.1 Clinical cohorts	26
3.1.1 CALM	27
3.1.2 FLINC	27
3.1.3 HepaT1ca	28
3.1.4 PREV	29
3.1.5 UK-AIH	29
3.2 Clinical data overview	30
3.2.1 Disease history	30
3.3 Acquisition of digital histology slides	32
3.3.1 Sampling variability and specimen size	37
3.3.2 Histological staining for liver fibrosis	39

3.3.3	Sirius red staining in thesis cohorts	40
3.3.4	Slide digitisation	42
3.4	Colour characterisation of digital slides	45
3.5	Conclusions	47
4	A New Digital Lens for Quantifying Fibrosis in Histology Slides	51
4.1	The Drivers of Uncertainty in CPA Measurement	52
4.1.1	In-situ localisation of collagen	53
4.1.2	Digital sampling	53
4.1.3	Segmentation errors	54
4.2	Data	55
4.3	Segmentation Models	56
4.4	Uncertainty Estimation	58
4.5	Results	61
4.5.1	Uncertainty computation	62
4.5.2	Interpreting uncertainty	66
4.5.3	Processing time	68
4.6	Discussion	68
4.6.1	Important nuances in reporting of CPA values	72
4.7	Conclusions	73
5	Automatic Staging of Liver Fibrosis in MAFLD	76
5.1	Previous work	77
5.2	Data	81
5.3	Methodology	81
5.3.1	Graph construction	83
5.3.2	Graph convolutional layers	84
5.3.3	Attention layer	84
5.4	Results	85
5.5	Conclusions	87
6	Beyond Staging: Localised Analysis of Disease Advancement	90
6.1	Characterisation of fibrosis sub-patterns	91
6.2	Methodology	93
6.2.1	Collagen tile subtyping	93
6.3	Results	98
6.3.1	CPA	101
6.3.2	Weighted-fibrosis score and tissue graph parameters	102
6.4	Visualisation of the learned disease space	105
6.5	Tracking disease progression	110
6.6	Conclusions	112

7 Future work	114
7.1 FLINC cohort - research collaboration with Novo Nordisk	114
Appendices	
A Additional information concerning the clinical cohorts	117
A.1 Example written pathology report from UK-AIH study	117
A.2 Full clinical information overview	119
A.3 Imaging artefacts in the FLINC cohort	123
A.4 PSR staining protocol for the FLINC cohort	125
A.5 Colour comparison between cohorts	128
B Collagen Identification Supplementary Material	129
B.1 Training parameters of the colour transfer pipeline	129
C UK-AIH longitudinal study supplementary figures	131
C.1 Comparison of mean CPA values by study	131
C.2 Tracking change in collagen metrics over time	131
References	134

List of Figures

2.1	A schematic of liver anatomy. Figure modified from [26].	9
2.2	Microscopic view of the human liver. Source: surgical liver resection, HepaT1ca cohort [29].	11
2.3	A schematic of the indivisible functional unit of the liver: a portal tract - central vein pair. Blood flows into the tissue from the portal venule and hepatic arteriole through the sinusoid and is drained via the central vein. Hepatocytes, which surround the sinusoid produce bile, which in turn is drained by the bile ductule in the portal field/tract. Image source: [30].	12
2.4	A schematic of the normal liver structure superimposed over a pathology slide image. 2.4a The smallest indivisible unit of the liver: a portal tract - central vein pair. Blood flows from the portal tract and is drained by the central vein. Bile produced by hepatocytes is drained by a bile duct located within the portal tract. 2.4b The liver lobule (dashed hexagon) is a model of liver tissue typically used by pathologists. The central vein takes its name from being at the centre of the lobule. The annotations have been drawn by myself.	13
3.4	Examples of H&E and PSR digital slides from each of the cohorts studied. Differences in sample size and colour, as well as background illumination colour can be observed when comparing samples from different studies.	34
3.5	The digital slide acquisition workflow. At least three human actors are involved in the digital pathology workflow. The sample is acquired by a physician, processed by a laboratory technician, and finally it is examined by a pathologist. The slide's quality is affected by the quality of processing at each individual step in the pre-examination workflow.	35

3.6 A visual representation of a selection of Sirius red-stained liver slides from the UK-AIH study.	
This image illustrates how the use of different protocols for PSR collagen staining results in dramatic differences in slide appearance within the cohort collected retrospectively from 20 NHS hospitals. The images are not to scale. In each digital slide the background was removed and colours adjusted through histogram equalisation for visual clarity.	41
3.7 The results of stain deconvolution on example digital PSR slides from all cohorts.	
a) CALM: the black colour of some collagen fibres is a result of non-specific binding of an unknown green stain to fibres stained with Sirius red. The slides contain a pale yellow residuum. b) FLINC: Sirius red and Weigert's haematoxylin, with little residuum. Upon close inspection one can see that cell nuclei are non-specifically stained red along with the collagen. c) HepaT1ca: Sirius red and two unknown stains, olive and blue. Upon close inspection one can see that cell nuclei are non-specifically stained red along with the collagen. d) PREV: only Sirius red stain was detected. e) UK-AIH: slide stained with Sirius red and a purple haematoxylin with an empty residual channel.	43
3.8 The results of stain deconvolution on example digital H&E slides from the five study cohorts.	
In each of the cohorts a haematoxylin and an eosin were used to produce H&E slides. The residual channel is empty or nearly empty in all of the slides. Slight differences in stain hue of both haematoxylin and eosin, as well as scan background colour can be seen between samples from different cohorts.	44
3.9 Identification of the primary colour of a digital slide.	
First, an appropriate downsample level is identified in the WSI pyramid. The slide image is then read from the downsampled level of the WSI file. Tissue/background intensity-based threshold is identified with the Otsu thresholding algorithm (applied to the greyscale-converted image). Background pixels are replaced with white values. Finally, the median cut algorithm is applied to thus pre-processed image to identify the primary RGB value in the digital slide.	46
3.10 Primary colour distribution in digital slides of PSR-labelled sections.	
Samples from the individual cohorts form separate colour clusters.	48
3.11 Primary colour distribution in digital slides of H&E-stained sections.	
	49

4.1 Limitations of creating human annotations of collagen fibres in microscopy images.	
The task is highly subjective and characterised by high inter- and intra-observer variability, particularly for thin collagen strands and in tissue slides with poor staining quality. It can be seen that while regions with thick, intensely stained collagen are captured, thinner fibres were missed by the annotator (MW).	55
4.2 Collagen annotation with human in the loop.	
Firstly, a representative region is cropped from the full resolution level of the WSI file. The annotator trains a built-in random trees (RT) pixel classifier to categorise individual pixels within the image as either: "collagen" or "ignore". The results of the RT classification are carefully examined visually. If needed, the human annotations are refined, and the process is repeated until the classifier produces a satisfactory pixel-level classification for the entire image. The annotation is then exported as a binary mask.	57
4.3 Collagen prediction maps output by the tested segmentation U-net models.	
The human annotation used as ground truth for model validation contains noisy false positive pixels around the annotated fibres. Each prediction map shows an output prediction from a single model trained on the pooled cohort. It is difficult to see any differences in the quality of the collagen maps predicted by different model types.	63
4.4 The results of U-net Mini ensemble model prediction with uncertainty estimations for examples of unseen tiles from each of the cohorts.	
Tiles with low contrast of stained collagen and unseen colours are characterised by high ensemble model uncertainty.	64
4.5 An example image prediction tile with high epistemic uncertainty.	
The air bubble visible in the original tile is highlighted in the epistemic uncertainty map, which visualises out of distribution data.	65
4.6 Correlation coefficients and corresponding statistical significance values for the prediction parameters of the model trained on the individual cohorts.	
The values are reported for predictions obtained using the U-net Mini ensemble.	66
4.7 Correlation coefficients and corresponding statistical significance values for the prediction parameters of the model trained on the pooled cohort.	
The values are reported for predictions obtained using the U-net Mini ensemble.	67

4.8 An example of stain transfer results between H&E and PSR stainings (HepaT1ca cohort).	
The scene shows a large bifurcating portal tract in two adjacent tissue slices stained with PSR and H&E, respectively. In the synthetic PSR image, while the shapes and textures of the tissue objects look highly realistic, the colour distribution is very different from that of the real PSR slide. A substantial part of the portal tract is coloured green rather than red, and fine collagen fibres within the parenchyma are not captured. Two pieces of debris (grey circles on the left side of the image) have been replicated from the original H&E slice. The synthetic H&E image looks highly realistic.	71
5.1 Visual characteristics of liver architectural landmarks in early and advanced fibrosis.	
Identified portal tracts are outlined in yellow. Annotations were drawn by MW in consultation with Prof. Kenneth Fleming.	79
5.2 Visual context in PSR tissue tiles obtained at different WSI pyramid levels.	
Tiles from pyramid levels 11-15 corresponding to the area indicated by the arrow are shown. Tile size is fixed at 256 by 256 pixels at each of the levels.	80
5.3 The distribution of training and testing samples across the liver fibrosis stages in the PREV cohort (METAVIR standard).	
Healthy liver (stage F0) shows little collagen (red stain), which is almost only present in portal tract walls. During the initial stages of fibrosis (F1-F2) a network of fine collagen strands appears in the parenchyma. In the higher stages (F3-F4) the pathological fibres form "bridges" connecting portal tracts and central veins, resulting in a cobweb-like macro-pattern. It should be noted that the highest stage (F4 - cirrhosis) is not present in the dataset.	81
5.4 Tile subtyping pipeline.	
4 clusters of collagen tiles have been identified, each representing a different level of collagen content. The classified tiles are coloured according to their predicted cluster. Regions with high collagen content (cluster 3) are later used as central nodes for Voronoi tessellation of tissue tiles. The tile size of 256x256 px was specifically chosen to capture the local pattern of the collagen fibres and at the same time to limit the field of view to only contain one type of collagen pattern at a time.	83
5.5 The slide classification pipeline.	
First, tissue landmarks are identified and are used as centres for a Voronoi tessellation of image tiles. Next, a tissue graph is constructed using features extracted from individual tiles and the tessellation weights. Finally, the graph representing the whole slide serves as an input to the GCN classifier.	84

5.6 Examples of tissue graphs superimposed on the original images.	
Node size and colour reflect attention activation for the predicted slide class. Notice that the edge length in the individual sub-graphs corresponds to the underlying pattern of fibrosis and that the sub-graphs capture biologically meaningful regions, e.g. (5.6a) Stage F0 with no signs of fibrosis. (5.6b) Stage F2 with early fibrotic clusters in the parenchyma. (5.6c) Stage F3 with prominent bridging.	88
5.7 The weighted fibrosis score across fibrosis stages and the distribution of tile subtypes across each METAVIR stage.	
The weighted fibrosis score, which is not informed by tissue topology, does not separate between classes F0 and F1 (Wilcoxon signed-rank test coefficients: $W=-1.798$, $p=0.072$) and classes F1 and F2 ($W=-0.775$, $p=0.438$). However, class F3 can be differentiated from all the other classes: F0 ($W=-4.436$, $p<0.001$), F1 ($W=-4.181$, $p<0.001$), F2 ($W=-3.559$, $p<0.001$), and class F2 can be differentiated from F0 ($W=-2.062$, $p=0.039$).	89
5.8 The number and size of the Voronoi cells across the fibrosis stages.	
It can be seen that the size of a Voronoi cell relative to the area of the biopsy can change by an order of magnitude, therefore allowing to capture biological objects of varying sizes.	89
6.1 The hierarchical structure of liver fibrosis.	
Liver fibrosis can be observed at different levels ranging from individual pixels in microscopic images through local textural patterns, biological regions and whole slides. Advanced fibrosis can be further observed at the scale of the whole organ, using modalities such as CT or MRI.	92
6.2 Examples tiles used for learning the subtypes of local collagen patterns.	
Each of the images shows a region of liver tissue stained with a variant of PSR staining for collagen. The size of each square is 256 by 256 pixels, corresponding to a tissue region of approximately 127 by 127 μm .	94
6.3 Tile clustering pipeline.	95
6.4 Collagen tile cluster properties as a function of the number of clusters.	96
6.5 The distribution of tile collagen content across the tile population.	
The distribution of tile collagen in the cohort follows the pattern of exponential decay. Most of the tiles in the cohort contain very little collagen. The smaller figure is zoomed-in to the tail portion of the whole distribution.	97

6.6 Example collagen tiles from k=8 clusters, ordered by mean collagen %.	99
6.7 Examples of tissue samples with different tile cluster distributions. It can be seen that the computed tile clusters, as well as the weighted cluster score (WFS) provide different information about the image than both CPA and pathologist fibrosis stage. In particular, the tile clusters allow to observe the quality of collagen fibres in different parts of the tissue. The histograms allow to visualise the proportions of tiles from each of the clusters in an image.	100
6.8 CPA by stage of fibrosis in the combined cohort.	103
6.9 Mean CPA by fibrosis stage in each of the individual studied cohorts.	104
6.10 Logistic regression plot fitted to WFS as a function of CPA. The points represent individual biopsies from the CALM, HEpaT1ca, PREV and UK-AIH cohorts.	105
6.11 Mean WFS by stage of fibrosis in the combined study cohort.	
6.11a cohorts staged according to the Ishak system. 6.11b individual study cohorts.	106
6.12 Mean fraction of Voronoi cells per tissue tile by stage of fibrosis in the combined study cohort. 6.12a cohorts staged according to the Ishak system. 6.12b individual study cohorts.	107
6.13 The results of principal component analysis of the analysed disease space. The datapoints are coloured according to: study, Ishak fibrosis stage, METAVIR fibrosis stage and the presence of bridging fibrosis in each of the respective plots.	108
6.14 The results of principal component analysis of the analysed disease space coloured by the numerical variables: CPA, WFS and Voronoi cell mean number and size.	109
6.15 Example cases from the UK-AIH cohort for which improvement in Ishak fibrosis stage was observed.	111
6.16 Progression of the cases depicted in Figure 6.15 along the learned disease space.	112
7.1 Example of a small tissue region from a WSI from the FLINC study. A comparison of the features visible at the previously used magnification - 20x and 40x.	115

A.1 An example written pathology report from the UK-AIH study.	
The content of the report is divided into three major parts:	
1) the clinical history of the patient, 2) observations from the	
macroscopic examination of the tissue and 3) the disease-related	
observations of the same specimen examined under the microscope,	
often accompanied by a diagnosis.	118
A.2 Examples of imaging artefacts in WSIs from the FLINC cohort.	124
A.3 PSR staining protocol used in the FLINC cohort.	
Instructions for manual staining (Danish).	126
A.4 PSR staining protocol used in the FLINC cohort.	
Instructions for staining using the Prisma automatic tissue stainer.	127
A.5 Distribution of the primary colours of slides from all studies in the CIELAB colourspace.	128
C.1 Mean CPA by study.	132
C.2 Time difference between the follow-up and the baseline biopsy in the longitudinal sub-cohort of the UK-AIH study.	
Blue colour indicates that time was recorded in days. Grey colour	
indicates that time was recorded in years. The median time difference	
between time-points was 19 months. The shortest time difference	
recorded was 9 months. The longest was 7 years.	132
C.3 Change in CPA in the longitudinal sub-cohort of the UK-AIH study.	
The participants were each following their individual	
treatments. Empty bars indicate excluded data.	133
C.4 Change in WFS in the longitudinal sub-cohort of the UK-AIH study.	
The participants were each following their individual	
treatments.	133

1

Introduction

This thesis is titled "Quantitative Histopathological Assessment of Liver Fibrosis". As the title suggests, the medical concern which forms the basis of all the presented research is liver fibrosis; that is the presence of an excess of collagen fibres in the tissue of a person's liver. The tools and methods used in each of the chapters of this thesis belong to the broad category of digital pathology.

Digital pathology is a relatively new discipline within engineering science which makes use of contemporary digital imaging technologies, such as whole slide imaging, to the study of histopathology. Histopathology is the study of diseases affecting organs, tissues and cells at the microscopic level and is essential in both biological research and in clinical diagnostics of diseases such as cancer.

In the traditional clinical pathology workflow, the pathologist assesses a tissue sample under a microscope and provides a written report to the leading physician (see example report in Appendix [A.1](#)). Such a report typically contains detailed observations about the quality of the tissue specimen with the use of categorical language (e.g. "mild inflammation" or "advanced bridging fibrosis"). These observations are then used for making the diagnosis and monitoring disease progression and regression.

As the language of such descriptions can be ambiguous, numerous scoring systems have been devised in the pathology community to create objective and

comparable standards for disease description. Pathology scoring systems codify disease features into discrete grades and stages, and therefore allow to perform statistical analyses within patient cohorts. However, the assignment of a categorical score to a tissue sample is still a highly subjective process. Pathological studies of scoring repeatability typically report high variability between different observers and even for the same observer on different days [1].

The digitisation of histopathological specimens offers the possibility of performing objective measurements of disease features in tissue images. This avenue of digital pathology is often called "quantitative pathology" and is essential in modern biological research. Quantitative pathology makes use of contemporary deep learning techniques for detection and delineation of tissue objects and of classical image analysis methods for performing measurements.

The recent advances in artificial intelligence have opened entirely new avenues in digital pathology, namely the possibility of training algorithms to infer the diagnosis or disease stage directly from an image. As deep learning algorithms can perform visual detection and classification tasks with high accuracy, a large body of research has been devoted to designing models which imitate the diagnostic inference performed by the pathologist. Such "AI pathologists" follow the diagnostic categories and scoring systems of traditional pathology and can be used to automate the detection and classification of known disease features.

Finally, deep learning can be employed as a tool for discovering aspects of disease which have not yet been described. Unsupervised learning methods can be leveraged to identify previously unknown patterns of disease development and therefore facilitate research in fields such as drug target discovery. In particular, the tools of digital pathology can be used to identify new disease phenotypes or sub-stages and link them to genetic and molecular information.

1.1 Contributions

The core aim of this thesis is the design of new digital pathology algorithms for assessment of liver fibrosis. The chapters are arranged in the following structure.

Chapter 2 - Background describes the context of the presented research from the perspective of the current knowledge on liver fibrosis and current methods of digital pathology.

Chapter 3 - Clinical Cohorts and Data Acquisition contains a detailed description of the patient cohorts studied in this thesis, along with information on data acquisition and characteristics.

The technical contributions of this thesis are then described in the subsequent research chapters. Each of these chapters concerns a different avenue of research in digital pathology.

Chapter 4 - A New Digital Lens for Quantifying Fibrosis in Histology Slides

stems from the fundamental shortcoming of the current quantitative pathology solutions: their lack of assessment of measurement uncertainty. Here we train a deep learning algorithm for the task of segmenting collagen fibres in whole slide images collected from an array of clinical sites in which different tissue pre-processing protocols were used. Subsequently, we leverage outputs from ensembles of such segmentation models to derive metrics of data-related and model-related uncertainty. Finally, we demonstrate how the obtained uncertainty maps can be used to estimate the measurement uncertainty of an established imaging marker of fibrosis - collagen proportionate area.

Chapter 5 - Automatic Staging of Liver Fibrosis in MAFLD presents a novel

deep learning pipeline for automatic detection of bridging fibrosis from a whole slide image. In this chapter we propose a tissue graph tailored for representation of the liver's unique tissue micro-architecture. The tissue graph is then utilised to perform automatic scoring of tissue slides into an established pathological staging system. In this chapter, we introduce a graph neural network set-up which outperforms the currently published methods.

Chapter 6 - Beyond Staging: Localised Analysis of Disease Advancement

presents a novel framework yielding a fingerprint of the fibrosis pattern in the liver whole slide image. In this chapter we perform clustering of local collagen patterns to learn the morphological subtypes of liver fibrosis at the level of individual cell neighbourhoods. Finally, we utilise the learned morphological subtypes along with image analysis methods developed in the previous chapters to create a novel, comprehensive representation of liver fibrosis at multiple image levels. We show how this combined representation can be used to form a succinct description of the current disease state and visualise the time progression of a patient in the learned disease space.

1.2 Publications

The following papers were published based on the research conducted in this project.

Wojciechowska, M., Malacrino, S., Garcia Martin, N., Fehri, H., Rittscher, J. (2021). *Early Detection of Liver Fibrosis Using Graph Convolutional Networks*. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science(), vol 12908. Springer, Cham.

We plan to submit a clinical publication based on the findings from Chapters 3 and 4. We also plan to submit a further article for publication based on the gathered data and the contents of Chapter 6.

2

Background

Contents

2.1 Metabolic-Associated Liver Disease: Clinical Burden and the Need for Precise Diagnostic Tools	5
2.1.1 Clinical implications in the global context	7
2.1.2 Clinical trials of anti-fibrosis drugs	7
2.2 Liver Function and Histology	8
2.2.1 Liver Under the Microscope	10
2.2.2 Development of fibrosis	14
2.2.3 Pathology staging systems	16
2.2.4 Complementary diagnostic tools	18
2.3 Digital pathology and the liver	18
2.3.1 Quantifying features of MAFLD	19
2.4 Proposal	22

2.1 Metabolic-Associated Liver Disease: Clinical Burden and the Need for Precise Diagnostic Tools

Liver disease is one of the chief contributors to the worldwide burden of disability and premature deaths. It is estimated that **non-alcoholic fatty liver disease (NAFLD)**, affects 25% of the world's population [2]. NAFLD is an umbrella term for a continuum of liver conditions, starting with "fatty liver" and ending with

non-alcoholic steatohepatitis (NASH). The prevalence of NAFLD has been growing exponentially, and in the United States of America alone it is projected to increase by 20% between 2015 and 2030 [3].

A patient is considered to have a fatty liver when their liver contains more than 5% of fat within its tissue [4]. Patients with fatty liver only carry a very small risk of adverse liver-related clinical outcomes. However, fatty liver can progress to NASH, which is characterised by liver inflammation, hepatocyte injury and liver fibrosis [5, 6].

Recently, an alternative name has been recommended by an international consensus group of experts as a more precise and informative term for the disease [7]. The updated name, **metabolic-associated fatty liver disease (MAFLD)** [1] refers to the chief comorbidity associated with fatty liver - the metabolic syndrome.

Metabolic syndrome is the medical term for a combination of conditions, which occur together: hyperglycemia, high blood pressure, and obesity [9]. Sedentary lifestyle and a diet composed of highly processed foods are the chief underlying causes of both MAFLD and the metabolic syndrome.

Liver fibrosis is the feature of MAFLD most strongly associated with adverse outcomes [10]. Firstly, it may progress and cause further liver damage in the form of cirrhosis. Cirrhosis may lead to end-stage liver disease, which necessitates a liver transplant [11]. Secondly, the presence of liver fibrosis has been associated with the development of hepatocellular carcinoma (HCC). Both liver fibrosis and the development of malignancy are a result of the same repeated cycle of cell damage, death and regeneration. The treatment of HCC usually involves a form of surgical intervention and/or radiation- or chemotherapy [12]. MAFLD itself

¹The first definitions of metabolic-associated fatty liver disease in pathology were based on discoveries of fat in tissue samples from patients who suffered from alcoholism. The disease was termed alcoholic liver disease (ALD). Soon afterwards similar changes were noted in patients with no history of alcohol misuse. The term "non-alcoholic" liver disease was designated to distinguish between the purported separate diseases. The current knowledge, however is that alcohol use is one of many factors which can play a role in the development of histological features of MAFLD, and a reference to alcohol should not be included in the name, as is it confusing and can be stigmatising to patients [8].

also carries an increased risk of adverse cardiovascular events, independently of its comorbidities [13].

2.1.1 Clinical implications in the global context

The clinical burden of MAFLD is a direct consequence of its very high prevalence. While high prevalence of obesity used to be associated with the countries of North America and Europe, the highest global MAFLD prevalence rates reported are those from the Middle East and South America (31% and 30% respectively) and the lowest are from Africa at 14% [2]. This makes the accurate diagnosis of metabolic-associated liver disease both a truly urgent and global challenge.

According to the current clinical guidelines, MAFLD can only be definitively diagnosed on the basis of a liver biopsy [14]. At the same time, evidence is mounting of a global shortage of pathologists, both in countries with high [15] and low [16] numbers of clinicians per inhabitant. One of the possible solutions to alleviate the diagnostic burden of MAFLD could be the introduction of AI-based digital pathology solutions in the clinic [17].

2.1.2 Clinical trials of anti-fibrosis drugs

Liver fibrosis has been found to reverse upon discontinuation of chronic hepatic injury (e.g. resolution of viral hepatitis or MAFLD) [18]. This important finding along with the need for curative options for MAFLD has prompted a high number of clinical studies in potential antifibrotic drugs. Major phase II and III drug trials targeting MAFLD and liver fibrosis are currently underway [19].

While the ultimate therapeutic goal for such drugs is the resolution of MAFLD and consequently of fibrosis, improvement in fibrosis stage is on its own a desirable endpoint [10]. Similarly to diagnosis, the current golden standard for assessing improvement in fibrosis is the categorical staging of the liver biopsy by a pathologist.

Fibrosis staging, which is described in detail in Section 2.2.3, is a highly subjective task. As such, it is inherently bound by high inter- and intraobserver variability

[1]. Furthermore, the usage of rigid, discrete stages by design does not allow to capture smaller, gradual changes in the state of the organ.

Recent advances in the field of digital pathology have opened new opportunities for precise patient stratification and disease tracking. In particular, a quantitative estimate of liver collagen in biopsies, **collagen proportionate area (CPA)**, has been shown to independently predict long term outcome in MAFLD patients [20]. The promising results from using CPA quantitation in clinical studies have led to arguments for inclusion of digital image quantitation as endpoints in clinical trials of MAFLD drugs [21].

Despite its successes as a predictor of outcome in single-centre clinical studies, CPA cannot yet be seen as an objective *measurement* of tissue collagen content from an engineer’s perspective [22]. Wide discrepancies in methodology of CPA quantitation between studies [23], as well as technical challenges in identifying collagen fibres in digital images of chemically-stained tissue sections [24] render CPA an experimental tool in need of further refinement and standardisation.

The core aim of this thesis is the development of new algorithms for automatic assessment of liver fibrosis from histology sections. My primary goal in this chapter is to explore the current challenges in clinical assessment of liver fibrosis, as well as in application of digital pathology methods for liver collagen quantification and automated fibrosis staging. Finally, my overarching motivation is to identify areas which can be improved using novel deep-learning-based image analysis tools.

2.2 Liver Function and Histology

The liver is the largest internal organ of the human body. It is located in the right upper quadrant of the abdomen. It is protected by the rib cage and can be to a limited extent palpated in medical examination.

The liver performs a number of essential biological functions, and is a crucial player in systemic metabolism and in the maintenance of homeostasis. The heat generated from the many metabolic processes occurring in the liver make it the prime contributor to the body’s homeostatic temperature maintenance [25].

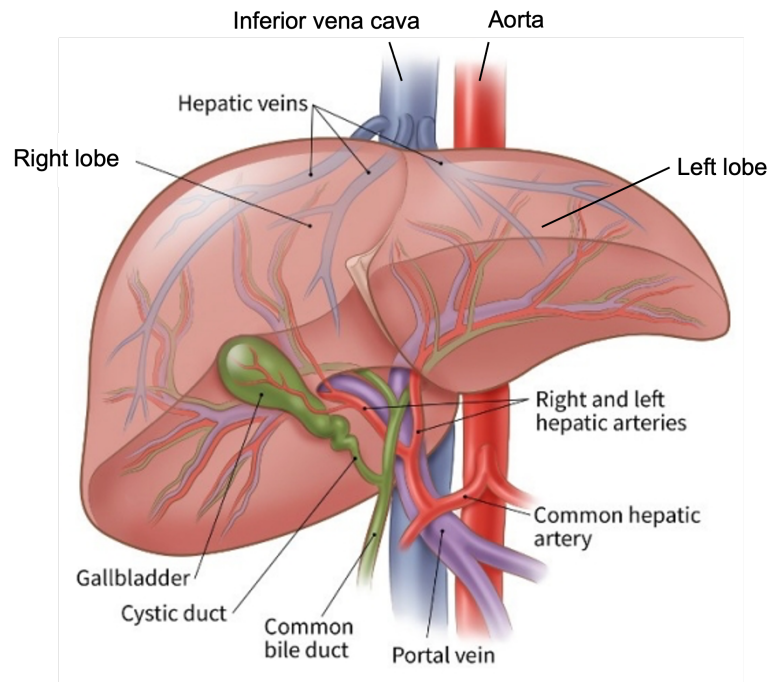


Figure 2.1: A schematic of liver anatomy. Figure modified from [26].

The liver is divided into two asymmetrical lobes: left and right (Fig. 2.1). It is supplied by two separate blood vessels: the common hepatic artery and the **portal vein**. Only approximately 20% of the blood entering the liver is the oxygen-rich blood from the hepatic artery. The remaining 80% is venous blood supplied by the portal vein from the intestines, pancreas and spleen. The liver produces several litres of bile daily. Bile is stored and concentrated in the gallbladder, which is situated directly below the anterior surface of the right lobe. Bile is secreted into the first section of the small intestine (the duodenum) through the common bile duct.

The bulk of the liver is made of **hepatocytes**, which are the chief functional cells of the liver. Hepatocytes perform a number of essential metabolic tasks. These include: production of certain proteins and cholesterol, conversion of excess glucose into glycogen, conversion of ammonia to urea, decomposition of red blood cells and clearing the blood of drugs and toxins. Bile, the alkaline fluid which helps the digestion of fats in the intestine, is also produced by hepatocytes.

The internal anatomy of the liver is defined around the branches of its vasculature [27]. The major vessels: the portal vein, hepatic artery and common bile duct divide

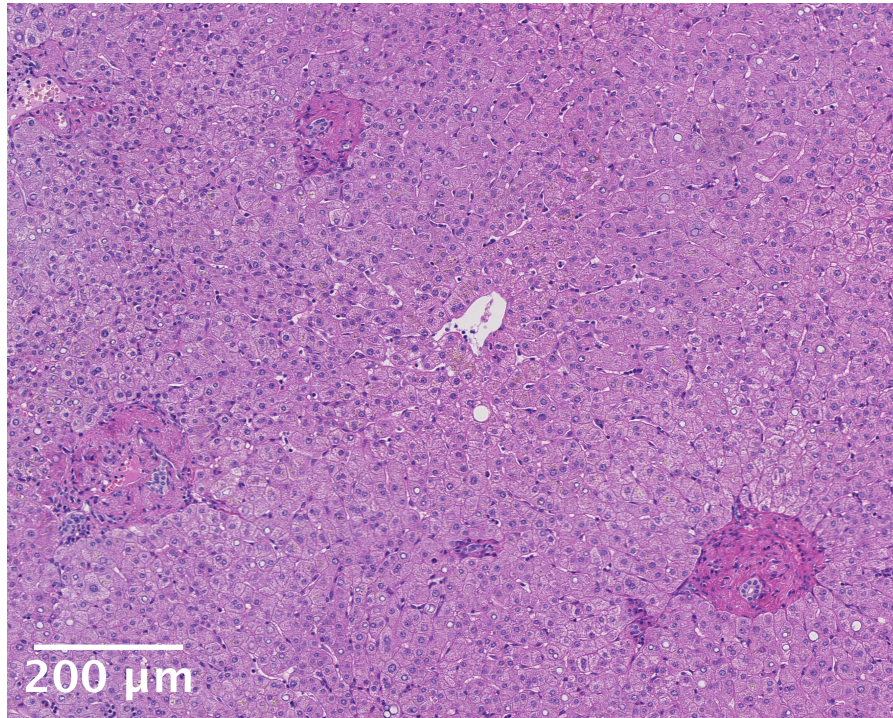
in a structure similar to branches of a tree [28]. This portal tree is simultaneously interwoven with the branches of the hepatic vein - a complimentary "upside-down" vascular tree which removes the blood from the liver and delivers it to the inferior vena cava. The functional units of the liver, the **lobules** are organised around liver vasculature at the microscopic level.

2.2.1 Liver Under the Microscope

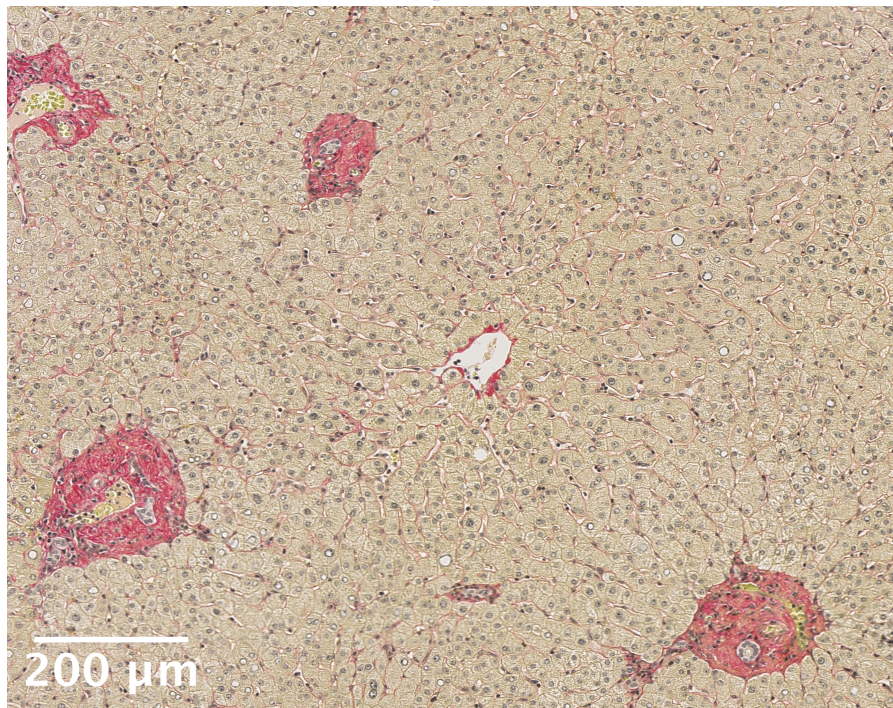
A microscopic view of the liver reveals the internal structure of the organ. In the normal liver, the majority of the visible area is covered by the liver parenchyma, composed mainly of hepatocytes (Fig. 2.2). Hepatocytes are regularly shaped cells with a large, round nucleus. The basic histological staining, haematoxylin and eosin (H&E) allows to distinguish between cell nuclei (haematoxylin - blue) and cytoplasm (eosin - pink). H&E staining is useful for examining basic tissue morphology. However, it does not create a high contrast between liver vasculature and the parenchyma.

Fig. 2.2b shows a slice of the liver stained with the PicroSirius Red (PSR) staining for collagen. In this view, liver vasculature (branches of the portal triad and of the central vein) is visible in red, while the parenchymal region is yellow. Cell nuclei are stained brown. In this view, two kinds of collagen structures can be seen. The white region with a red rim in the centre of the image is the lumen of a branch of the central vein. The solid-looking red regions are in turn cross-sections of the branches of the portal tree.

Fig. 2.3 shows a schematic of the hepatic structure at the cellular level. A portal field (also known as a portal tract or portal triad) contains branches of its three components: a portal vein, a hepatic artery and a bile duct. These three structures are together wrapped in a layer of collagen, which forms the wall of a portal tract. The blood from the portal vein and the hepatic artery enters the parenchyma through a network of sinusoids - narrow channels between hepatocytes. Inflammatory cells, such as the resident macrophages of the liver (Kupffer cells) can traverse the parenchyma through the sinusoid network. In figure 2.2b, sinusoids



(a) Normal liver slice stained with the basic histological staining: **haematoxylin and eosin (H&E)**. Cell nuclei are stained blue. Cell cytoplasm, as well as connective tissue are stained pink.



(b) An adjacent slice of the same region, stained with **PicroSirius Red (PSR)** staining for collagen. Here the red collagen stain highlights the normal portal tracts and central veins of the liver. Liver parenchyma is stained yellow. Cell nuclei are dark brown.

Figure 2.2: Microscopic view of the human liver. Source: surgical liver resection, HepaT1ca cohort [29].

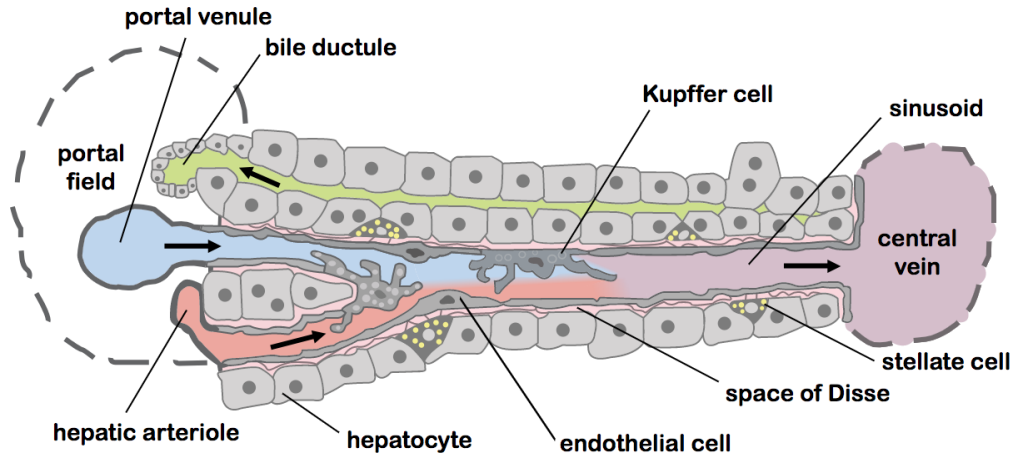


Figure 2.3: A schematic of the indivisible functional unit of the liver: a portal tract - central vein pair. Blood flows into the tissue from the portal venule and hepatic arteriole through the sinusoid and is drained via the central vein. Hepatocytes, which surround the sinusoid produce bile, which in turn is drained by the bile ductule in the portal field/tract. Image source: [30].

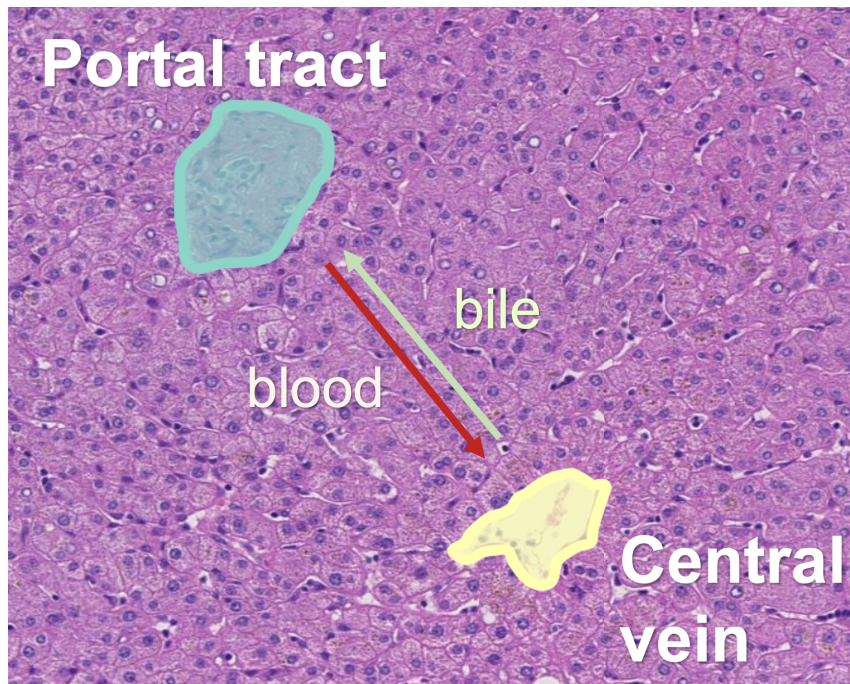
can be seen as thin pink lines within the parenchyma, sometimes with a visible white lumen. At their ends, the sinusoids join into the branches of the central vein, which drains the blood from the organ.

Bile, which is secreted by hepatocytes, flows in the opposite direction from the blood. It is gathered into small bile ductules which are part of the portal tract.

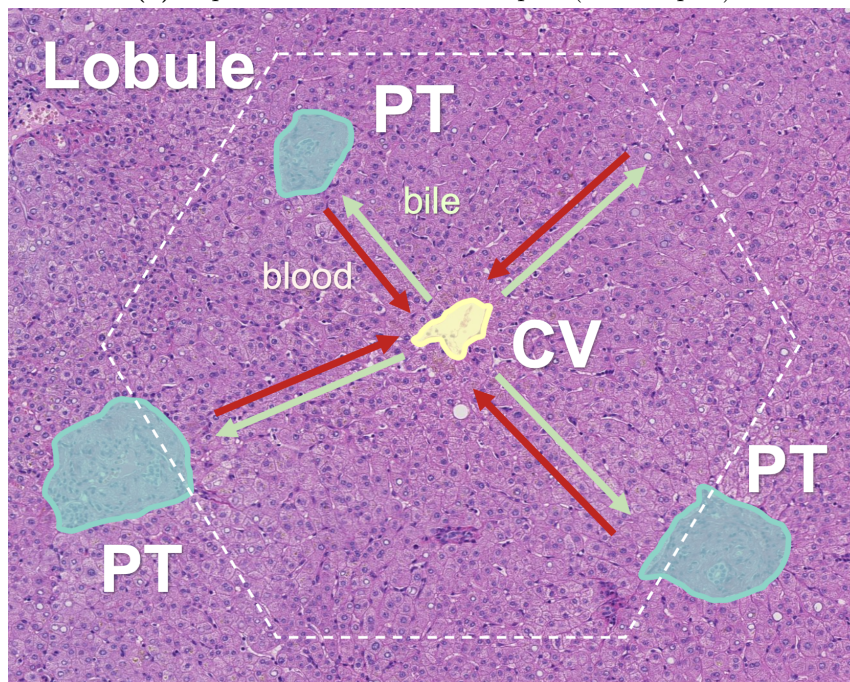
This pairing of a portal tract (PT) and a central vein (CV), along with the hepatocytes between them forms the smallest indivisible functional unit of the liver, the **portal tract - central vein pair (PT-CV pair)**. Hepatocytes specialise in different metabolic tasks depending on their position along the PT-CV axis [31].

A PT-CV pair is a one-dimensional model of the liver tissue, defined at a single perpendicular cross-section of two long tubular vessels. Such small linear regions within the tissue can be identified and isolated, and are useful for observing liver function at the scale of individual cells [32]. The model which is typically used in pathology, the **lobule**, is an extension of the PT-CV pair into a planar cross-section of the liver, where multiple neighbouring portal tracts and central veins can be seen (Fig. 2.4).

A textbook lobule is made of a central vein and six surrounding portal tracts arranged in the corners of a hexagon. Such an idealised model is rarely seen in real



(a) A portal tract - central vein pair (PT-CV pair).



(b) Liver lobule. Annotations drawn over the image from Fig. 2.2a.

Figure 2.4: A schematic of the normal liver structure superimposed over a pathology slide image. 2.4a The smallest indivisible unit of the liver: a portal tract - central vein pair. Blood flows from the portal tract and is drained by the central vein. Bile produced by hepatocytes is drained by a bile duct located within the portal tract. 2.4b The **liver lobule** (dashed hexagon) is a model of liver tissue typically used by pathologists. The central vein takes its name from being at the centre of the lobule. The annotations have been drawn by myself.

tissue sections. Nonetheless, the lobule is a visual concept crucial for understanding liver zonation [33]. As blood flows out of the portal tract region, the concentration of oxygen, glucose and hormones is highest around the lobule's portal tracts. Similarly, the concentration of metabolites produced by the hepatocytes is the highest around the "sink" of the lobule – the central vein. This "watershed" of the tissue is the underlying reason for local anisotropies in the liver parenchyma and consequently in the pathologies affecting it [34].

There also exists an alternative definition of an indivisible functional unit of the liver in a planar cross-section, called the liver **acinus**. A textbook acinus is defined as a region containing a portal tract and six surrounding central veins. The acinar model is used by pathologists alongside the lobular model to describe liver tissue topology.

2.2.2 Development of fibrosis

The development of liver fibrosis is a complex process in which the healthy parenchyma is gradually replaced by scar tissue. The extracellular matrix (ECM) is a complex network of proteins which make up a scaffold for cells forming the tissue of an organ. Various liver cell types can produce ECM proteins, however hepatic stellate cells (HSCs) are the chief players in the process of fibrogenesis [35].

In a healthy liver, the ECM is constantly maintained and fine-tuned by both synthesis and degradation processes. Following liver injuries, HSCs become activated to deposit an increased amount of ECM proteins. Constant liver injury, such as caused by chronic inflammation, can lead to an abnormal perpetuation of fibrogenesis and accumulation of the ECM over time [19].

Table 2.1 illustrates the chronological progression of liver tissue from normal to fibrotic. As hepatocytes within the parenchyma become apoptotic and die as a result of sustained liver injury, the ECM which surrounds them accumulates. Due to the inherent zonation of the liver, this accumulation of ECM fibres is anisotropic and follows a well-described pattern. These accumulated ECM fibres

are mostly composed of type III collagen and can be stained with histological stainings such as PSR.

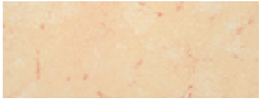
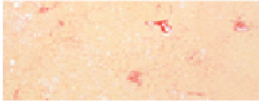
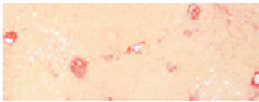
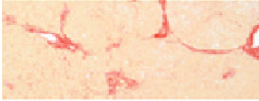
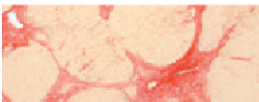
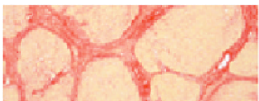
During visual assessment of the liver under the microscope, the first change signalling the development of fibrosis in MAFLD is the fibrous expansion of the portal tracts. Following many cycles of hepatocyte death and proliferation, fibrotic depositions can also be seen within the parenchyma.

The development of collagen **bridges** linking portal tracts with other portal tracts or central veins is a serious alteration in the architecture of the tissue. Collagen bridges interfere with the blood flow within the liver, altering the blood supply to the parenchymal cells [36]. Because of this, the development of bridging fibrosis is considered a tipping point separating early fibrosis from advanced fibrosis.

As the disease progresses and bridges thicken, individual portal tracts and central veins can no longer be delineated as they become incorporated into the fibrotic septa. In the most advanced stage of liver fibrosis, one can observe thick fibrotic septa enclosing nodules of the regenerating parenchyma. This final stage is called **cirrhosis**, from the Greek term *κίρρως*, meaning tawny yellow. The name was first chosen because the surface of a cirrhotic liver, when seen during surgery, is visibly nodular and tan or yellow in colour [37]. A cirrhotic liver is also stiff, which can be felt through palpation.

Liver cirrhosis has a negative impact on the work of the organ and causes a cascade of further negative effects on the rest of the body. As extensive fibrosis within the liver increases its internal vascular resistance, consequently blood pressure increases within the portal vein, a condition known as portal hypertension. Portal hypertension causes vasodilation in the abdominal gastrointestinal organs connected to the portal vein and eventually leads to shunting of portal blood into the systemic circulation [38]. Porto-systemic shunting causes life threatening complications in the form of bleeding gastric or oesophageal varices, ascites (leaking of fluid into the abdominal cavity leading to swelling of the abdomen and bacterial peritoneal infections). Furthermore, the failure of the biosynthetic function of the liver

Table 2.1: Stages of liver fibrosis in the Ishak system with equivalent NAS stages. Images show fragments of liver histology slides stained with the Sirius Red staining for collagen. Table modified from [40].

Ishak	Appearance	Ishak stage: categorical description	NAS
0		No fibrosis. (Normal.)	F0
1		Fibrous expansion of some portal areas. ± Short fibrous septa.	F1
2		Fibrous expansion of most portal areas. ± Short fibrous septa.	F2
3		Fibrous expansion of most portal areas with occasional portal to portal bridging .	F3
4		Fibrous expansion of most portal areas with marked portal to portal + portal to central bridging.	
5		Marked bridging (PT-PT and/or PT-CV) with occasional nodules. (Incomplete cirrhosis.)	F4
6		Cirrhosis , probable or definite.	

caused by cirrhosis can lead to jaundice, hepatic encephalopathy, and hepatorenal syndrome which can result in kidney failure.

While symptoms of liver failure due to cirrhosis decompensation are serious and acute, early (compensated) cirrhosis often goes undetected for extended periods of time [11]. Measures such as serum markers and hepatic vein pressure gradient (HVPG) are commonly used to assess the level of cirrhotic decompensation [39].

2.2.3 Pathology staging systems

In the clinic, the standard of histological assessment has been and still is a detailed pathological description of a tissue specimen. The first semi-quantitative score for liver histological activity, the Histology Activity Index (also referred to as HAI, IASL or Knodell score), was introduced in 1981 to allow a statistical analysis

Table 2.2: A summary of existing systems for liver histology assessment.

System	Year	Disease	Fibrosis	
			Stages	Subscores
HAI [41]	1981	chronic active hepatitis	4	
Scheuer [43]	1991	chronic viral hepatitis	5	
Batts & Ludwig [44]	1995	chronic hepatitis	5	
Ishak [45]	1995	chronic hepatitis	7	
METAVIR [46]	1996	hepatitis C	5	
NAS [4, 47]	1999	non-alcoholic fatty liver disease	5	1A, 1B, 1C
Laennec [48]	2000	hepatic fibrosis	5	4A, 4B, 4C
Beijing [49]	2017	chronic hepatitis B	-	progressive, indeterminate, regressive
SALVE [42]	2021	alcohol-related liver disease	5	4A, 4B, 4C

of serial changes in liver histology [41]. The system was meant to enable group comparisons in patients with asymptomatic chronic active hepatitis. The HAI score consisted of a sub-score for fibrosis, but it also included scores for periportal necrosis, intralobular necrosis and portal inflammation.

The study of different chronic liver diseases has led to revisions in the original system, resulting in an introduction of other semi-quantitative scores for liver histology. The current existing pathological systems for scoring liver histology are summarised in Table 2.2. In each of the scoring systems, stage 0 signifies a liver with no fibrosis and the highest respective stage, cirrhosis. As can be seen, different systems utilise different numbers of stages for fibrosis, some providing additional scores for subtyping early fibrosis (e.g. NAS [4]) or for substaging cirrhosis (e.g. SALVE [42]). Two systems commonly used in clinical trials, Ishak and NAS scoring of fibrosis, are compared in Table 2.1.

Histopathological literature also documents different spatial and textural patterns of fibrosis which can be specific to particular diseases, and are different from fibrosis stages. For example, the **chicken-wire fibrosis** pattern is often found in patients with ALD and/or MAFLD. It is incorporated into the NAS score in the form of subscores 1A (mild) and 1B (moderate) [50]. Similarly, the Laennec and SALVE

systems divide cirrhosis into subtypes, depending on the thickness of the fibrotic septa [42, 49]. Such textural patterns are important in the assessment of current disease activity. The quality of the septum forming fibres (whether they are densely compacted and parallel or loose and sprawling across the parenchyma) is the basis of the Beijing score [49]. The Beijing score divides textural patterns of cirrhosis into predominantly progressive, indeterminate and predominantly regressive.

The presence of multiple staging systems for liver fibrosis testifies to the inherent subjectivity of the task of assigning categorical labels to disease states which exist on a continuum. Similarly, the usage of different systems by different research groups presents a challenge when comparing results across studies.

2.2.4 Complementary diagnostic tools

The biopsy has several important limitations as a diagnostic tool for liver fibrosis. Firstly, it is a painful procedure, which in some cases may be dangerous to the patient [51]. It is first performed for the purpose of diagnosis, but in many cases, follow-up biopsies may be required to monitor the progression or regression of disease. Preparation and storage of tissue sections is also time consuming and expensive. However, the most important limitation of the liver biopsy is that the excised specimen may not be representative of the organ on the whole, due to the very small size of the sample.

A large number of non-invasive diagnostic tests have been developed to estimate the degree of liver fibrosis such as serum markers and imaging-based measures [52]. These include among others: MRI-derived metrics (e.g. cT1 for fibro-inflammation [53], MR-elastography [54]), and ultrasound-based tests (i.e. transient elastography [55]). However, such methods are ultimately only validated against the discontinuous stages of fibrosis as defined by the pathology staging systems [10].

2.3 Digital pathology and the liver

The advent of the whole slide imaging (WSI) technology along with the recent developments in deep learning have led to a large interest in applying computational

image analysis tools to the analysis of tissue slides. Quantitative pathology solutions rely on identification of biological objects within the tissue (e.g. whole cells or cell nuclei) and quantification of their directly interpretable morphological features (e.g. area, eccentricity, density of occurrence). Such tools are already available as both open software and commercial products and are commonly used in biomedical research [56].

Similarly, advanced image analysis pipelines using deep neural networks (DNNs) have been successfully applied to detect disease states directly from medical images [57]. This approach has been adapted by many digital pathology studies in areas such as cancer research [58].

In the case of liver disease, most of the work in digital pathology has been concentrated around quantification of features of MAFLD [59]. The purpose of this section is a review of methodologies applied to tissue feature quantitation, with a special focus on methods targeting the liver.

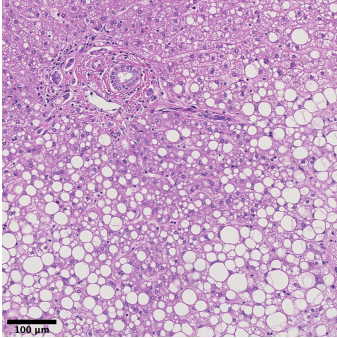
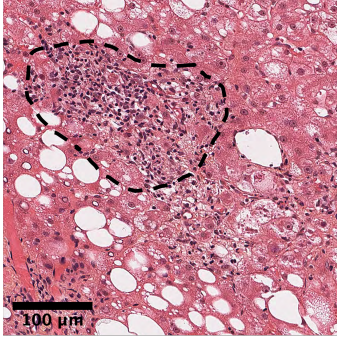
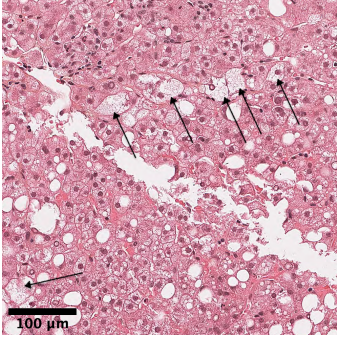
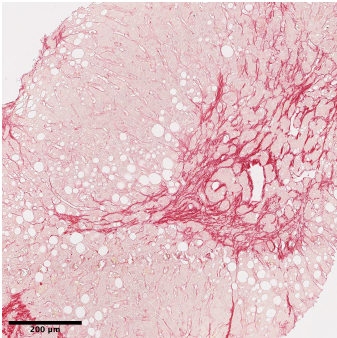
2.3.1 Quantifying features of MAFLD

Quantification of histopathological features of MAFLD is an area of active research with many method papers published in the recent years. Scores for visual features of liver histology have been previously developed by the pathology community for each of the tissue markers of MAFLD (steatosis, inflammation, ballooning and fibrosis) separately [4, 60]. The digital pathology literature draws from that expertise, providing individual methods for quantifying each of the MAFLD features. The current approaches to MAFLD feature quantification are briefly summarised in Table 2.3.

The published methods can be divided into two general categories. Firstly, tissue objects (such as cell nuclei or bubbles of fat) can be segmented and their area percentage can be reported. Secondly, advanced deep-learning methods can be applied to classification of whole slides into the previously defined histology categories.

The benefit of the first approach is that it is fully intuitive and interpretable by the human reader. However, it becomes challenging when it is not clear how an object

Table 2.3: An overview of published methods for quantification of MAFLD features.

Feature	Example Image	Staining	Quantitation Methods
Steatosis White round vacuoles within the parenchyma.		H&E, others	Fat vacuole segmentation, fat area % [61, 70]. Automatic grading [71].
Inflammation Parenchyma infiltrated by inflammatory cells (dashed region).		H&E, CD45	Small nuclei segmentation, area % (H&E) [72], CD45-positive area % [70]. Area % of "dark regions" (H&E) [73]. Automatic grading [72].
Ballooning Enlarged hepatocytes with rarefied cytoplasm (pointing arrows).		H&E	Ballooned cells area % [73]. Automatic grading [72].
Fibrosis Collagen fibres in the parenchyma (here stained red).		PSR, others	Collagen fibre segmentation + area % (CPA) [70, 73]. Manually annotated collagen area % [74]. Automatic staging/subtyping [71, 75, 76]. Independent quantitative scores [77, 78].

or region should be defined. This is for example the case with quantitative scores of inflammation (Table 2.3). While some groups define the area of inflammation as the area of the segmented immune cell nuclei [72], others report the area of regions which have nuclear density higher than a chosen threshold [73]. Similarly, marking of other features such as cell ballooning is highly subjective. A recent study of pathologists' agreement on annotation of ballooning cells has reported a substantial divergence between the scores assigned by individual experts [79]. Such findings suggest that not all image features of MAFLD can be reliably quantified from basic H&E-stained tissue slides, even if they were previously included in the histology scoring systems.

The other group of methods relies on application of deep-learning networks for classification of the whole slide into histology scores. Such methods do not need precisely defined biological features to achieve high correlation with histological scores, but in turn their results are not easily explainable [80]. As categorical classification of liver histology slides is burdened by high inter-observer variability [1], the inherent subjectivity of scoring is propagated onto the models during algorithm training.

Finally, some groups have developed entirely novel quantitative metrics of NAFLD features, which are different from both basic area quantitation and histopathological scores. An example of such a metric is QFibrosis, in which hand-crafted imaging features of liver fibrosis are extracted for different predefined topological subtypes of collagen and combined into a single numerical score [78]. Another example is quantification of fractal dimension of the segmented collagen in the histology slide [77]. Such methods, though boasting fine-grained metrics, still need to be validated either by their ability to predict histological stages or against clinical outcomes. Similarly to deep-learning based methods, models which rely on combining high numbers of handcrafted imaging features or on use of abstract mathematical concepts often aren't directly interpretable to human readers.

Due to current high interest in precision tools for liver histology assessment from the pharmaceutical industry, a number of medical imaging companies offer such tools as commercial products. Such commercially developed software constitutes

protected intellectual property which is rarely disclosed publicly in the form of original method papers. Nonetheless, some studies which make use of commercial image analysis algorithms have recently been published in clinical journals [81, 82]. It is important to note that studies which do not disclose the details of image analysis method implementation cannot be replicated in an academic setting.

2.4 Proposal

The need for highly robust methods of collagen identification in histological slides

Almost all of the method papers published in the subject of digital pathology of liver fibrosis present histology data processed within a single institution. Differences in tissue staining procedures between laboratories and even batch-to-batch staining variations are a known issue which often leads to unreliable results of collagen area quantification [24]. While advanced microscopy methods, such as SHG/TPEF microscopy allow for detection of collagen fibres from unstained tissue sections [83], these are not widely available and are currently unlikely to replace the use of whole-slide imaging of stained tissue sections in the clinic.

Deep-learning based segmentation models like such as the U-net [84] have been a popular solution for delineation of objects in medical image analysis. The advantage of deep-learning based models is that they can be trained to be robust to differences in sample colour [85]. Segmentation models are trained against accurate annotations drawn manually by field experts. However, it is known that CPA measurement using annotations made by pathologists is burdened by high inter-observer variability [24].

In this thesis I present a collagen segmentation model trained on a highly variable dataset composed of whole-slide images from three individual clinical studies. The model is additionally tested against samples from an unseen cohort.

Subsequently, deep learning ensembles are utilised to estimate the measurement uncertainty of the segmented collagen area in the histology slide. The results of applying this method to samples from four different clinical studies are presented. To my knowledge, this is the first attempt at estimating measurement uncertainty

of collagen proportionate area in whole-slide images of liver histology sections. The work from this chapter has been submitted for publication in the proceedings of the MICCAI'23 conference under the title "A New Digital Lens for Quantifying Fibrosis in Histology Slides".

Automatic detection of bridging fibrosis in clinical liver biopsies

The development of bridging fibrosis is an important clinical milestone separating early fibrosis from advanced fibrosis. Automatic classification of liver needle biopsies into fibrosis stages is a challenging task due to the small size and very narrow shape of the samples [86]. Deep-learning-based classification methods published so far benefited from the use of whole liver resections from animal studies [75].

In this thesis we propose a novel graph-based representation for liver tissue and consequently apply it to classification of PSR-stained liver needle biopsies into bridging fibrosis/ no bridging fibrosis. The proposed graph neural network based prediction model achieved higher F1 scores for classification of samples with bridging fibrosis than the previously published methods.

This work has been published and presented at the MICCAI'21 conference under the title "Early Detection of Liver Fibrosis Using Graph Convolutional Networks" [87].

Data-driven characterisation of fibrosis patterns

Collagen area quantitation does not allow to capture topological changes in liver fibrosis, the observation of which is the basis of standard pathological assessment. While semi-quantitative staging systems are widely used in clinical trials to capture major changes in collagen architecture, they do not allow to observe incremental improvements which may indicate that a therapy is producing desirable outcomes. Because of these limitations, there now is a large interest in devising fine-grained metrics to describe the quality of fibrotic patterns in liver biopsies.

This interest is reflected in liver pathology literature with introduction of scores for chicken-wire fibrosis [4], progressive/regressive patterns [49] and substages of

cirrhosis [42, 48]. Similarly, attempts have been made at capturing these textural changes through digital quantification of selected regions of the tissue sections.

A notable example of this is the reporting of collagen proportionate area computed from the whole slide *excluding* manually annotated portal tract and central vein regions [74]. Other studies have been conducted where in turn only collagen area *from* regions identified as portal tracts (with the fibrotic expansions) was quantified [88]. Another example is the quantitation of width of annotator-identified cirrhotic septa [89].

All of the above methods use human input to recognise the different textures of collagen in the slide and select regions of interest. As was previously reported, classification of fibrotic structures is highly subjective, even among expert pathologists [1]. Furthermore, analysing the different regions of collagen in separation from each other selectively removes information from the pathology slide and so introduces bias to the analysis. Finally, while creating manual tissue annotations by expert pathologists is feasible in small studies, it requires additional workload, which is undesirable both in clinical practice and in large clinical trials.

In this thesis I use deep-learning-based methods to learn the different textural patterns of collagen in PSR-stained slides of the liver in an unsupervised fashion. The collagen features extracted on the tile level by a DNN are clustered to learn the basic subtypes of collagen patterns in the liver in all stages of fibrosis, from healthy liver to cirrhosis. Finally, whole slides are analysed to produce heatmaps of distribution of the identified collagen textural subtypes in the slide.

The results of this quantification are applied to samples from the CALM, HepaT1ca, PREV and UK-AIH cohorts. The results of fibrosis pattern analysis are compared against pathologist-assigned fibrosis grades. The potential of applying collagen pattern subtyping methods to disease tracking over time is explored using longitudinal samples from the UK-AIH cohort.

3

Clinical Cohorts and Data Acquisition

Contents

3.1 Clinical cohorts	26
3.1.1 CALM	27
3.1.2 FLINC	27
3.1.3 HepaT1ca	28
3.1.4 PREV	29
3.1.5 UK-AIH	29
3.2 Clinical data overview	30
3.2.1 Disease history	30
3.3 Acquisition of digital histology slides	32
3.3.1 Sampling variability and specimen size	37
3.3.2 Histological staining for liver fibrosis	39
3.3.3 Sirius red staining in thesis cohorts	40
3.3.4 Slide digitisation	42
3.4 Colour characterisation of digital slides	45
3.5 Conclusions	47

This chapter describes the histological and clinical data which were made available to this thesis project. Thorough understanding of the characteristics, inherent limitations, as well as quality of the available histopathological data was necessary for formulation of each of the research questions posed in this thesis.

Here five separate cohorts of participants who have undergone liver biopsy or surgical liver resection are described. Each of the cohorts was originally gathered

to address a specific research question - concerning a particular liver disease or diagnostic method. The cohorts were then made available to me for the purpose of identifying novel ways of examining liver disease using digital pathology methods.

This combined cohort of five different studies offers a unique insight into clinical practices around liver fibrosis assessment from tissue sections. The specimens were acquired in more than 20 healthcare institutions from 3 different countries. Together they represent the real world's diversity of laboratory techniques and protocols in research as well as clinical care.

In this chapter I briefly introduce each cohort, referencing the original studies. I then provide a brief summary of the clinical data available. Finally, I describe the process of producing a digital slide from a liver biopsy, which is important for understanding how exactly the samples from various cohorts and hospitals are comparable to one another.

To my knowledge, this is the first time where such a diverse grouping of digital liver histology slides stained for collagen was gathered and described. The goal of this chapter is not only to introduce digitised liver histology samples as data in this thesis, but also to use this combined cohort to reflect more broadly on the data-related challenges on the way to clinical introduction of quantitative pathology.

3.1 Clinical cohorts

Five separate datasets of liver digital slides are used in this project, henceforth referred to by the names of their respective original studies. Table [3.1](#) contains a brief summary of the cohort characteristics.

In total I have received access to 4211 digital slides from 844 liver needle biopsies and 45 surgical liver resections. The digital slides were stored on the BMRC cluster, save for the FLINC cohort which was stored and processed on a specially set up BMRC private cloud. To organise such numerous data points, I have developed a unified slide database, using Pandas [\[90\]](#). In this structure, each digital slide is referred to by a unique slide ID, matched with a unique case ID. The database was also used to store all other slide information including file meta-data, pathology

Table 3.1: An overview of the cohorts studied within the project.

Cohort	Country of Origin	Years	Study Identifier	Data Owner	Disease Studied	Study Aim	N digitised biopsies	N clinical sites
CALM	UK	2014 - 2015	ISCRTN39463479	Perspectum	-	MRI method validation	153	2
FLINC	Denmark	2017-2026	NCT04340817	Novo Nordisk	MAFLD	Evaluation of biomarkers and potential targets in MAFLD	203	1
HepaT1ca	UK	2017-2019	NCT03213314	Perspectum	liver cancer, MAFLD	MRI method validation	44	1
PREV	USA	2015-2019	NCT03142867	Perspectum	MAFLD	Disease prevalence estimation	279	1
UK-AIH	UK	2014 -	IRAS ID: 144806	UK-AIH	AIH	Retrospective analysis of clinical care	210	20

reports, as well as the quantitative results of my analyses. The data characteristics and results were visualised using the seaborn package [91].

3.1.1 CALM

The CALM study was a prospective study for the validation of multiparametric MRI, LiverMultiScan® in quantifying liver disease [92, 93].

The study invited participants scheduled for a standard-of-care liver biopsy in two large tertiary UK liver centres to investigate known or suspected liver disease. Patients were recruited from Queen Elizabeth Hospital in Birmingham and the Royal Infirmary of Edinburgh between February 2014 and September 2015. Overall, 161 adult patients, who had clinically indicated liver biopsies underwent LiverMultiScan® proton density fat fraction (LMS-PDFF) and corrected T1 (cT1) scans, as well as transient elastography (TE) and Enhanced Liver Fibrosis (ELF) test.

The collagen-stained slides were scored by four independent expert pathologists using modified Ishak score. It is unclear from the available publications how reader consensus was achieved in this study. Additionally, the pathology reports contain scoring of liver fat, inflammation and iron. The cohort was made available by Perspectum Ltd.

3.1.2 FLINC

Fatty Liver Disease in Nordic Countries (FLINC): a Prospective Cohort Study is a study conducted at the Copenhagen University Hospital, Hvidovre, in Denmark, in

collaboration with Novo Nordisk, Denmark. The overarching aim of the study is to systematically evaluate biomarkers and potential targets in MAFLD.

Digitised liver histology slides from 203 patients and healthy volunteers were made available to this project, along with the relevant clinical information. The majority of the collected biopsies were percutaneous liver biopsies, however 32 participants were subjected to transjugular liver biopsies. The histology slides were graded according to the NAS scoring system and the Kleiner-Brunt system for fibrosis.

3.1.3 HepaT1ca

HepaT1ca was an observational cohort study in two tertiary-referral liver surgery centres in the United Kingdom - Royal Infirmary of Edinburgh and Basingstoke and North Hampshire Hospital [29]. Because of the high prevalence of MAFLD worldwide, it is estimated that approximately 1 in 5 liver resection candidates also have MAFLD [94]. The purpose of this study was to quantify liver health in surgical candidates for liver malignancies and to predict post-intervention outcomes following liver resection.

The study participants were recruited from all the patients referred to the study hospitals who were considered for liver resection or transarterial chemoembolization (TACE), between August 2017 and January 2019. In addition to the standard care the patients received during their treatment, they were examined using LiverMultiScan™ before and after surgical interventions to determine the the ability of LiverMultiScan™ to predict risk of post-operative morbidity and mortality.

The study included the collection of surgical resections containing non-cancerous tissue, which was used to assess the health of the parts of the organ unaffected by malignancies. 44 liver resections from the study were digitised and independently scored by two expert pathologists according to the NAS scoring system [4] along with Ishak assessment of fibrosis. It is unclear from the available publications how reader consensus was achieved in this study.

3.1.4 PREV

The cohort hereafter referred to as the “PREV” cohort comes from a prospective study of the prevalence of non-alcoholic fatty liver disease (NAFLD) and steatohepatitis (NASH) in a large middle-aged US cohort [95]. The purpose of the study was to establish the prevalence and severity of NAFLD/NASH in a cohort of asymptomatic middle-aged Americans.

The participants were recruited from a colonoscopy class at the San Antonio Military Medical Center Gastroenterology clinic in San Antonio, Texas, USA and were subsequently screened for MAFLD using several non-invasive imaging modalities. The methods used included MR-based LiverMultiScan® proton density fat fraction (LMS-PDFF) and corrected T1 (cT1) as well as elastography for liver stiffness measurement (LSM) and FibroScan®. Participants with an abnormal reading on either of these measurements were proposed a liver biopsy.

In total 279 liver digital slides from this study were made available for this project through our collaboration with Perspectum. The slides were accompanied by standardised pathology reports based on the consensus of two expert pathologists. The comprehensive reports include grading of liver inflammation, steatosis, and fibrosis according to the NAS standard, as well as additional observations, such as cell ballooning and iron content.

3.1.5 UK-AIH

UK-AIH is a multi-centre, national cohort study, funded by the NIHR Rare Diseases Translational Research Collaboration (RD-TRC) and is coordinated by Newcastle University and Newcastle upon Tyne Hospitals NHS Foundation Trust [96]. The overarching goal of UK-AIH is to improve clinical care across the UK for people with autoimmune hepatitis (AIH). The study began recruitment in 2014 and has included 54 sites.

The histological data from the cohort consists of 210 cases with digitised liver pathology slides from 20 different NHS hospitals. Out of the 210 cases, there are 35 longitudinal cases with at least 2 time points available.

Because of the retrospective nature of the study, no single standardised scoring system was used to grade the samples. Instead, the original written pathology reports were made available along with the digitised slides. An example of a written pathology report from the UK-AIH study is attached in Appendix [A](#).

3.2 Clinical data overview

Below is a summary of the basic characteristics of the cohorts included in this thesis. Figure [3.1](#) presents a comparison of the cohort sizes and the year of acquisition of study samples. In particular, the retrospective UK-AIH cohort contains samples collected over a span of almost 20 years, which is reflected in visible discrepancies in slide quality across the study.

The population characteristics of the cohorts are shown in Figure [3.2](#). The human liver is characterised by sexual dimorphism which is known to influence the progression of MAFLD [\[97\]](#). Liver histology also changes significantly over the course of the organ's development into its adult form. This is reflected in differences in disease phenotype between paediatric and adult patients in MAFLD and other diseases [\[98\]](#). The combined cohort contains 4 samples from patients below the age of 18. Furthermore, race is an important factor in liver disease and there are known racial differences in the presentation of MAFLD [\[99\]](#).

The overall cohort of 885 participants contains 386 men and 380 women. The gender data of the remaining participants is unknown. The mean participant age at the time of biopsy collection is 53. Racial information was only gathered in two of the cohorts: HepaT1ca and PREV, with the majority of the participants being Caucasian/white (Fig. [3.2b](#)). The remaining cohorts were gathered within the UK and Denmark and are assumed to contain predominantly Caucasian/white participants.

3.2.1 Disease history

The combined cohort consists of 595 hospital patients and 290 nominally healthy volunteers. Each of the studies was performed on a population with a different disease distribution.

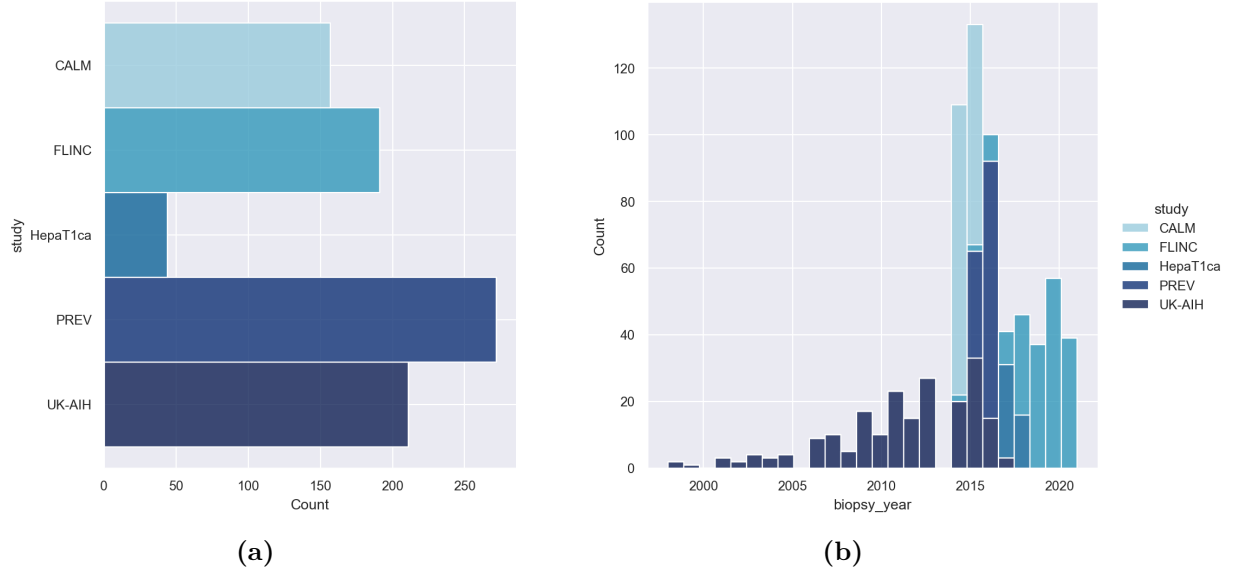


Figure 3.1: Information on acquisition of the study cohorts. a) Comparison of study population sizes. b) Distribution of year of biopsy acquisition by study.

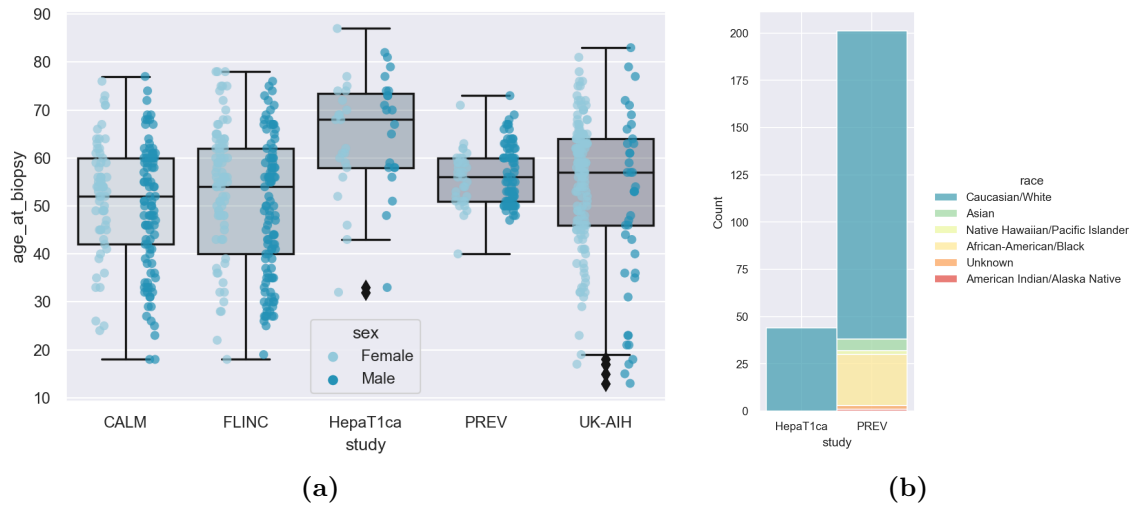


Figure 3.2: Population characteristics of the study cohorts. a) Age and sex distribution by study. b) Racial statistics within the HepaT1ca and PREV cohorts. The CALM, FLINC and UK-AIH studies did not record racial information and are assumed to be containing mostly Caucasian/White participants.

The CALM cohort consists of hospital patients with any liver disease. Over 20 different diagnoses are present in this cohort, including but not limited to: MAFLD, alcoholic liver disease, viral hepatitis and autoimmune liver disease. The FLINC cohort consists of MAFLD patients and 11 healthy controls. Many of the study participants hold additional diagnoses; the most common ones being type 2 diabetes mellitus and systemic hypertension. The HepaT1ca study consists of patients referred to liver surgery because of either a primary or metastasised liver tumour. In this case, their MAFLD diagnosis is secondary to their primary diagnosis with cancer. The PREV cohort consists entirely of volunteers with no known liver disease. The UK-AIH study included individuals recruited from NHS hospitals, who have been diagnosed with autoimmune hepatitis. The available clinical information for each of the cohorts is summarised in Appendix [A](#).

Because of the differences in the available disease information across the studies, it isn't possible to estimate the overall disease distribution in the combined cohort. Instead, here I present solely the distribution of pathologist-assigned fibrosis stages within each study (Figure [3.3](#)).

The CALM, HepaT1ca and UK-AIH studies contain the Ishak staging of liver fibrosis, which consists of 7 stages [\[45\]](#). In contrast, the FLINC and PREV studies were scored according to METAVIR and MAFLD-specific Kleiner-Brunt fibrosis staging systems with 5 stages [\[4, 46\]](#). In each of these systems, stage 3 marks the onset of bridging fibrosis (advanced fibrosis). Overall, there are 575 participants with no bridging fibrosis and 144 participants with bridging fibrosis. Fibrosis staging is unavailable for the remaining participants.

3.3 Acquisition of digital histology slides

The entirety of histology and clinical data referred to in this thesis was collected by the individuals conducting the respective projects. I have not participated in any part of the histology slide acquisition process and have only had access to the digitised versions of the tissue slides. In this section I am describing the process of preparation of digital histology slides for the purpose of explaining the

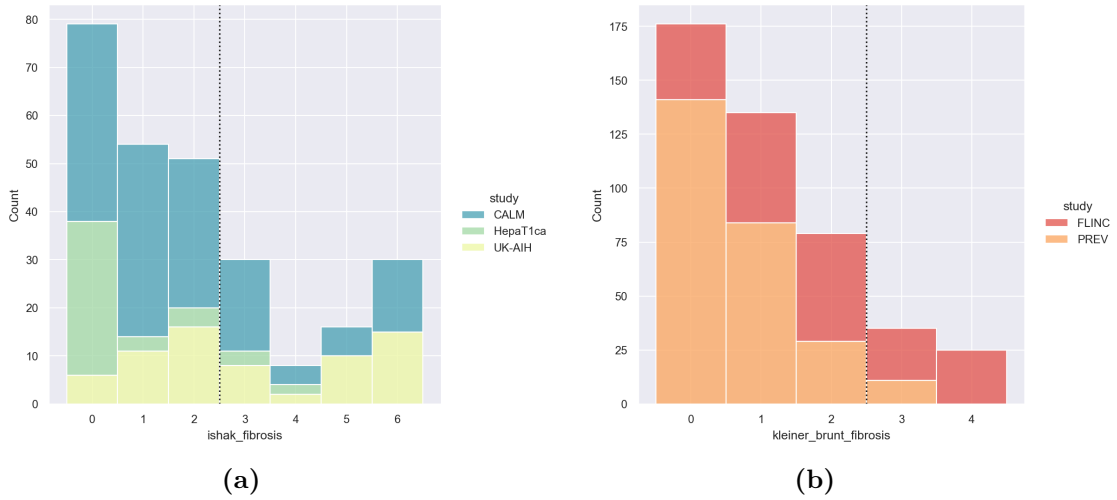


Figure 3.3: Distribution of fibrosis stages within the cohorts. a) Cohorts scored using the Ishak system. b) Cohorts scored using the Kleiner-Brunt system. The dotted line indicates the onset of bridging fibrosis in each of the scoring systems.

characteristics of the images studied in this DPhil project. Figure 3.4 presents digital slides from each of the individual studies which contributed data to this thesis. For every study participant at least two stainings were used: the basic morphology staining - haematoxylin and eosin (H&E) and the collagen staining - PicroSirius Red (PSR). As can be seen in Fig. 3.4, the specimens vary in both size and colour in each of the stainings. This is a result of differences in processing in the different healthcare institutions in which they were obtained. Below is a discussion of each of the factors which contribute to the differences in data quality and appearance between the cohorts.

Because of the number of hospitals engaged in preparation of the specimens studied in this thesis, it was unfeasible to obtain the exact protocol used for tissue processing in each of the cohorts. However, the processing of human tissue sections is highly normalised across healthcare institutions and is regulated by both domestic and international norms. In particular, the processing of formalin-fixed, paraffin-embedded (FFPE) tissue sections is currently regulated in the UK by the EN ISO 20166-4:2021 norm, which is identical to the International Standard ISO 20166-4:2021 [100]. These documents specify requirements and give precise recommendations for the collection, handling, documentation, transport, storage

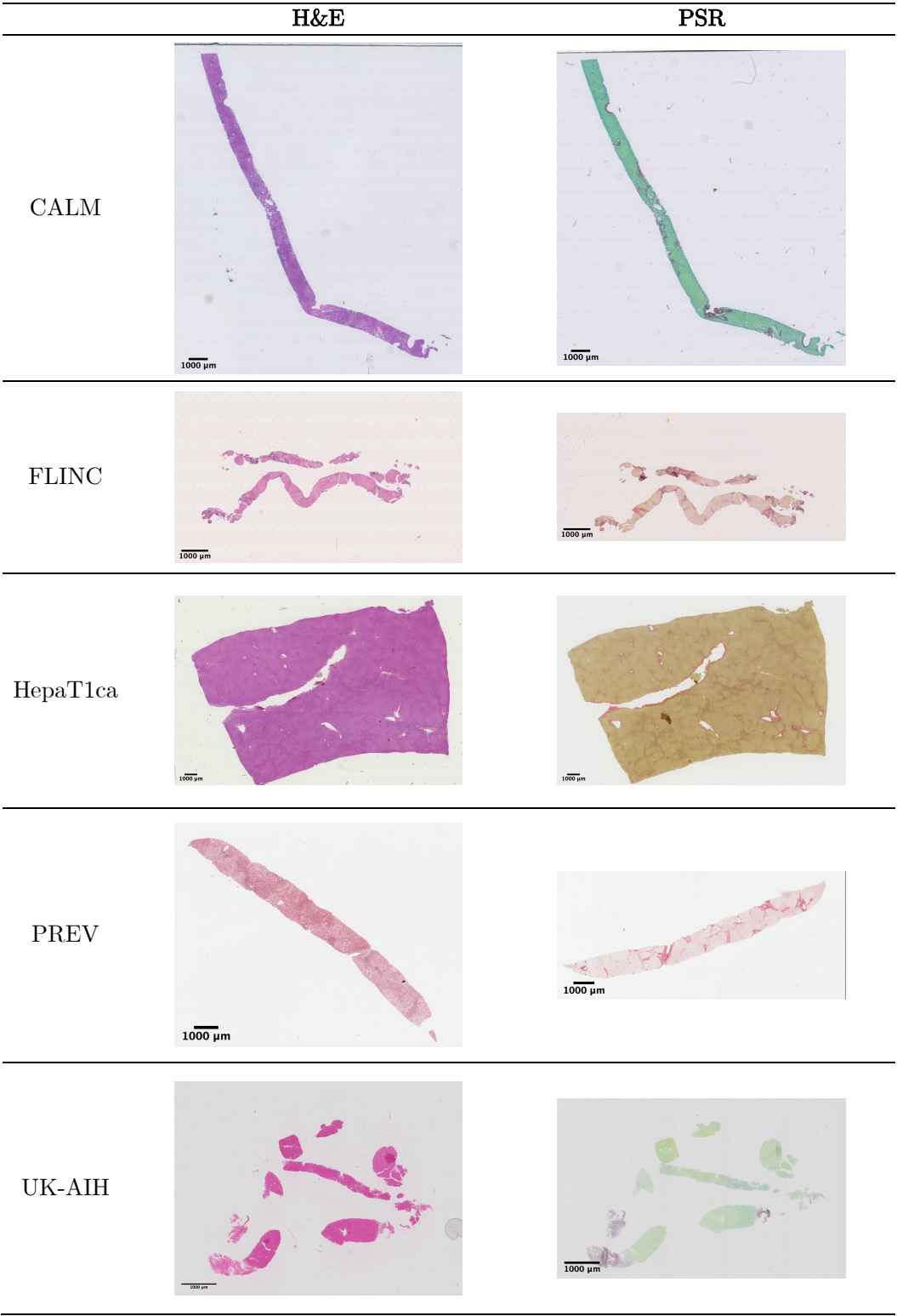


Figure 3.4: Examples of H&E and PSR digital slides from each of the cohorts studied. Differences in sample size and colour, as well as background illumination colour can be observed when comparing samples from different studies.

and processing during the pre-examination phase of FFPE tissue specimens intended for qualitative and/or (semi-)quantitative in situ examination.

The data in each of the prospective study cohorts: CALM, FLINC, HepaT1ca and PREV was collected according to the respective appropriate norms, and a single tissue processing protocol was used within each cohort. In the case of the retrospective cohort, UK-AIH, each site within the study employed its own protocol for processing tissue sections, with some sites changing their procedures over the 20 years in which the samples were collected. This understandably leads to large differences in the quality of samples in this cohort.

Fig. 3.5 summarises the process of obtaining and examining a needle biopsy slide in a clinical workflow.

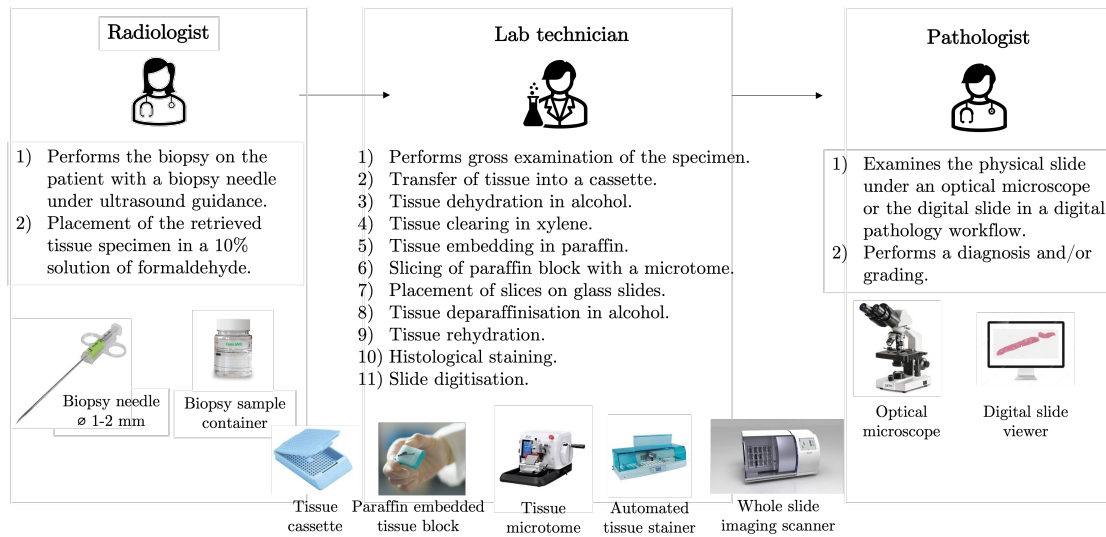


Figure 3.5: The digital slide acquisition workflow. At least three human actors are involved in the digital pathology workflow. The sample is acquired by a physician, processed by a laboratory technician, and finally it is examined by a pathologist. The slide's quality is affected by the quality of processing at each individual step in the pre-examination workflow.

The first step in the pipeline is the acquisition of liver tissue from a human donor. A percutaneous liver biopsy is performed through insertion of a biopsy needle and recovery of a small piece of the organ with it. In the UK this is currently typically done by a radiologist under ultrasound guidance [51]. In case of older biopsies, as found in the UK-AIH study, the biopsies were most likely taken by

a physician without ultrasound assistance. Alternatively, rather than inserting the needle percutaneously, a transjugular liver biopsy can also be performed. In a transjugular biopsy, the needle is inserted through the jugular vein in the neck. The technique is used to limit the risk of surface bleeding after the biopsy. Larger blocks of tissue can also be resected by a pathologist during surgery, as was in the case of the samples from the HepaT1ca study.

The extracted tissue goes through multiple steps of processing before an image of it can be obtained. Firstly, the clinician who performs the biopsy places the tissue in a solution of formaldehyde to preserve the sample as it is being transported to the tissue processing facility.

The tissue specimen is then processed by a laboratory technician. Figure 3.5 lists the multiple steps of tissue processing in a laboratory. The pipeline can be summarised into three parts: specimen preservation in paraffin, preparation of glass slides and histological staining. The quality of the resulting histology slide is impacted by each individual step in the pre-examination workflow [101].

For the purpose of long-term storage, the specimen is embedded in paraffin. Paraffin is hydrophobic (it is not water-soluble), therefore in order to allow it to infiltrate the tissue, the sample needs first to be dehydrated in ethyl alcohol and cleared in xylene, after which paraffin is added. After tissue paraffinisation the sample is left to cool until it forms a solid block. A specimen preserved this way can be stored for decades. The paraffin block is sliced with a microtome and single tissue slices are placed on glass slides. Typically, several slices are taken, and the rest of the block is left preserved.

A paraffin-embedded slice cannot absorb any stains, and therefore the tissue is then processed to reverse the embedding. This is done by washing the sample in ethyl alcohol with increasing proportions of water in the solution. Once the tissue has absorbed the water, the biological structures inside it are able to absorb and bond to the histological stains. The process of tissue rehydration and staining can either be performed manually or automatically. The current UK norm for FFPE processing requires the use of an automated tissue stainer for histology staining, as manual

processing is burdened with low reproducibility [100]. Samples from all cohorts in this thesis, except for the UK-AIH cohort have been stained automatically. Finally, after slide staining, coverslips are applied using a resinous mounting medium.

In the classical pathology workflow, such histochemically stained slide is then ready to be examined by a pathologist under a bright-light microscope. In a digital pathology workflow, the tissue is scanned with a Whole-Slide Imaging system at the appropriate magnification (typically 20x or 40x) and the slide is uploaded to and stored in a secure digital pathology system. A pathologist can then log into the system and view the digitised slide on a digital display.

3.3.1 Sampling variability and specimen size

When performing any histological analysis, one must be aware that the excised tissue specimen may not be representative of the organ on the whole. The liver is the largest gland in the human body and is estimated to weigh 1.5 kg in adult men on average [102]. A needle biopsy can only retrieve a tiny portion of the organ. The most important limitation of the procedure is that it isn't possible to verify how representative of the entire organ the acquired sample is. This limitation is particularly important in assessment of liver fibrosis, where the pathological structures observed (fibrotic bridges or nodules) often exceed in size the width of a typical biopsy [86]. In spite of these limitations, liver biopsy remains the golden standard in the diagnosis of multiple diseases, including MAFLD.

To mitigate the risk of inaccurate readings, The European Association for the Study of the Liver (EASL) and its American counterpart - The American Association for the Study of Liver Diseases (AASLD) publish guidelines for the process of obtaining liver samples from patients. According to the newest UK and US guidelines for the percutaneous liver biopsy, a specimen is considered "adequate" for histopathological assessment if it contains more than 10 portal tracts [51, 103]. An adequately sized sample is meant to enable the pathologist to perform a reading and diagnose the patient with very high certainty. However, this does not mean that smaller samples are to be rejected from analysis in either clinical

Table 3.2: A summary of tissue slice areas in each cohort.

Cohort	Mean Tissue Area [mm^2]	Standard Deviation [mm^2]
CALM	13.5	5.1
FLINC	13.8	16.2
HepaT1ca	187.8	67.5
PREV	11.8	4.2
UK-AIH	10.0	15.9

practice or research studies. It is left to the pathologist's discretion to decide how confident they are reading from a particular sample and to note the certainty of their reading in the pathology report.

To achieve adequately sized samples, the guidelines recommend the use of a 16-gauge needle to retrieve at least 20 mm of tissue. A 16-gauge needle has a nominal inner diameter of 1.194 ± 0.038 mm. Assuming that a tissue slice is cut out at a cross-section of the biopsy cylinder with a width corresponding to $3/4$ of the diameter, these measurements correspond to a slide area of approximately 17.9 mm^2 .

In this project, the data on the number of pathologist-observed portal spaces was only collected in one of the five cohorts - the PREV study. Out of 276 biopsies in the cohort only 79 (29%) had the minimum desired number of portal spaces, with the mean value being 9.5. The mean tissue areas computed for each of the cohorts are summarised in Tab. 3.2. Here, it can be seen that the surgical tissue sections collected in the HepaT1ca trial are larger than needle biopsies collected in the other cohorts, by about an order of magnitude (see example samples in Figure 3.4). Because of this difference, results of examination from the HepaT1ca cohort cannot be directly compared to those from the other cohorts. Due to a large number of WSIs with imaging artefacts in the FLINC cohort, the tissue area values here are reported for 146 out of the 203 tissue slides (72%). Further information about slide quality in the FLINC cohort is available in Appendix A. Only 137 out of 879 (16%) of all PSR tissue slides in this thesis have tissue area equal to or higher than that recommended by the most recent guidelines.

In this DPhil project, each of the samples was individually assessed by an expert pathologist with regard to slide adequacy. Each of the research chapters

in this thesis contains a brief summary of the samples rejected from analysis in the context of the particular research question.

3.3.2 Histological staining for liver fibrosis

Thin sections of biological tissues are almost entirely transparent to visible light and the structures within them are not easily differentiable under an optical microscope. Histochemical stains are used to create visual contrast between tissue structures through in-situ localisation of their chemical components. Stains used in histology work on the basis of high chemical affinity between the dye and its target molecule and low affinity to all the other substances present in the specimen. For example, the most widely used histochemical staining, H&E is based on the differential uptake of haematoxylin and eosin depending on the pH of different cellular compartments [104].

Liver fibrosis is a condition characterised by the pathological deposition of collagen fibres within the liver lobule, and therefore it is assessed using collagen-binding stains. Other connective tissue fibres, such as elastin fibres are also present in fibrotic structures, however in comparatively small quantities [105]. Several stainings for identifying connective tissue in general and collagen specifically have been independently developed and traditionally used in pathology studies and in the clinic [23]. These include: PicroSirius Red, Masson's Trichrome staining for connective tissue, van Gieson's stain, Orcein and others. Each of these stainings is used in pathology to make the structure of connective fibres visible and can be used to visually assess the grade of fibrosis. However, the only histochemical stain which is known to be collagen-specific and so can be used to quantify collagen is Sirius Red.

PicroSirius Red (PSR) is a staining developed in the 1960's with the purpose of selectively staining collagen in histology samples [106]. The primary agents in this staining are Sirius red F3B (CI 35780, Direct red 80) and picric acid (2,4,6-trinitrophenol, TNP). Sirius red F3B is a red acidic dye which is commonly used as a colourant for textiles, leather and even as a hair dye. Picric acid is a yellow dye used in histochemical staining for two simultaneous purposes: as a fixing agent and to improve the intensity of staining using acidic dyes. It can also

be used on its own as a yellow general cytoplasmic and connective tissue fibre stain. In the presence of picric acid, Sirius red F3B selectively binds with type I collagen - thicker collagen fibres, and type III collagen - thinner fibrillar collagen (reticulin fibres) [107]. Because of the selective staining for type I and III collagen fibres and its sensitivity to fine collagen strands, the PSR protocol is used for the quantification of collagen in histology slides.

In this thesis, only Sirius red-stained slides were used in analysis of liver histology, both for pathologist staging and in the development of digital pathology methods.

3.3.3 Sirius red staining in thesis cohorts

In spite of its wide-spread use as a collagen-specific stain, there isn't a single standard protocol for PSR staining, which is used globally across institutions. The classical form of PSR staining includes only the mixture of Sirius Red and picric acid for collagen detection. However, laboratories often add additional stains to the same tissue slice, to simultaneously examine other forms of tissue pathology. Commonly, two other stains may be added: a nuclear stain (a haematoxylin) and a cellular plasma stain (typically a green dye). This lack of PSR staining standardisation between healthcare institutions leads to very significant differences in the colour of the stained slides. This phenomenon can be well observed within the UK-AIH cohort where 20 NHS hospitals contributed liver FFPE slides prepared within their pathology laboratories (Fig. 3.6). The figure illustrates how deep the visual differences between PSR-stained slides can be when comparing samples collected across different sites. Below is a comparison of the ways in which the samples from each of the cohorts in this study were stained for fibrosis.

The samples within each of the four prospective cohorts in this thesis: CALM, FLINC, HepaT1ca and PREV have been processed according to internally agreed processing and staining protocols. The specimens from all of these studies have been stained with automatic stainers. In contrast to the above cohorts, the UK-AIH study was conducted retrospectively, gathering pathology slides from 20 UK-based hospitals, with the oldest samples collected in 1998. The FFPE slides in this cohort



Figure 3.6: A visual representation of a selection of Sirius red-stained liver slides from the UK-AIH study. This image illustrates how the use of different protocols for PSR collagen staining results in dramatic differences in slide appearance within the cohort collected retrospectively from 20 NHS hospitals. The images are not to scale. In each digital slide the background was removed and colours adjusted through histogram equalisation for visual clarity.

were processed according to the internal standards within each of the participating hospitals at the time they were collected.

Only one of the PSR staining protocols was shared with me during the duration of this thesis project - for the FLINC study conducted by Novo Nordisk, Denmark. The protocol is attached in full in Appendix [A](#). In this protocol the tissue is submerged in a water solution of Sirius Red with picric acid (PicroSirius Red in the protocol). Additionally, Weigert's iron-haematoxylin staining was performed on the same slides, according to the modified Lillie's recipe, along with Alcian Blue staining for mucin.

In order to identify the stains used in each of the respective studies in the absence of staining protocols, stain deconvolution was performed on selected digitised slides from each cohort [\[108\]](#). Stain deconvolution is a transformation of the image from the RGB colourspace into a colourspace in which each of the three image channels encodes the proportion of a colour of a single stain. This processing was performed manually by me using QuPath software [\[109\]](#).

The results of stain deconvolution for example Sirius red stained slides from all cohorts are presented in Figure [3.7](#). It can be seen that while in each of the

Table 3.3: A comparison of the stains used in the Sirius red-stained slides in each of the clinical cohorts. In the UK-AIH cohort many different staining protocols were employed.

Cohort	SR	Haematoxylin	Other
CALM	✓	-	Unknown Green
FLINC	✓	Weigert's Haematoxylin	Picric Acid, Alcian Blue
HepaT1ca	✓	-	Unknown Olive, Unknown Blue
PREV	✓	-	-
UK-AIH	✓	Purple Haematoxylin/-	✓/-

cohorts, a red collagen dye was used, different stains were used alongside it. This is in stark contrast to the results of stain deconvolution applied to the corresponding H&E slices from the same samples [3.8]. In the H&E-stained slides, the number of detected stains was constant across study cohorts: in every cohort only a blue or purple haematoxylin and an eosin was used. In the PSR-labelled slides, the number of dyes used was ranging from one (just Sirius red) to four (e.g. Sirius red, picric acid, Weigert's haematoxylin and Alcian blue). In the case of the non-homogeneous UK-AIH cohort, different numbers and colours of stains were used for PSR staining. Table [3.3] summarises the stains detected in the digital slides from each of the studies.

3.3.4 Slide digitisation

The glass slides were digitised using whole slide imaging (WSI) scanners. Each of the cohorts was scanned internally using parameters chosen by the individuals conducting the respective studies. The information about scanners and slide parameters for each of the cohorts is summarised in Table [3.4].

As can be seen, scanners from several WSI manufacturers were used for data acquisition in this thesis. WSI systems from different vendors are equipped with different optical instruments and digital image sensors. These differences in hardware mean that the whole-slide images were acquired at slightly different resolutions, given the same nominal magnification (e.g. 20x). Similarly, scanning the same histology slide with two WSI scanners from different vendors produces digital images with slightly different colours [110]. This can be a result of both hardware differences

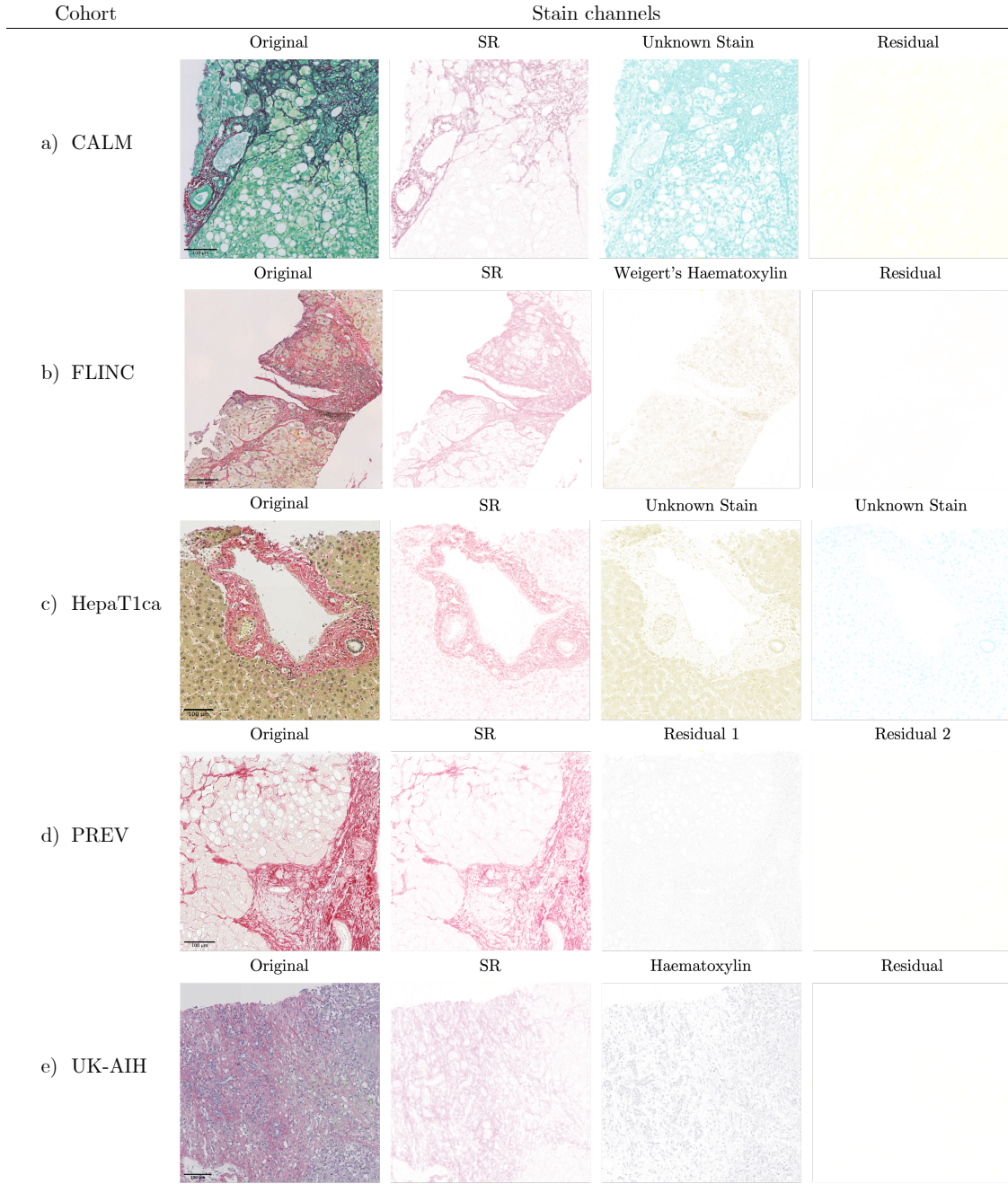


Figure 3.7: The results of stain deconvolution on example digital PSR slides from all cohorts. a) CALM: the black colour of some collagen fibres is a result of non-specific binding of an unknown green stain to fibres stained with Sirius red. The slides contain a pale yellow residuum. b) FLINC: Sirius red and Weigert's haematoxylin, with little residuum. Upon close inspection one can see that cell nuclei are non-specifically stained red along with the collagen. c) HepaT1ca: Sirius red and two unknown stains, olive and blue. Upon close inspection one can see that cell nuclei are non-specifically stained red along with the collagen. d) PREV: only Sirius red stain was detected. e) UK-AIH: slide stained with Sirius red and a purple haematoxylin with an empty residual channel.

Table 3.4: Overview of digital slide parameters for each of the cohorts.

Cohort	Scanner	Magnification	Pixel Size [μm]	File Format	Compression
CALM	Hamamatsu C12000-02	20.0 x	0.4533, 0.4533	*.ndpi	JPEG <i>level unknown</i>
FLINC	Hamamatsu C13210	40.0 x	0.2209, 0.2209	*.ndpi	JPEG <i>level unknown</i>
HepaT1ca	Hamamatsu C12000-02	20.0 x	0.4533, 0.4533	*.ndpi	JPEG <i>level unknown</i>
PREV	Aperio <i>model unknown</i>	20.0 x	0.4954, 0.4954	*.svs	Aperio JP2000 RGB
UK-AIH	Leica SCN400	20.0 x	0.5000, 0.5000	*.scn	JPEG <i>level unknown</i>

and the colour calibration of a particular scanner. Each vendor also typically uses a proprietary image format. While the differences in colour could not be directly normalised due to lack of appropriate data, the differences in resolution were taken into account when performing area measurements on digital slides.

In this thesis all slides were processed using the vendor-neutral libraries: OpenSlide [111] and wsi-reader [1]. Wsi-reader is a Python library developed by Stefano Malacrino within the Rittscher group at the University of Oxford, designed to offer a unified interface for the most common WSI libraries, such as openslide, tiffslide [2] and Philips Pathology SDK. The library was being actively developed throughout the time of this DPhil project and some of the image analysis pipelines shown in this thesis were used to test some of its functionalities.

In total, 4211 liver WSIs (1338 GB of total disk space), including 882 PSR-stained slides (223 GB) were gathered in this thesis.

3.4 Colour characterisation of digital slides

In order to visualise the spectrum of Sirius-red-stained digitised FFPE section appearances, I here propose the concept of a primary colour of a digital slide (Fig. 3.9).

¹<https://github.com/stefano-malacrino/wsi-reader>

²<https://github.com/cgohlke/tiffslide>

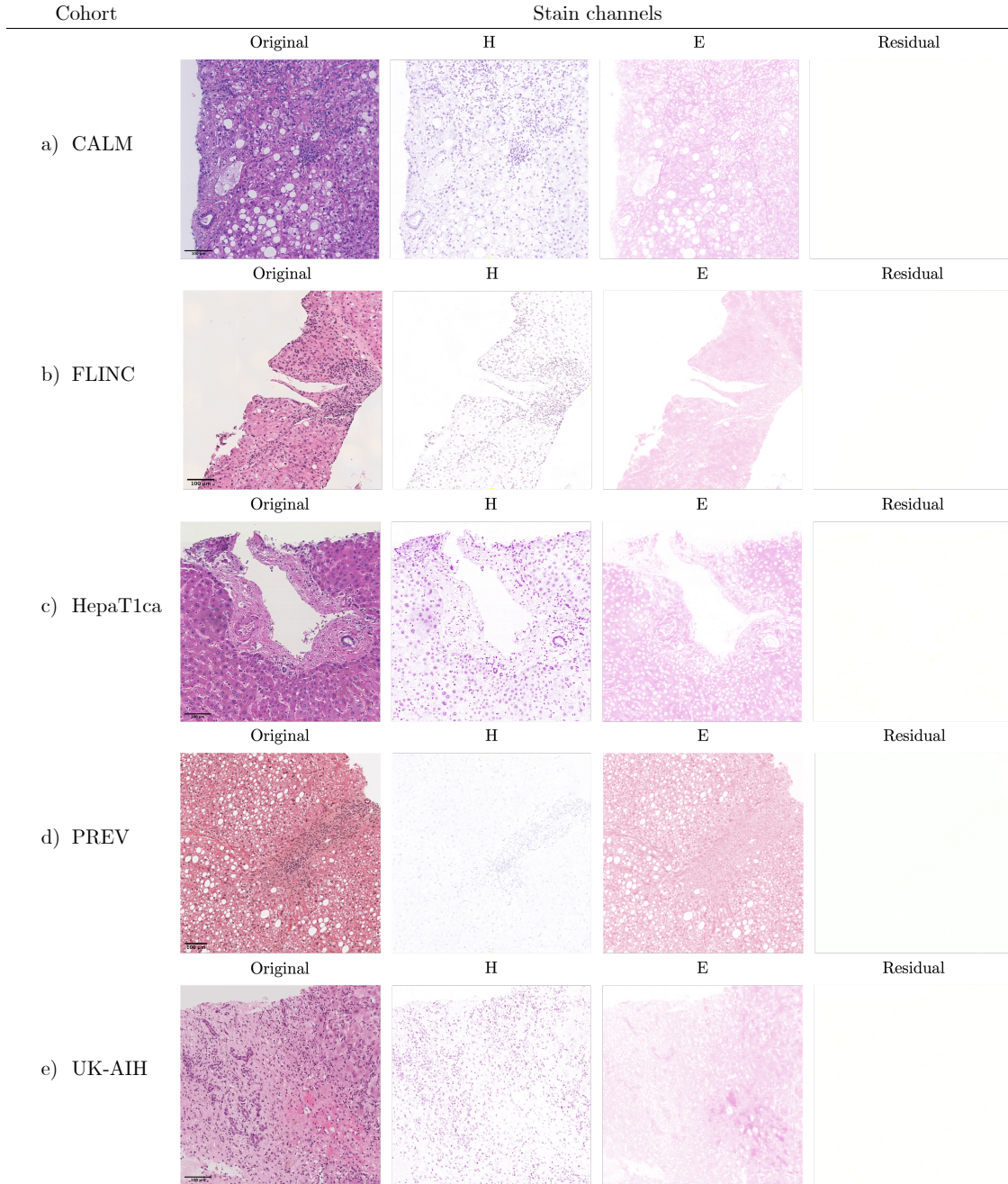


Figure 3.8: The results of stain deconvolution on example digital H&E slides from the five study cohorts. In each of the cohorts a haematoxylin and an eosin were used to produce H&E slides. The residual channel is empty or nearly empty in all of the slides. Slight differences in stain hue of both haematoxylin and eosin, as well as scan background colour can be seen between samples from different cohorts.

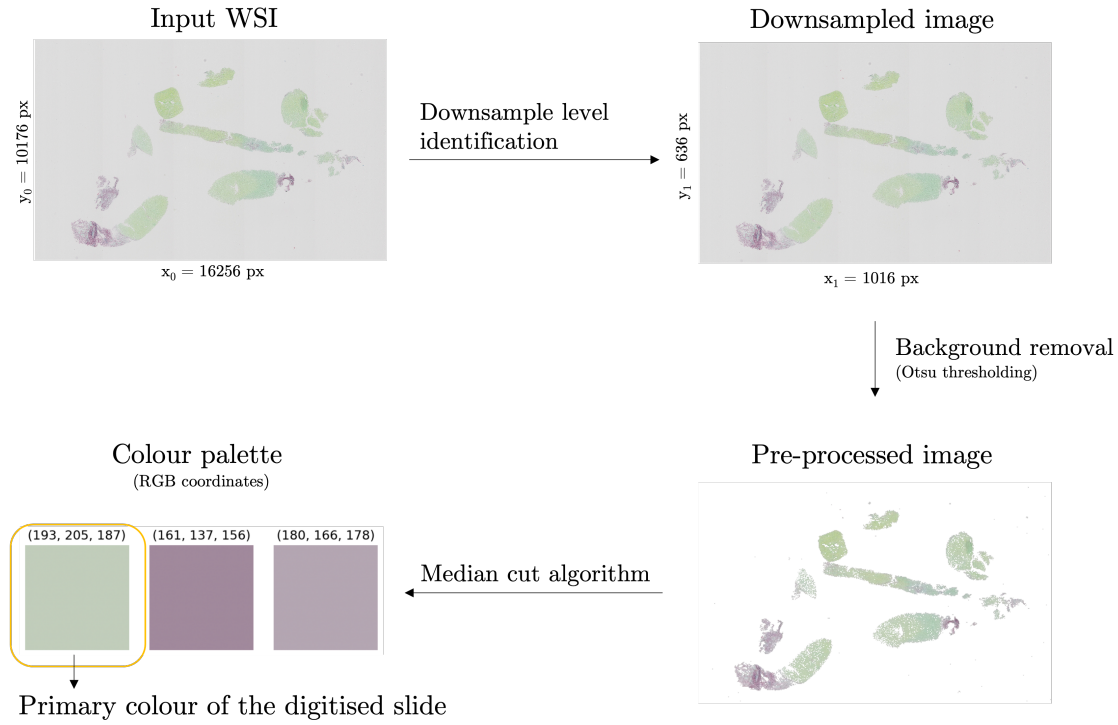


Figure 3.9: Identification of the primary colour of a digital slide. First, an appropriate downsample level is identified in the WSI pyramid. The slide image is then read from the downsampled level of the WSI file. Tissue/background intensity-based threshold is identified with the Otsu thresholding algorithm (applied to the greyscale-converted image). Background pixels are replaced with white values. Finally, the median cut algorithm is applied to thus pre-processed image to identify the primary RGB value in the digital slide.

One of the most commonly used approaches to quantisation of colours in digital images is through the employment of median cut algorithm [112]. Median cut algorithm is a method of sorting data of an arbitrary number of dimensions into series of sets by recursively cutting each set of data at the median point along the longest dimension. When applied to an image histogram, it outputs the most commonly occurring pixel values. Here, the WSI is first pre-processed by extraction of a downsampled version of the image and background removal. A pixel intensity threshold between tissue and background is established using the Otsu method applied to a grey-level version of the image [113]. All pixels of value above thus established intensity threshold are replaced with white values. Median cut algorithm is applied to the pre-processed image to extract a three-colour colour palette, and the RGB values of the most commonly occurring colour encode the

primary colour of the digital slide.

The primary colour has been computed for all of the SR and H&E-stained digitised slides available to this project. One PSR-stained slide from the FLINC cohort was excluded from colour analysis due to an ink artefact. For purposes of comparison the RGB values of the primary slide colours were converted to the more perceptually uniform CIELAB colourspace [114]. The CIELAB colourspace expresses each colour as three values: L^* for lightness, a^* where negative values indicate green and positive values indicate red, and b^* where negative values indicate blue and positive values indicate yellow. The results of primary colour estimation for the WSIs from all five thesis cohorts are shown in Figures 3.10 and 3.11. The distribution of the numeric values of the primary colours is shown in Appendix A.5.

3.5 Conclusions

In this chapter I've presented a description of the digital liver pathology data studied in this thesis. The collected data is unusual in its breadth, comprising samples obtained in over 20 health institutions across three countries. The participant cohort is predominantly white and middle-aged, with an equal representation of males and females. The combined cohort is highly heterogeneous both in terms of liver disease representation and the methods applied to data preprocessing. The cohort represents all stages of liver fibrosis, with a four-times higher representation of early fibrosis than advanced fibrosis.

The heterogeneity of the cohort has two kinds of consequences from the perspective of this thesis. Firstly, it limits the number of research questions which can be asked of the data as a whole. For example, fibrotic samples from patients with AIH cannot be directly compared to those from patients with MAFLD in terms of disease staging. However, the multiplicity of healthcare institutions involved in the gathering of the data provides us with an opportunity to examine the problem of accurately measuring collagen content in clinical, Sirius red-stained liver biopsies.

A further limitation of the cohort results from the use of three established, competing fibrosis staging systems as a basis to perform the histological scoring.

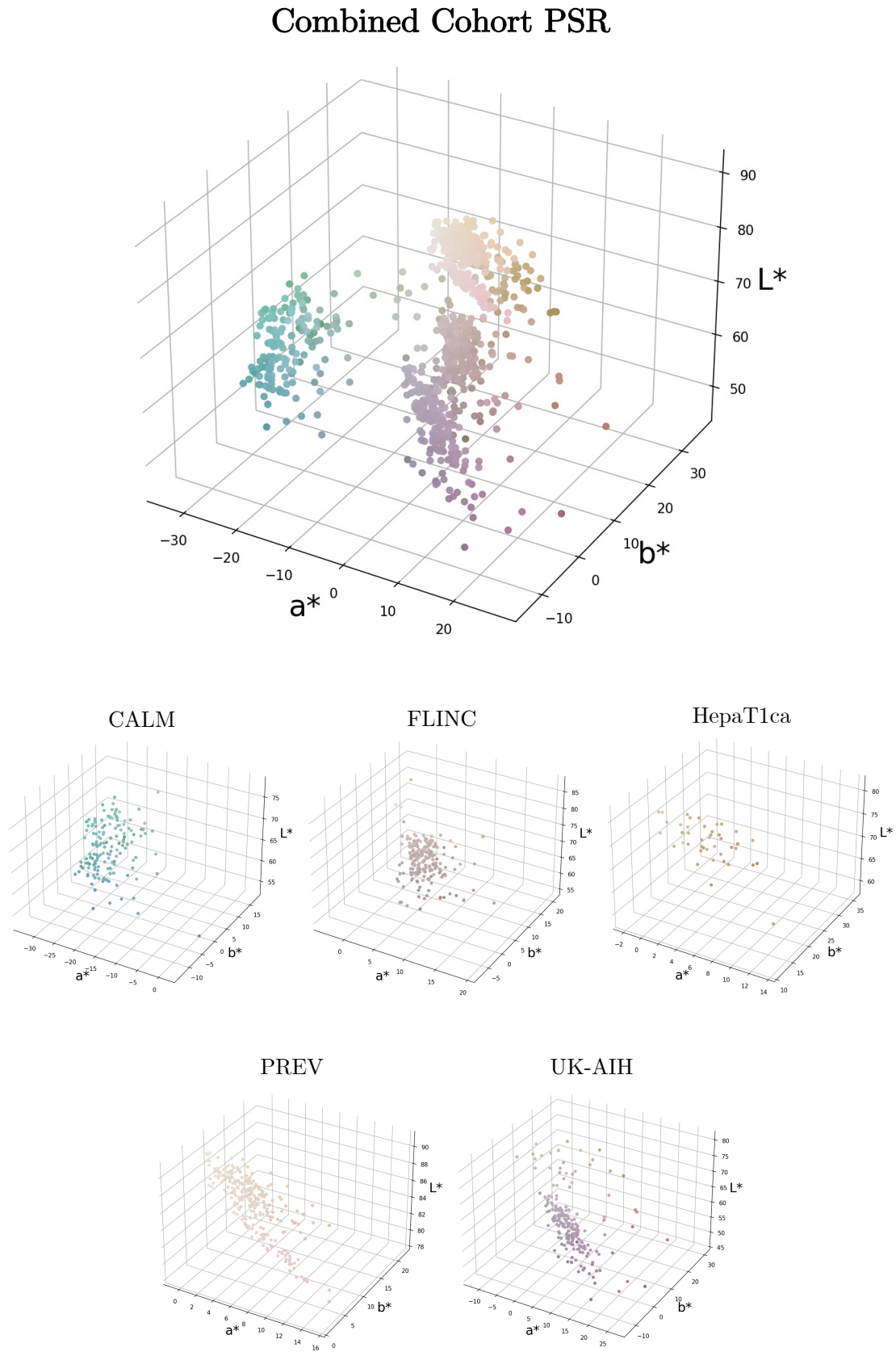


Figure 3.10: Primary colour distribution in digital slides of PSR-labelled sections. Samples from the individual cohorts form separate colour clusters.

Combined Cohort H&E

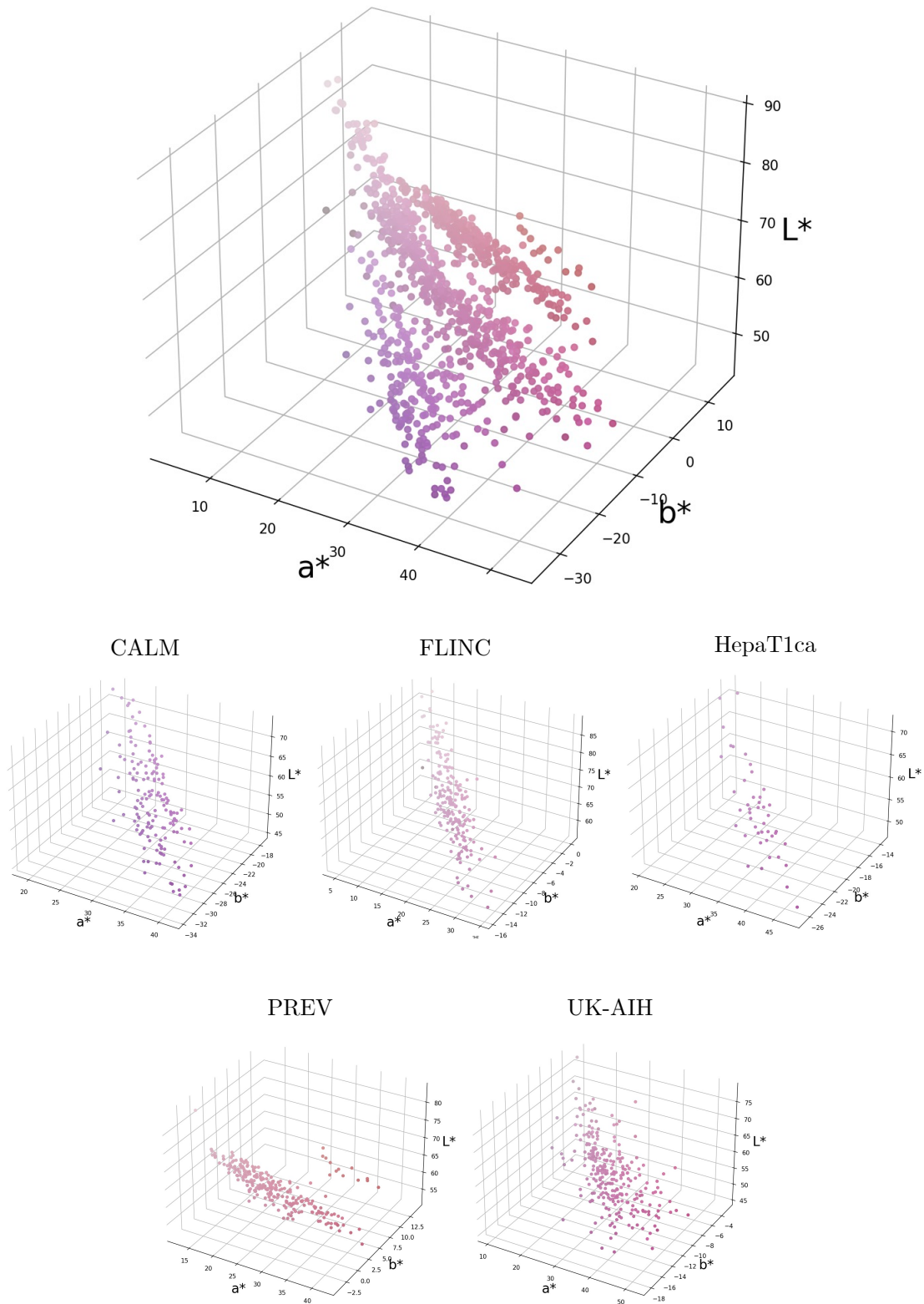


Figure 3.11: Primary colour distribution in digital slides of H&E-stained sections.

The samples in each of the five studies were scored by individual groups of expert pathologists, who would not have been able to hold a consensus conference [115]. This means that there are likely minor discordances in disease staging between the studies, even where the same scoring system was used.

Based on the characteristics of the data described in this chapter, different cohorts were chosen as the basis of investigation in the consecutive research chapters of this thesis.

Chapter 4 - Identification of Collagen in Liver Histology Slides focuses on the challenge of precise identification of collagen in histochemically stained liver histology slides. The methods developed in this chapter make use of Sirius red-stained liver slides from the described cohorts scanned at 20 x nominal magnification (all except the FLINC cohort). Chapter 5 concerns automatic grading of liver fibrosis in MAFLD and is based on the examination of liver sections from the PREV cohort.

Finally, in Chapter 6, CPA quantification is applied to samples from all thesis cohorts. Furthermore, collagen pattern subtypes are learned in an unsupervised manner from samples from CALM and PREV studies. The learned collagen pattern subtypes are then predicted for samples from CALM, HepaT1ca, PREV and a subset of the UK-AIH cohort. The possibility of applying the methods developed in this chapter to fibrosis progression tracking are explored, using a subset of 35 participants from the UK-AIH cohort, for whom longitudinal data is available.

4

A New Digital Lens for Quantifying Fibrosis in Histology Slides

Fibrosis is increasingly recognised as a core functional component of the tissue microenvironment. Collagen proportionate area (CPA) is an important image-based measure of fibrosis content in whole slide images (WSIs) of liver tissue samples [21]. Collagen proportionate area has already been shown to be an independent predictor of long-term clinical outcomes in patients with metabolic-associated liver disease (MAFLD) [20] and alcoholic liver disease (ALD) [116]. Because of this, arguments have been made for the inclusion of the change in CPA as an endpoint in clinical drug trials [10]. However, in practice the published clinical findings on CPA as an image-based biomarker are often inconsistent in their collagen detection and quantification methodologies [23]. This leads to major discrepancies in the quality of reported results and hinders comparisons between individual studies.

In a typical engineering setting reporting of the result of any measurement is expected to be accompanied by an estimate of the result's uncertainty. I observe that reporting uncertainty values is not currently common practice in the literature concerning computation of stain positive tissue area, be it PSR collagen staining or proportions of any other stains. Obtaining such estimates is challenging due to the high complexity of the measurement task.

A critical view of CPA as an imaging biomarker needs to employ a combination of methods and approaches from disciplines spanning biochemistry, microscopy and finally metrology and computer vision. In this research chapter, I describe the multiple factors which affect the process of performing a measurement of collagen percentage and propose a way of estimating the measurement's uncertainty. This is done by applying an ensemble of deep networks to model a distribution of the possible solutions to the segmentation problem. The range of ensemble predictions is then used to obtain an average pixel-wise collagen likelihood as well as the confidence intervals of the prediction.

4.1 The Drivers of Uncertainty in CPA Measurement

Firstly, it must be acknowledged that any biopsy sample may not be representative of the examined organ on the whole. This limitation, caused by the small size of the acquired specimen, affects all histology-derived metrics and predictions, and its examination goes beyond the scope of this thesis. However, from the moment of the acquisition of the biopsy, there is still a number of confounding factors which need to be considered in order to obtain a reliable measure of collagen area in the analysed sample.

The process of CPA measurement can be broken down into three separate detection and quantification tasks. These are: 1) collagen in-situ localisation (i.e. chemical staining), 2) the acquisition of the digital image and 3) the quantification of the stained area relative to the area of the tissue section. Each of these processes is burdened by its own errors which together contribute to the final magnitude of the uncertainty. It can be therefore seen how the quality of collagen area quantification from a formalin-fixed paraffin-embedded (FFPE) tissue specimen is affected by each of the steps in the pre-examination digital pathology workflow (see Fig. [3.5](#)).

4.1.1 In-situ localisation of collagen

While PSR staining is known to create differential contrast between collagen strands and other tissue components, the exact sensitivity and specificity of PSR stain binding to collagen in a liver tissue section is not known [107]. Additionally, parameters such as the freshness of the specimen, stain concentration, stain exposure time and ambient temperature can all affect the intensity of staining.

Furthermore, staining contrast can be increased or decreased by application of additional chemicals. An example of this is the addition of the yellow-coloured picric acid to the Sirius Red solution in order to increase the perceived visual contrast between tissue collagen and parenchyma. However, tissue staining is a destructive process and different stains can block one another. This is for example the case in slides in which a green counterstain has been applied (such as from the CALM cohort). In these slides it can be seen that collagen fibres are stained both with Sirius Red and the green counterstain, making the collagen regions appear black (see Figure 3.7). It is not clear how the interactions of different chemicals and even the order in which different stains are applied affects the sensitivity and specificity of stain-collagen binding.

Finally, histological stains are known to fade over the course of years and therefore retrospectively collected slides may not be comparable to freshly obtained ones in terms of colour intensity.

4.1.2 Digital sampling

The thickness of an individual collagen fibre is well below the x,y imaging resolution of a typical WSI scanner [117, 118]. Collagen fibres which can be seen in WSIs are aggregates of many collagen fibrils and are characterised by highly complex fractal-like shapes. Because of this fractal-like quality of the analysed shape the measure of collagen area is intrinsically tied to the x,y resolution of the microscope objective lens and the size of the pixel of the imaging sensor [119]. In fact the dependency of the quantified collagen area on the imaging magnification is so

strong that fractal metrics of dimensionality have been successfully applied to predict the overall stage of fibrosis [77].

Beyond the imaging resolution in the xy -plane, parameters such as slice thickness, lighting conditions and sensor exposure time directly affect the intensity of the colour in the digitised slide [22]. Clinical WSIs are typically acquired at either 20x or 40x magnification, which corresponds to x,y pixel sizes of approximately $0.5 \times 0.5 \mu\text{m}$, and $0.2 \times 0.2 \mu\text{m}$, respectively (see Tab. 3.4). By contrast, typical clinical biopsies are sliced at the thickness range between 4 and $10 \mu\text{m}$. Because of this a WSI pixel can contain a wide range of collagen thicknesses within the corresponding z volume of the tissue [120].

The final colour intensity of an imaged pixel is therefore a function of 1) the thickness of the tissue slice, 2) the thickness of the imaged collagen fibre in the tissue volume and 3) the relative intensity of collagen staining.

4.1.3 Segmentation errors

Finally, the results of CPA quantification depend on the quality of the segmentation of collagen fibres from the WSI acquired at a given resolution. Fig. 4.1 illustrates the difficulty of annotating collagen on an example tile from a SR stained liver biopsy specimen. In this section from the PREV study, only one stain, Sirius Red, has been applied to the tissue slice. The stain is present in all pixels belonging to the tissue section, though its intensity differs from region to region. While collagen fibres have an intensely red colour, it can be seen that other regions within the tissue parenchyma (cell cytoplasm and even more so cell nuclei) are also coloured pink. In many regions it is not clear if the colour results from the presence of thin collagen fibres or from unspecific staining of other tissue components. Furthermore, as we analyse each image at full resolution, compression artefacts start to affect its quality. Because of this, annotating collagen fibres in liver histology slides is burdened with high inter- and intra-observer variability [24].

Performing manual or even semi-automatic annotations of collagen area is a tedious task which can easily be automated by the employment of image analysis

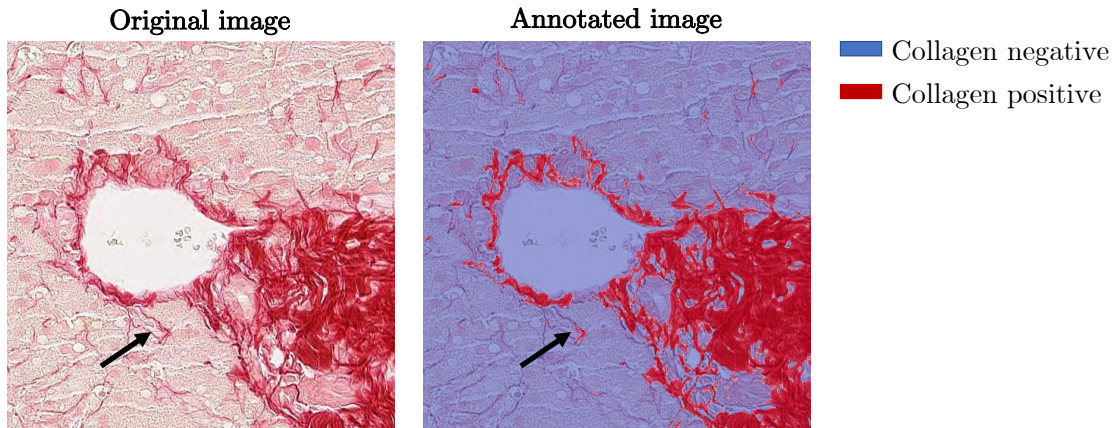


Figure 4.1: Limitations of creating human annotations of collagen fibres in microscopy images. The task is highly subjective and characterised by high inter- and intra-observer variability, particularly for thin collagen strands and in tissue slides with poor staining quality. It can be seen that while regions with thick, intensely stained collagen are captured, thinner fibres were missed by the annotator (MW).

methods. While simple annotation tools often rely on the manual selection of a pixel intensity threshold by the annotating pathologist, such pipelines are known to be vulnerable to stain intensity gradients which are often found in histology slides. Deep learning segmentation models can be trained against pathologist segmentations to be robust to minor and even major differences in stain colour.

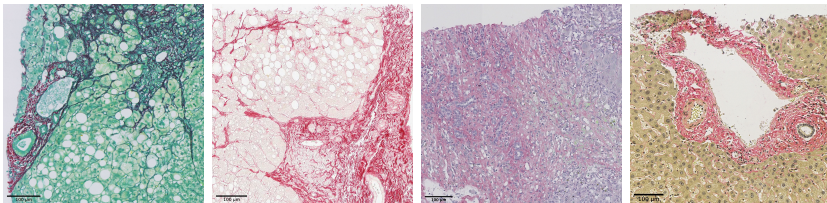
4.2 Data

Four separate sets of liver histology WSIs have been used in this study. Three sets, referred to by their original study names: CALM [92], PREV [95] and HepaT1ca [94] were made available by Perspectum, Oxford, UK. An additional cohort of retrospectively collected clinical FFPE WSIs was made available by the UK-AIH research consortium [96].

The liver tissue slides in each of the cohorts were all stained for collagen using Sirius Red (SR) staining, which dyes tissue collagen red [107]. The individual studies were performed with the use of differing SR staining protocols, where different variants of counter-stainings were applied. The glass slides were all digitised at 20x nominal magnification at the facilities used in each of the individual studies. Table

[4.1](#) shows a comparison of tissue appearances in each of the cohorts along with detailed information about the training and validation data in this paper.

Table 4.1: A summary of training and validation data for each of the study cohorts. The images show the appearance of the SR-stained slides in each cohort.



Cohort	CALM	PREV	UK-AIH	HepaT1ca
Disease	Mix	MAFLD	AIH	Cancer/MAFLD
N Annotated Cases	12	13	13	5
N Annotated Tiles	495	1313	408	151
N Train (Tiles)	442	1172	356	0
N Validation (Tiles)	53	141	52	151
Training Set Collagen	21%	15%	24%	-

In total 38 whole slide images of liver biopsy sections were densely annotated by MW using QuPath software [\[109\]](#). The pipeline for semi-automatic annotation of the slides with human in the loop is shown in Figure [4.2](#). Tiles containing less than 5% of annotated collagen area were discarded from the training and validation set to improve the stability of model training. All tiles from the HepaT1ca study were retained as an independent validation set. A total of 2367 annotated 512x512 px tiles were used in the analysis. Of these 1970 were used for model training and 397 for model validation.

4.3 Segmentation Models

The U-net is a popular deep learning model used for medical image segmentation [\[84\]](#). In this thesis three variants of the U-net model have been trained for the collagen segmentation task using the implementation available at [\[1\]](#). The first two models (here called U-net Tiny and U-net Mini) are each down-sized versions of

¹<https://github.com/yingkaisha/keras-unet-collection>

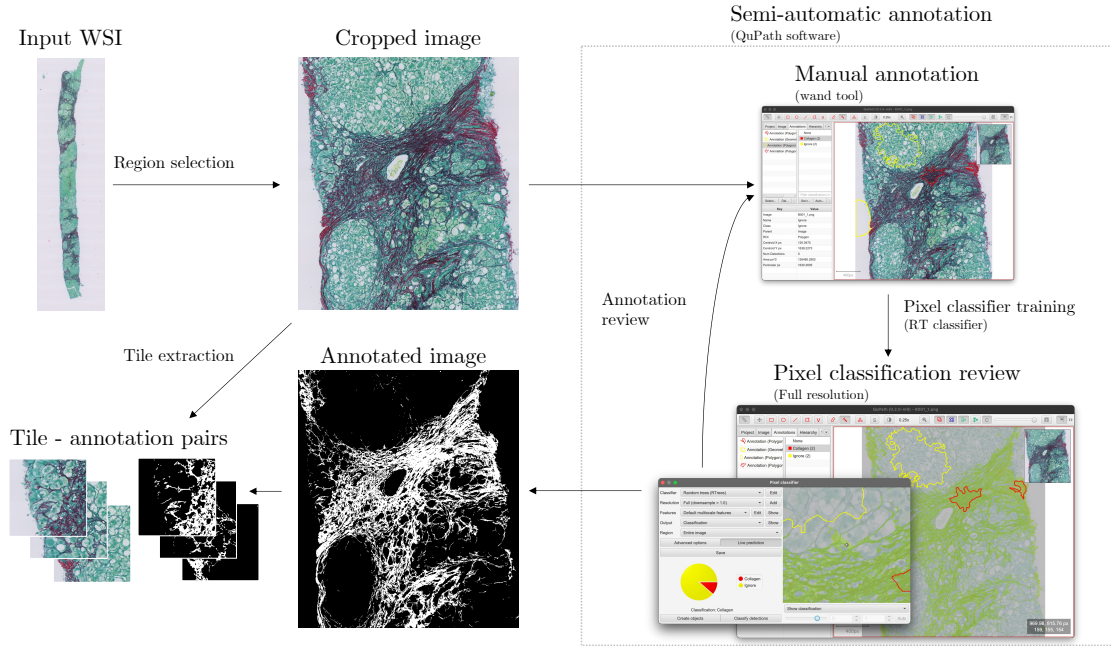


Figure 4.2: Collagen annotation with human in the loop. Firstly, a representative region is cropped from the full resolution level of the WSI file. The annotator trains a built-in random trees (RT) pixel classifier to categorise individual pixels within the image as either: "collagen" or "ignore". The results of the RT classification are carefully examined visually. If needed, the human annotations are refined, and the process is repeated until the classifier produces a satisfactory pixel-level classification for the entire image. The annotation is then exported as a binary mask.

the original U-net. Additionally an Attention U-net [121] was trained for the same task. The parameters of the individual networks are summarised in Table 4.2.

Table 4.2: A summary of the parameters of the trained segmentation models.

Parameter	U-net Tiny	U-net Mini	Att U-net
Depth	3	3	3
Number of Filters	[8, 16, 32]	[32, 64, 128]	[32, 64, 128]
Stacked Layers (Down)	2	2	2
Stacked Layers (Up)	2	2	2
Trainable Parameters	32 665	517 729	612 691

Each of the trained segmentation networks has an input which matches the size of 512x512 px RGB image tiles. The ReLU activation function was used in all of the networks. U-net Tiny and U-net Mini used max pooling as the pooling strategy and a sigmoid activation function at the network output. Each of the models used bilinear unpooling. The models were trained using binary crossentropy

loss and Adam optimiser. Each of the networks was trained from scratch with a randomised initialisation seed and a batch size of 8. A Quadro RTX 6000 GPU core was used for all of the experiments conducted.

4.4 Uncertainty Estimation

In order to estimate the real uncertainty of PSR measurement, one would need to perform an uncertainty estimation for each of the steps in sample processing and image acquisition separately. Unfortunately, as only digitised slides were made available to this project, such an analysis goes beyond the scope of this thesis. Here we propose a way of estimating the uncertainty inherent to the digital image itself, which can be inferred by the image segmentation deep network.

It has been shown that uncertainty estimation in the context of deep learning can improve the quality of the predictions [122, 123]. Bayesian inference is a statistical framework which can be used to quantify the uncertainty present in deep-learning models. The concept is based on Bayes' theorem, which links the posterior distribution of our observations to the prior probability of the hypothesis and the likelihood of the data given the hypothesis:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (4.1)$$

where: $P(A|B)$ is the posterior probability of hypothesis A given the observed data B, $P(B|A)$ is the likelihood of the observed data B given hypothesis A, $P(A)$ is the prior probability of hypothesis A and $P(B)$ is the marginal likelihood of the observed data B.

Uncertainty can be quantified in the Bayesian framework, although exact Bayesian inference with deep neural networks is computationally intractable. It can be approximated by Bayesian Neural Networks (BNNs), which are neural networks with probabilistic weights and biases. In the BNN framework, the weights and biases of a network are treated as random variables with a prior distribution over their values. Bayesian inference is then used in the process of training the network

to update the prior distribution with new data. The posterior distribution over weights and biases obtained during network training can be used to compute the predictive distribution of the model. This predictive distribution of the model is a representation of the range of possible model outputs given an input. The uncertainty of the model can be inferred from the model’s predictive distribution and can be a measure of the model’s confidence in the predictions. An alternative approach to Bayesian neural networks is Bayesian model averaging, in which predictions of multiple models are averaged. In practice, several methods of applying Bayesian inference to uncertainty estimation in deep learning have been proposed. Those include Markov chain Monte Carlo [124], variational inference [125], Monte Carlo dropout (MC dropout) [122] and deep ensembles [126].

Predictive uncertainty can be decomposed into two components: aleatoric, which captures inherent randomness in the data (such as noise in the labels), and epistemic, which represents the uncertainty of the model (e.g. when making predictions on out-of-distribution samples) [127]. Kendall et al. [123] proposed an uncertainty decomposition method using a neural network that outputs two values representing the mean and the variance of the prediction and using Monte Carlo dropout to approximate Bayesian inference. A similar approach was used by Kwon et al. [128] who proposed a new method of estimating aleatoric and epistemic uncertainties using Monte Carlo dropout without the need of any additional parameters. A recent study by Fort et al. has shown that deep learning ensembles with random initialisations allow for a much more robust approach to uncertainty estimation than methods applied to a single predictive model [129]. This is attributed to the fact that deep ensembles explore entirely different modes, while methods which are based on exploring the subspace of a single model (such as MC dropout) converge to a single mode.

Following [126], here we used deep ensembles as an alternative to Bayesian inference. In our experiments, all the models were trained for the same duration of 30 epochs. We trained separate ensembles of $M = 10$ models for each of the studied cohorts and an additional one for the combined cohort (CALM + PREV +

UK-AIH). Overall 12 ensembles of $M = 10$ models each were trained. Slides from the HepaT1ca cohort were retained as a separate validation set.

Given an ensemble Ω of M trained models $\Omega = \{\omega_m\}_{m=1}^M$, we computed the pixel-wise collagen likelihood as the average $\hat{p}$ of the ensemble predictions p_m for an input pixel i .

$$p_m = p(y|i, \omega_m) \quad (4.2)$$

$$\hat{p} = \frac{1}{M} \sum_{m=1}^M p_m \quad (4.3)$$

We then used the method proposed by [128] to estimate the pixel-wise predictive uncertainty for an input pixel i and the output prediction y as the variance of y given i over the predictive distribution $p(y|i, \Omega)$:

$$\text{Var}_{p(y|i, \Omega)}(y) = \underbrace{\frac{1}{M} \sum_{m=1}^M [p_m(1 - p_m)]}_{\text{aleatoric}} + \underbrace{\frac{1}{M} \sum_{m=1}^M (p_m - \hat{p})^2}_{\text{epistemic}} \quad (4.4)$$

Two approaches for computing collagen percentage area in a tile were employed; the first made use of the mean ensemble prediction P and the other of the voted ensemble segmentation B . The mean collagen prediction P for image I with dimensions W, H is given by the equation:

$$P = \frac{1}{WH} \sum_{x=1}^W \sum_{y=1}^H \hat{p}(x, y) \quad (4.5)$$

Collagen segmentation value $\hat{b}$ was obtained for each pixel via thresholding of the corresponding ensemble prediction at collagen likelihood equal or higher than 0.5. Let $\hat{p}(x, y)$ be the ensemble prediction value for pixel i at location (x, y) in the image I , and let $\hat{b}(x, y)$ be the corresponding binary value after thresholding. Then $\hat{b}(x, y)$ can be defined as follows:

$$\hat{b}(x, y) = \begin{cases} 1, & \text{if } \hat{p}(x, y) \geq 0.5 \\ 0, & \text{otherwise} \end{cases} \quad (4.6)$$

And the segmented collagen area proportion B is given by:

$$B = \frac{1}{WH} \sum_{x=1}^W \sum_{y=1}^H \hat{b}(x, y) \quad (4.7)$$

The uncertainty in the mean tile collagen prediction is reported as the interquartile range (IQR) of the ensemble collagen predictions P . Similarly, the uncertainty of the segmented collagen area in a tile is reported as the IQR of the collagen segmentation maps produces by thresholding of the predictions by each of the models in the ensemble.

4.5 Results

The validation metrics for the trained segmentation models are summarised in Table 4.3. Here the scores are provided separately for ensembles trained and validated on the individual study cohorts and for the ensemble models trained on the pooled cohort (CALM + PREV + UK-AIH).

Table 4.3: Segmentation network Dice scores on validation tiles ($M = 10$ ensemble average $\pm$ standard deviation).

Training Set	CALM	PREV	UK-AIH	HepaT1ca
Individual Cohorts				
U-net Tiny	0.847 ± 0.022	0.877 ± 0.012	0.793 ± 0.036	-
U-net Mini	0.905 ± 0.026	0.882 ± 0.005	0.852 ± 0.019	-
Att U-net	0.820 ± 0.022	0.825 ± 0.059	0.759 ± 0.070	-
Pooled Cohort				
U-net Tiny	0.823 ± 0.021	0.861 ± 0.013	0.767 ± 0.037	0.594 ± 0.075
U-net Mini	0.834 ± 0.029	0.864 ± 0.014	0.795 ± 0.035	0.555 ± 0.104
Att U-net	0.785 ± 0.036	0.834 ± 0.017	0.623 ± 0.067	0.539 ± 0.123

The U-net Mini has achieved the highest validation metric scores for all of the assigned tasks. It can be seen that for all model architectures there is a drop in Dice scores between the ensembles trained on individual cohorts and those trained on the pooled cohort. The higher Dice scores for individual cohorts can be explained by the models over-fitting to the data within each of the cohorts.

It can be seen that while very high Dice scores are achieved for the CALM and PREV cohorts, performance of each of the network architectures is slightly worse for the UK-AIH cohort. This can be explained by the fact that the UK-AIH samples were collected retrospectively from a large number of hospitals and are characterised by frequent staining artefacts and large intra-cohort differences in slide colour. As can be expected, the performance of the segmentation models is significantly lower for the unseen cohort (HepaT1ca).

Figure 4.3 shows model predictions for an example tile from the CALM cohort. It can be seen that the human annotation used as ground truth for model validation contains noisy false positive pixels around the annotated fibres. These noisy labels can be attributed to falsely positive labelling by the annotator (MW) of the JPEG compression-induced noise in the original WSI. All prediction maps generated by the tested U-net models provide a realistic representation of the collagen pattern in the original image. It can be seen that neither of the trained segmentation models has replicated the image compression noise in the final prediction, which is a desirable outcome. It is difficult to notice any visual differences in the quality of the prediction maps output by each of the models. This shows that each of the tested networks is appropriate for the task of collagen segmentation.

4.5.1 Uncertainty computation

Fig. 4.4 shows the results of model predictions and uncertainty maps for the U-net Mini ensemble trained on the pooled cohort. It can be seen that almost all of the aleatoric uncertainty is concentrated on the boundaries of collagen fibres. This is expected, as boundary pixels are commonly affected by partial volume effects. Because of this phenomenon, the aleatoric uncertainty is directly tied to the length of the collagen outline in the tile. We observe that tiles with little to no collagen are characterised by uncertainty values near to 0 and tiles with much collagen have high inherent collagen area uncertainty.

Table 4.4 presents the mean ensemble uncertainty scores obtained from model predictions on the validation set. It can be seen that the values of epistemic

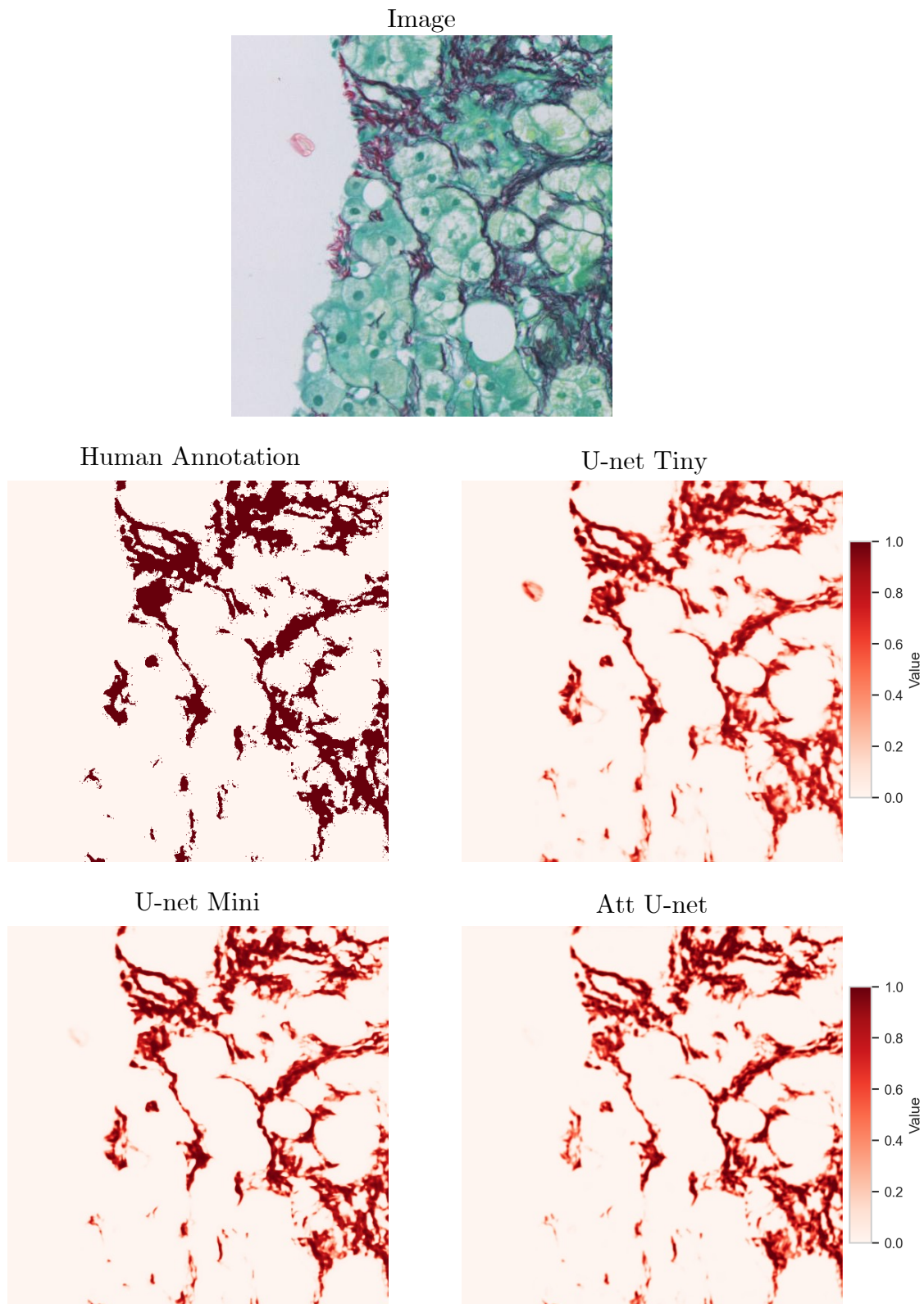


Figure 4.3: Collagen prediction maps output by the tested segmentation U-net models. The human annotation used as ground truth for model validation contains noisy false positive pixels around the annotated fibres. Each prediction map shows an output prediction from a single model trained on the pooled cohort. It is difficult to see any differences in the quality of the collagen maps predicted by different model types.

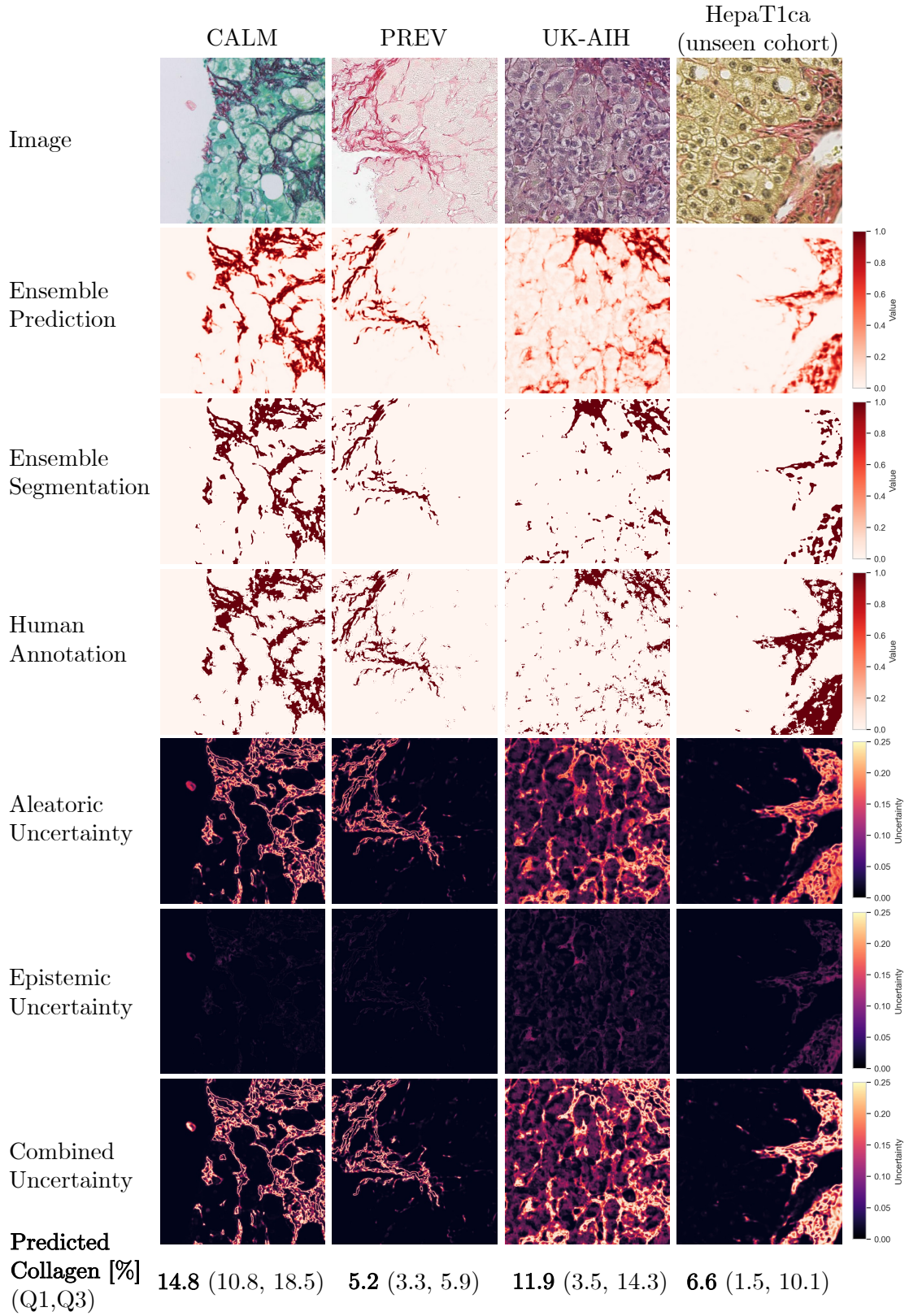


Figure 4.4: The results of U-net Mini ensemble model prediction with uncertainty estimations for examples of unseen tiles from each of the cohorts. Tiles with low contrast of stained collagen and unseen colours are characterised by high ensemble model uncertainty.

Table 4.4: Ensemble uncertainty scores on the validation set for each of the trained model types ($M = 10$ ensemble average $\pm$ standard deviation).

Model ensemble	U Aleatoric	U Epistemic	U Combined
U-net Tiny	0.020 ± 0.025	0.002 ± 0.003	0.022 ± 0.027
U-net Mini	0.017 ± 0.022	0.002 ± 0.003	0.019 ± 0.025
Att U-net	0.024 ± 0.036	0.003 ± 0.006	0.027 ± 0.041

uncertainty are very low compared to the values of aleatoric uncertainty for all of the image ensembles. This shows that it is the characteristics of the analysed image which drive the uncertainty of the collagen area measurement. The variability in measured collagen percentage due to differences in predictions by individual segmentation models within each ensemble is relatively very small.

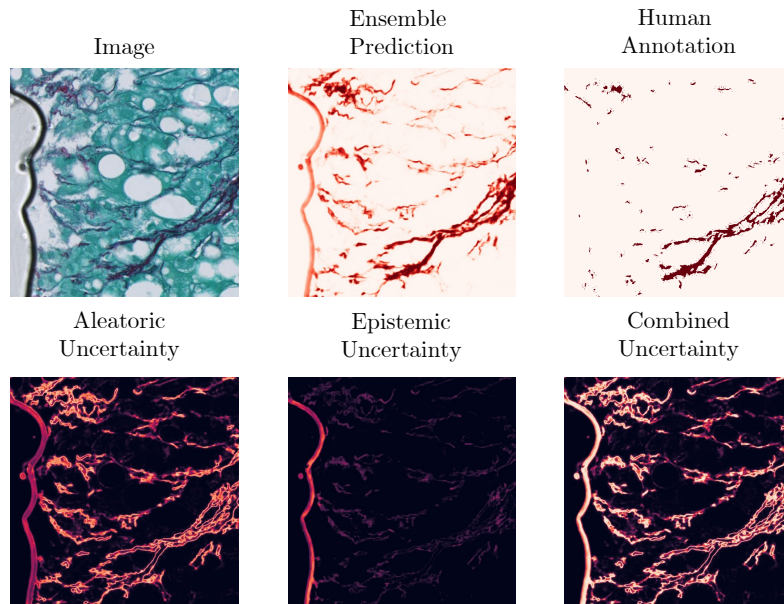
**Figure 4.5: An example image prediction tile with high epistemic uncertainty.** The air bubble visible in the original tile is highlighted in the epistemic uncertainty map, which visualises out of distribution data.

Figure 4.5 shows ensemble model prediction and uncertainties for a tile with a common imaging artefact, an air bubble. The bubble can be clearly seen in the epistemic uncertainty map, meaning that some of the models in the ensemble have classified its pixels as regions of collagen, and some did not. The visualisation of epistemic uncertainty can provide the pathologist with a map of regions in

the slide which were not represented in the model training set, and therefore are likely to contain prediction errors.

4.5.2 Interpreting uncertainty

The results of uncertainty scores computed for each of the cohorts are summarised in Table 4.5.

Table 4.5: Ensemble prediction uncertainty scores for validation samples from each of the cohorts, U-net Mini (M = 10 ensemble average $\pm$ standard deviation).

Training Set	CALM	PREV	UK-AIH	HepaT1ca
Individual Cohorts				
U Aleatoric	0.016 ± 0.016	0.006 ± 0.008	0.024 ± 0.020	-
U Epistemic	0.002 ± 0.003	0.001 ± 0.003	0.003 ± 0.005	-
Pooled Cohort				
U Aleatoric	0.026 ± 0.028	0.008 ± 0.011	0.044 ± 0.033	0.018 ± 0.022
U Epistemic	0.002 ± 0.002	0.001 ± 0.001	0.001 ± 0.005	0.002 ± 0.004

Figures 4.6 and 4.7 show the correlation statistics for computed collagen areas (the average prediction P and the segmented collagen area B), as well as uncertainties for the models trained on the individual cohorts and the model trained on the pooled cohort, respectively.

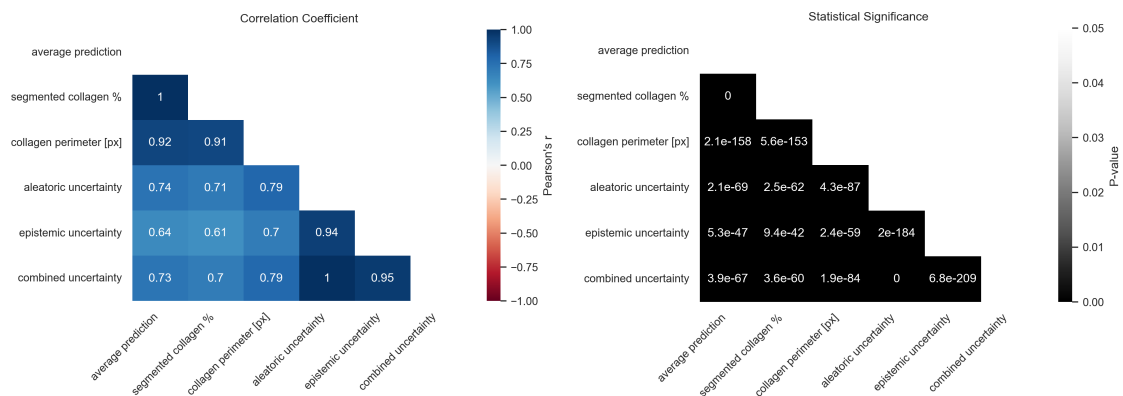


Figure 4.6: Correlation coefficients and corresponding statistical significance values for the prediction parameters of the model trained on the individual cohorts. The values are reported for predictions obtained using the U-net Mini ensemble.

The PREV cohort has the lowest aleatoric uncertainty values of the cohorts studied. It can be seen that aleatoric uncertainty is highly correlated to the area of

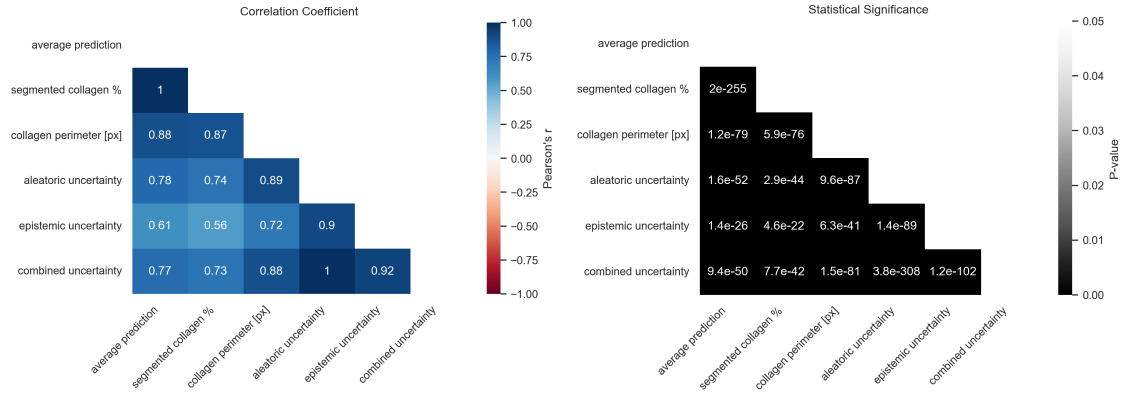


Figure 4.7: Correlation coefficients and corresponding statistical significance values for the prediction parameters of the model trained on the pooled cohort. The values are reported for predictions obtained using the U-net Mini ensemble.

collagen in an image and therefore the lower uncertainty in the PREV cohort can be partly explained by the lower content of collagen in this cohort overall (see Table 4.1).

Interestingly, after cohort pooling the mean aleatoric uncertainty scores increase for all the models, especially in the cohorts where more than one stain colour was used (i.e. CALM and UK-AIH). This indicates that the segmentation models introduced to different variants of collagen staining have lower confidence at the collagen boundaries where more than one stain is present. Such interpretation would correspond to the intuition that placing additional stains on the tissue specimen multiplies the noise in the data. The highest aleatoric uncertainty values are reported for the UK-AIH cohort in which the purple haematoxylin staining was used in the PSR slides. Collagen tiles in this cohort have overall a lower visual contrast than those from other cohorts.

When comparing the correlation coefficients between the ensembles trained on single cohorts and those trained on the pooled cohort, it can be seen that the correlation between the length of the collagen perimeter and the aleatoric uncertainty has increased after cohort pooling. At the same time, the correlation between the area of the obtained segmentation and the epistemic uncertainty has decreased.

It can be seen that the correlation between the the average prediction P and the segmented collagen B is equal to 1, meaning that these measures can be considered equivalent to each other. This means that it may be acceptable to

present pathologists with the collagen prediction maps rather than the binarised segmentation maps. The prediction maps have a closer resemblance to the original input image as they capture the colour intensity gradients characteristic of collagen fibres of different thicknesses. Collagen prediction maps are also visually more similar to the images acquired using SHG/TPEF microscopy, where similarly, an inferred collagen density map rather than a binary segmentation map is obtained [78]. Such continuous collagen prediction maps could be interpreted as a kind of "ideal" stain deconvolution, where only pixels containing any collagen can have positive values, and regions of tissue which have no collagen (i.e. cell cytoplasm, cell nuclei) have a collagen probability of 0.

4.5.3 Processing time

Table 4.6 shows a comparison of the training and inference times for each of the tested models. The times are reported for the processing of the training set (training over 30 epochs) and of the validation set, respectively. It can be seen that tile processing time is strongly related to the size of the network used, with U-net Tiny outperforming other models in our study. Processing time should be an important factor to consider when choosing deep learning models for use in clinical practice.

Table 4.6: Training and inference time comparison for the trained models (M = 10 ensemble average $\pm$ standard deviation).

Model	Training [s]	Inference [s]
U-net Tiny	417.7 $\pm$ 16.0	1.3 $\pm$ 0.2
U-net Mini	1358.8 $\pm$ 24.5	2.5 $\pm$ 0.6
Att U-net	1590.4 $\pm$ 37.0	3.3 $\pm$ 0.8

4.6 Discussion

The problem of accurately performing histological quantification and inference from differently stained slides is a known issue, to which several approaches can be found within the digital pathology literature. The first of them is the concept of staining normalisation.

The first attempts to normalise stain intensity between histology slides relied on the process of stain deconvolution. Stain deconvolution is the transformation of a tissue image from the RGB colourspace into a representation in which each image channel encodes the proportion of a particular histological stain [108]. Stain deconvolution can be used directly for stain content quantification in each pixel, in stains which conform to the Beer-Lambert law, i.e. those in which the relation between the stain concentration and light absorbance is linear.

To perform stain deconvolution (SD) on a digital slide, one needs to know the colour fingerprints of the particular pure stains, represented as stain vectors in the RGB colourspace. A stain vector can be measured directly for a given combination of stain and tissue scanner or alternatively, it can be estimated from a stained digital slide. Such an estimated stain vector can then be applied to deconvolve the stain channels of a histology image and to perform a colour normalisation between slides [130]. Deconvolution-based colour normalisation has been successfully used to correct for minor stain representation differences between WSI scanners from different vendors [110, 131].

However, there are several known limitations of applying stain deconvolution to whole-slide images. Firstly, WSI scanners typically make use of poly-chromatic illumination, as opposed to white light. The assumption of SD is that the light is white, which can lead to significant deconvolution errors [132]. Furthermore, automated stain vector estimation is known to fail in slides which contain only very small areas stained with a particular stain.

From the SD results presented in our study, it can be seen that stain-deconvolved slides (Figure 3.7) still need to be further processed before they can be used for collagen quantification. Especially regions of non-specific stain binding in the tissue (e.g. cell nuclei) need to be identified and excluded. This means that the computed stain distribution map needs to be either binarised using a stain intensity threshold or segmented using a deep-learning approach. Performing colour deconvolution adds an additional step into the analysis, which needs to be justified given the high performance of the state-of-the-art deep-learning models. Our results

demonstrate that even a relatively simple deep learning network, such as the U-net, can compensate for the use of different scanners and even major differences in sample colour, as long as sufficiently varied data is used for model training.

Independently from stain deconvolution, the concept of colour-transfer has been developed. The first methods of automatic colour transfer between images were developed in the early 21st century using simple statistical approaches, such as image histogram transfer [133]. Subsequently, deep-learning based methods were developed to achieve the same goal. Sparse autoencoders have been successfully applied to normalise colours between H&E slides without the need for estimating a stain matrix [134]. Finally, the introduction of generative autoencoders and the CycleGAN model [135] in particular has opened a new direction for digital pathology research.

Autoencoders are a type of artificial neural network which can learn a compressed representation of an input such as an image in an unsupervised manner. An autoencoder is built of two main parts; an encoder which encodes the compressed data representation and a decoder which reconstructs the original input. The whole autoencoder is trained using a process called backpropagation in which the network aims to minimise the difference between the input and the output.

Generative autoencoders are a special type of autoencoders which have the ability to generate new data similar to the received input data. In addition to an encoder and a decoder, these networks additionally contain a module called a generator, which accepts random noise as an input and is trained to generate new data which resembles the original input.

Generative autoencoders have been successfully applied to normalising between different variants of H&E staining and different scanners [85]. Similar models have enabled the generation of realistically looking coloured images from unstained tissue slides [136]. The concept has been further extended to predict the distribution of specialist stains from the basic H&E-stained slide [137].

Most recent deep-learning models are capable of creating very realistic looking colour-transformations between different stainings [138]. However, it doesn't mean that such simulated images carry biologically meaningful and clinically usable

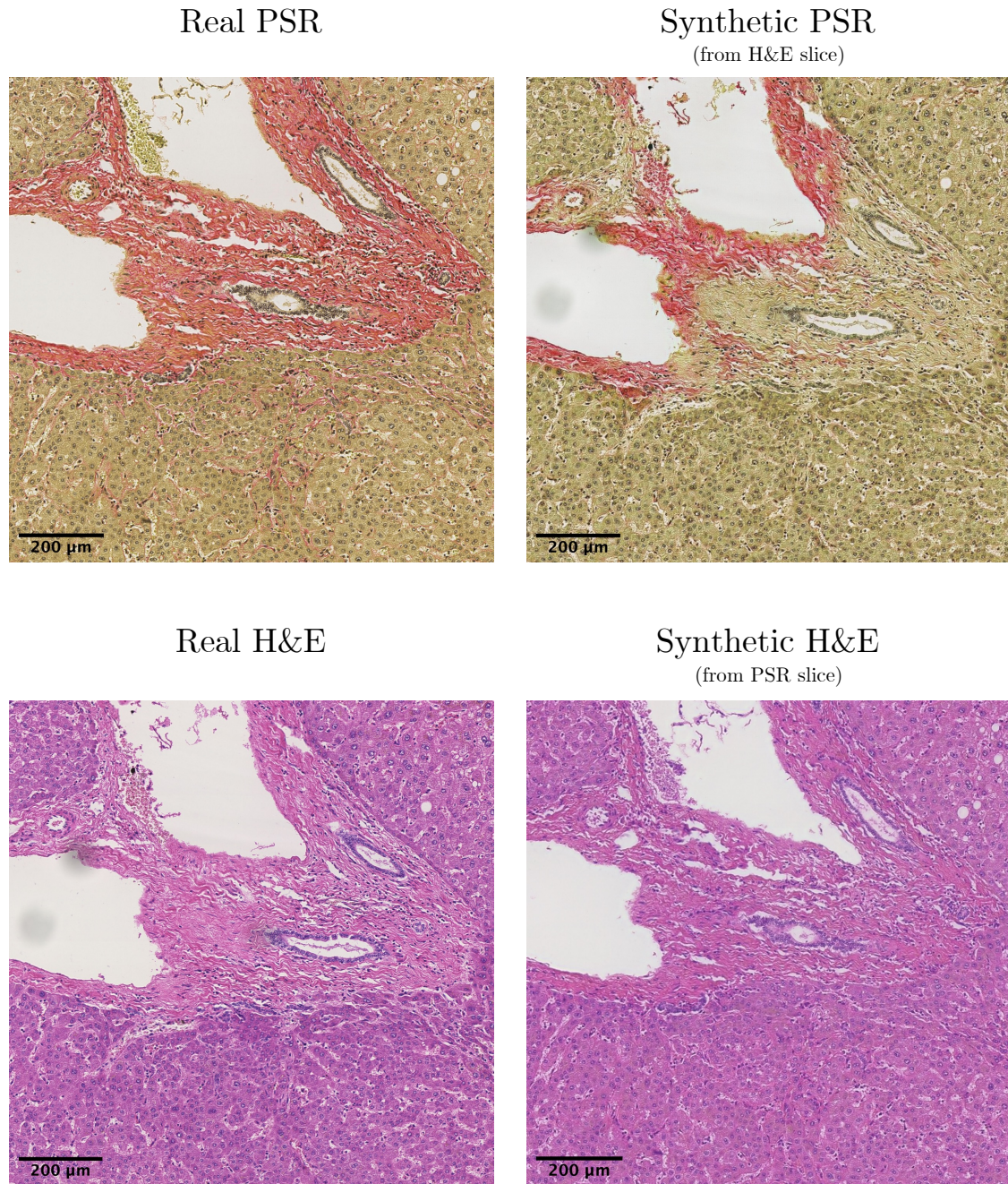


Figure 4.8: An example of stain transfer results between H&E and PSR stainings (HepaT1ca cohort). The scene shows a large bifurcating portal tract in two adjacent tissue slices stained with PSR and H&E, respectively. In the synthetic PSR image, while the shapes and textures of the tissue objects look highly realistic, the colour distribution is very different from that of the real PSR slide. A substantial part of the portal tract is coloured green rather than red, and fine collagen fibres within the parenchyma are not captured. Two pieces of debris (grey circles on the left side of the image) have been replicated from the original H&E slice. The synthetic H&E image looks highly realistic.

information. Figure 4.8 shows an example of stain transfer performed between H&E and PSR sections from the HepaT1ca cohort, using the CycleGAN model [135]².

H&E staining does not create a differential contrast between collagen fibres and other biological tissue objects as neither haematoxylin nor eosin binding is specific to collagen. The use of special stainings in pathology workflows enables the pathologist to examine the distribution of substances (such as glycogen, tissue iron or viral antigens) which cannot be seen in the basic H&E staining. While histological deep fake images may appear realistic to an untrained eye, they cannot be used to infer information which isn't embedded in the source image.

4.6.1 Important nuances in reporting of CPA values

CPA is defined as the proportion of the detected collagen area to the area of the detected tissue in a whole-slide image. While this concept is very intuitive, different definitions of both collagen area and tissue area exist in digital pathology literature, rendering it difficult to perform any direct comparisons of CPA values obtained within different studies.

While some studies report collagen area as the total area of stained collagen in the slide, others have reported it as the area of collagen regions which the annotating pathologists have identified as abnormal [23]. The variation in scenes (i.e. the presence or the absence of large portal tracts, central veins and the liver capsule) between individual biopsies means that the sum of collagen pixels in a slide does not reliably predict fibrosis stage especially in the early stages of fibrosis. The act of excluding of regions of "normal" collagen (i.e. the walls of portal tracts and large central veins) is an attempt at manually correcting the computed CPA value for the inherent sampling variation of the biopsy. While the motivation behind performing such annotations is clear, they introduce observer subjectivity into the analysis.

Even in the simple case where CPA is reported as the sum of all collagen pixels in the slide, the results of collagen quantification via thresholding are affected by high inter-observer variability [1]. Furthermore, while performing such manual

²Further details on the implementation of the stain transfer pipeline are provided in Appendix B.

annotations may be viable in a small study, it requires additional work from the pathologist which may not be desirable in everyday clinical practice.

The same complication exists in the definition of tissue area which can either be defined as the area of the biopsy including or excluding the holes present in every liver specimen (i.e. lumens of portal tracts and central veins, and/or emptied fat vesicles). We observe that the computation of tissue area by the second definition is more complicated as it opens additional questions around the precise size and nature of the holes which need to be included or excluded. Secondly, as was demonstrated in this chapter, long object boundaries introduce high aleatoric uncertainty. Excluding holes may add to the uncertainty of the quantified tissue area especially in cases which are highly steatotic, as the outline of each fat bubble will add to the length of the boundary between tissue and background.

Finally, we note that it is essential that each result of CPA quantification is accompanied with information about the magnification at which the WSI was acquired, due to the fractal nature of collagen shapes. Obtaining whole slide images at the highest magnification (40x) takes longer and results in larger output file sizes than for lower magnifications (e.g. 20x). The large size of whole slide images is a challenge not only from the perspective of cost of storing digital slides but also the cost of the required computational resources. While the lowest viable magnification for digitisation of PSR slides is not known, a recent clinical study on the subject has concluded that WSIs acquired at a resolution below 4x do not adequately capture the thin fibres characteristic of pericellular fibrosis [20].

4.7 Conclusions

The differences in CPA quantification results between pathologists were found to be associated with the variability of collagen staining, and high subjectivity of the readers. However, we postulate that the precision of quantification of CPA from WSIs is primarily limited by the fractal-like shape of collagen fibres, very long object boundaries, partial-volume effects in the z-axis, and the imaging resolution of the digitised slides.

In this paper we have successfully applied a commonly used semantic segmentation model, the U-net, for identifying tissue collagen in a large multi-centre patient cohort. The data in our study is unique in its scope, comprising four fully separate groups of participants. In each of those groups samples were preprocessed according to different staining protocols, resulting in entirely different colours of the analysed tissue specimens. While for the purpose of this study, the slides from the different cohorts are here treated as equivalent in their staining, it cannot be assumed that that is actually the case.

Here a framework based on Bayesian inference is applied to estimate the uncertainty in the predicted collagen map and the quantified collagen area. The method is applied to the digitised slides to obtain collagen uncertainty estimates at the levels of individual pixels and tiles. The values obtained through this method likely underestimate the real uncertainty of the whole collagen percentage measurement process, which is additionally affected by errors in each of the prior steps of tissue processing.

While the presented solution is limited in its scope, the examples provided still show its utility. The results of our uncertainty estimates are broadly in line with those from the previously published literature [128]. The separation of the segmentation uncertainty into its aleatoric and epistemic components allows to further split the visualisation of segmentation errors into those which come from quality of the analysed image and those which come from the quality of the applied model. These uncertainty maps can be used to provide the pathologist with a direct interpretation of the model performance on different parts of the sample. In particular, epistemic uncertainty maps can be used to visualise out of distribution data (e.g. artefacts and unseen colours). In future use, such metrics can be used to verify the performance of the segmentation model on new cohorts and indicate that additional training may be necessary to obtain reliable collagen segmentation results.

Previous work on stain transfer has enabled the generation of highly realistic histology images of different staining colours and patterns. However, innovations like these have not advanced our ability to quantify stain-positive regions in real

clinical data, as simulated stains do not reflect the real biological state of the tissue. Furthermore, while alternative segmentation models could be applied, here we demonstrate that the use of a standard U-net is appropriate.

We demonstrate that providing an estimate of the measurement error is critical. Without such error bounds the resulting quantitative results become meaningless. Great care should be applied when using such estimates for diagnostic purposes. In situations where CPA is measured from human annotations of collagen, the measurement error is estimated through the quantification of disagreement between the individual annotators. Here we propose that an ensemble of deep-learning segmentation models can be used to create a set of likely annotations and thus model the distribution of the possible quantification results. We argue that similar uncertainty estimations should be applied to all stain quantification tasks in modern digital pathology, such as the quantification of areas stained with immunohistochemistry (IHC) stains. Finally, we demonstrate that as long as differing definitions of collagen area and tissue area are used, CPA results will not be directly comparable between studies.

5

Automatic Staging of Liver Fibrosis in MAFLD

In the context of chronic diseases, the development of fibrosis refers to the deposition of excess collagen in tissue. Fibrosis can lead to organ dysfunction and even to organ failure. Typically, histochemical stains are being used to visualise collagen in whole slide images of tissue slides.

Assessing the level of fibrosis plays an important role in the diagnosis of liver disease in general and Metabolic-Associated Fatty Liver Disease in particular. MAFLD is a progressive liver disease characterised by deposition of excess fat in liver cells, inflammation and fibrosis, and has strong links to diabetes, obesity, and cardiovascular diseases.

While the measurement of the Collagen Proportionate Area (CPA) [139] provides a very basic assessment of the collagen content per unit area, it does not describe the topological changes in liver tissue architecture which form the basis of pathology staging systems. It has already been shown that the assessment of morphometrical features of collagen fibres allows for a more accurate histological staging of fibrosis than on the basis of CPA alone [74]. Here, we propose to utilise deep learning to identify collagen deposition patterns that form the basis of a more objective

and standardised assessment of the fibrosis progression in liver biopsies as defined by the METAVIR standard (see Figure 5.3).

Our proposed method utilises three steps. Firstly, it relies on collagen segmentation from the whole slide image to remove the influence of stain gradients and artefacts on the final fibrosis stage prediction. Secondly, we build a tissue-graph which connects the tissue tiles in a way which resembles the natural microscopic-level structure of the liver. Finally, we apply a graph-neural network to predict the overall fibrosis stage of the sample based on the features embedded in the tissue graph.

5.1 Previous work

Whole Slide Images (WSIs) are typically very large and therefore cannot be directly input into an image classifier, such as a convolutional neural network due to computer memory limitations. A common workaround for processing of histology slides relies on splitting the WSI into smaller square patches and processing the whole slide on a tile-by-tile basis. The tile-level representations obtained from deep-learning models can later be aggregated to predict the label of the entire slide. Such an approach has been used in diseases where information present at the cellular-level is important to form a diagnosis - e.g. in cancer. The downside of the patch-based methods is that such representations typically do not take into account the tissue topology at the macroscopic scale.

The assessment of fibrosis, however, is unique in that the staging systems describe changes (bridging) which occur at a macroscopic scale as opposed to the scale of individual cells [140]. In the case of liver disease, a single fibrotic bridge might even span the width of a typical biopsy. For this reason, the most recent approaches use either downsampled versions of the whole WSI [75] or very large patches of tissue [71] to obtain a slide-level fibrosis classification. Downsampling of a collagen WSI results necessarily in loss of detail at the pixel level. In turn, dividing tissue samples into large tiles which show much spatial context may be impractical in small clinical needle biopsies.

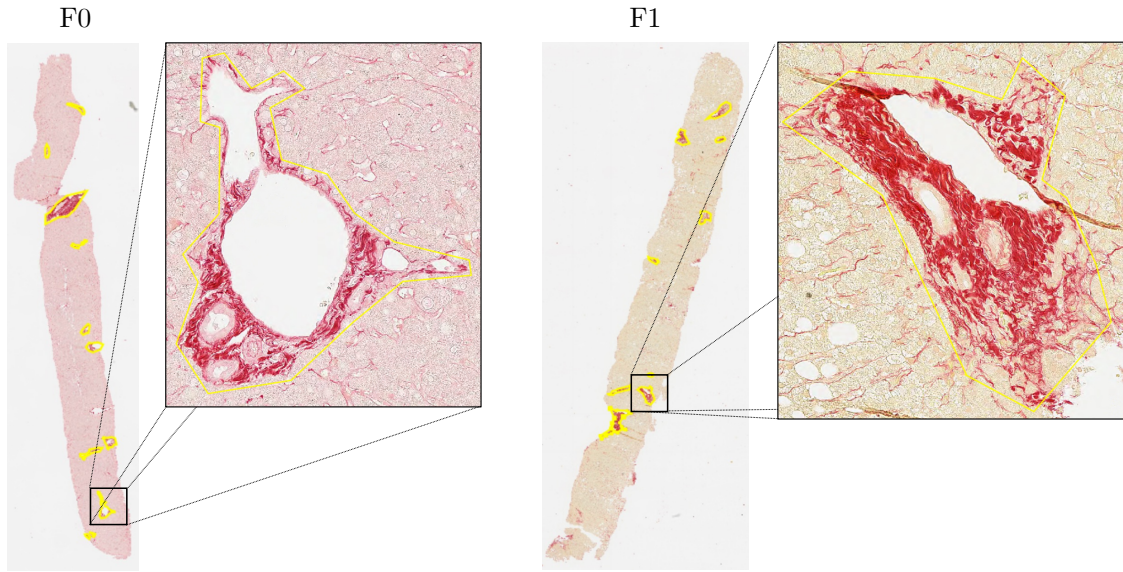
Xu et al. developed an alternative approach to collagen slide classification into fibrosis stages using handcrafted morphological features of collagen fibres in a fully-magnified slide. In their approach, the segmented collagen fibres were divided into local subtypes based on their textural patterns and the proximity to the nearest architectural landmark [78]. Such an approach allowed to output image classification maps which helped to interpret the results of fibrosis stage classification.

Pathology staging systems for fibrosis describe large-scale topological changes in liver architecture. In these systems portal tracts and central veins are treated as anatomical landmarks which can become connected with fibrotic bridges. Liver cirrhosis can be seen as a state in which all portal tracts lose their individual boundaries and become incorporated into the network of bridges (fibrotic septa). A pathologist examines individual tissue landmark regions and interprets and classifies them into their different types. However, such an examination is highly subjective largely due to the destruction of the assessed collagen objects during tissue slicing.

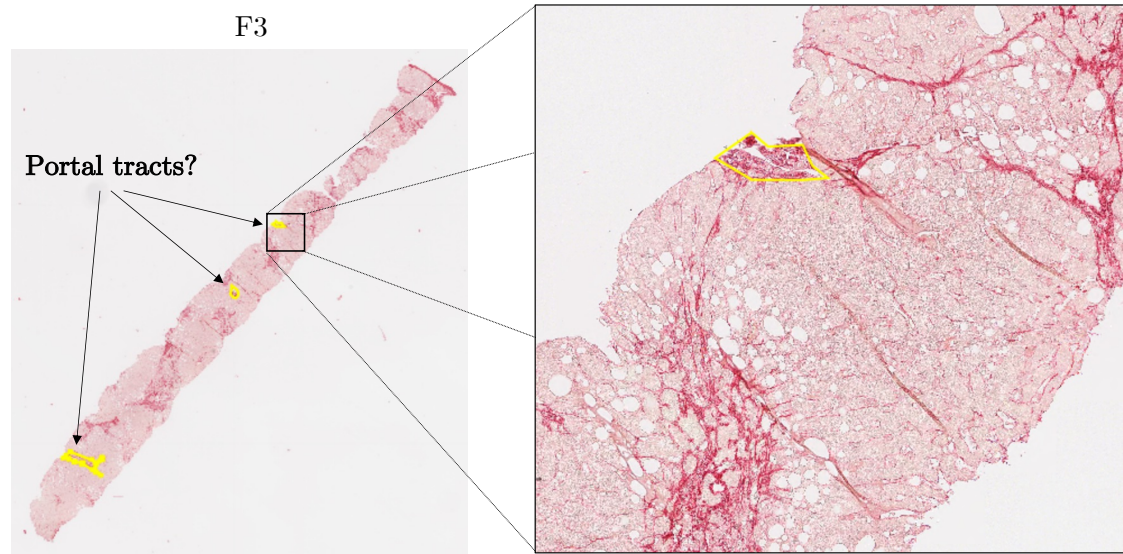
Figure 5.1 shows the challenge of identifying and classifying discrete objects such as portal tracts and central veins in needle liver biopsies. In a normal liver, as well as in the early stages of fibrosis (e.g. METAVIR F1), portal tracts and central veins can be identified in the stained tissue section with high confidence (Fig. 5.1a). However, as fibrosis progresses, individual portal tracts and central veins become more and more difficult to identify and delineate. Similarly, often it is difficult to discern if a long strand of thick collagen is a fibrotic bridge or a fragment of a longitudinally cut portal tract.

Dividing histological slides into small tiles results in further limiting of the visual context of the analysed image. As tile boundary allocation is random, anatomical objects often end up split between two or more neighbouring tiles, further complicating the task of classifying liver tissue objects (Fig. 5.2).

Graphs provide a unique way of describing spatial relationships between objects, and therefore have a long history of application to histology data [141]. More recently, Graph Convolutional Networks (GCNs) have been applied to histopathology. While conventional CNNs perform convolutions on 8-connected patches in each row of



(a) A normal liver (F0) is characterised by portal tracts with clearly visible definitive boundaries. Collagen structures such as portal tracts and central veins remain easily identifiable in the early stages of fibrosis (F1).



(b) Individual collagen objects such as portal tracts are often impossible to define in a liver with advanced (bridging) fibrosis (F3).

Figure 5.1: Visual characteristics of liver architectural landmarks in early and advanced fibrosis. Identified portal tracts are outlined in yellow. Annotations were drawn by MW in consultation with Prof. Kenneth Fleming.

an image, GCNs apply the convolutions to tile neighbourhoods defined by graph edges. Such an approach was recently used to create a database of regions of cancerous tissue with similar appearances [142].

Tissue graph nodes can also represent biological entities rather than image

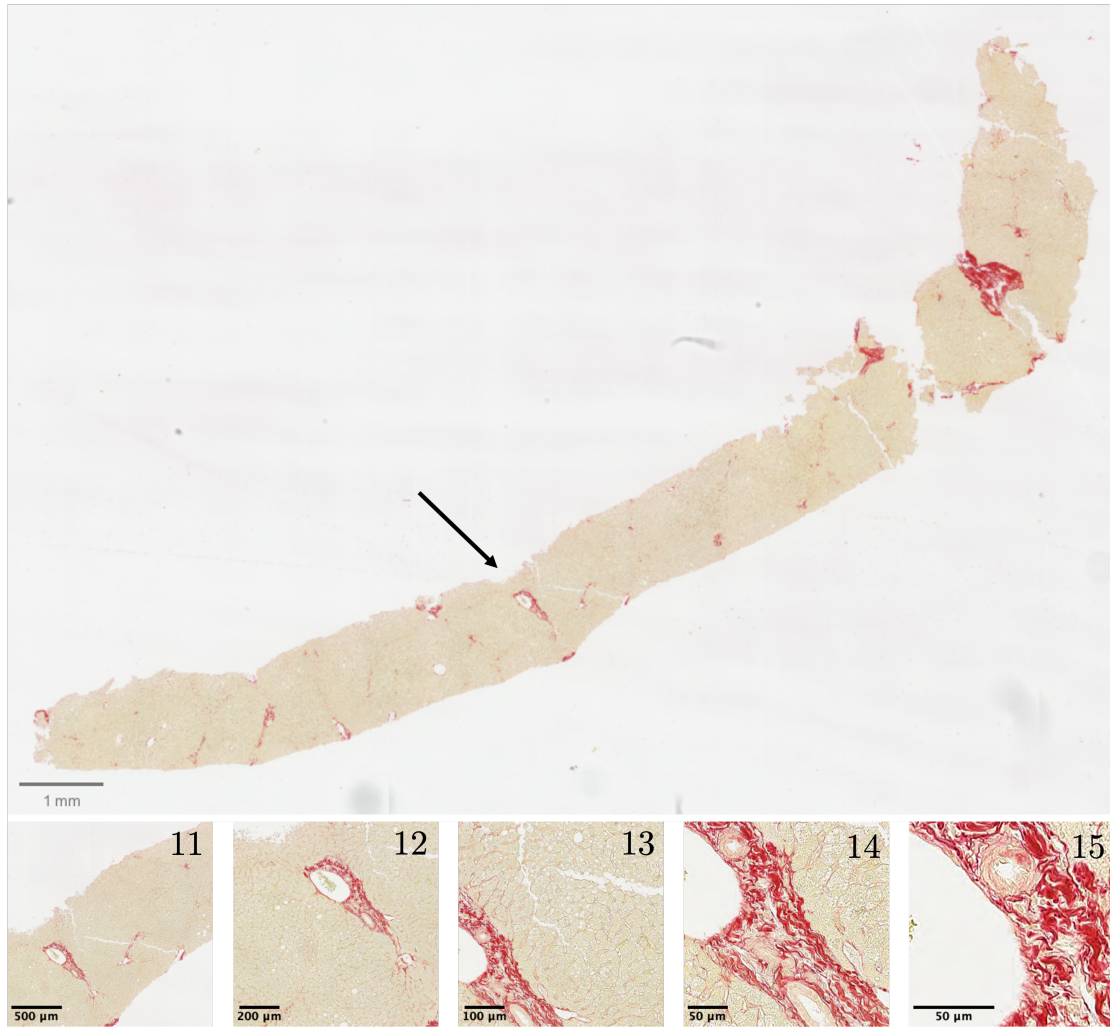


Figure 5.2: Visual context in PSR tissue tiles obtained at different WSI pyramid levels. Tiles from pyramid levels 11-15 corresponding to the area indicated by the arrow are shown. Tile size is fixed at 256 by 256 pixels at each of the levels.

tiles as is in the case of cell-graphs [143]. Furthermore, graph neural networks can perform convolution operations over neighbourhoods of differing sizes, which can be used to accommodate irregularly-shaped regions of tissue. Furthermore, graphs can be used to organise the tissue features in a hierarchical manner, such as in the case of HACT-Net, where embeddings of individual cell features were stacked with embeddings of tissue superpixels to perform breast cancer classification [144]. Here, we propose a novel tissue representation that combines the robustness of deep learning with the prior knowledge of tissue architecture by employing a graph-based model.

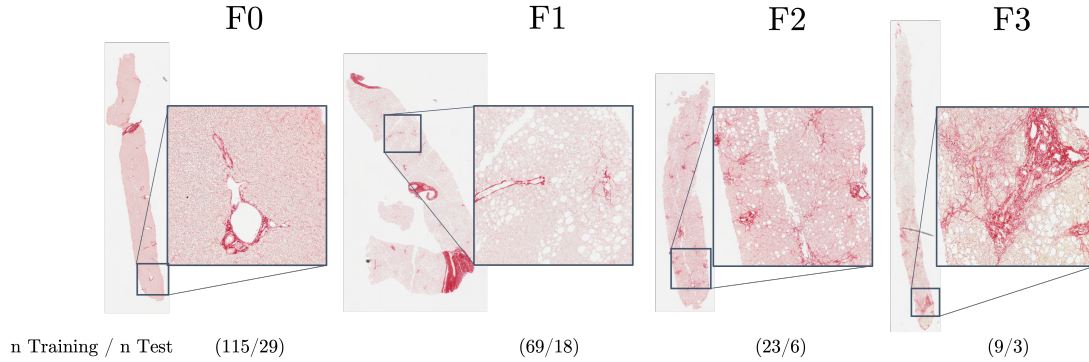


Figure 5.3: The distribution of training and testing samples across the liver fibrosis stages in the PREV cohort (METAVIR standard). Healthy liver (stage F0) shows little collagen (red stain), which is almost only present in portal tract walls. During the initial stages of fibrosis (F1-F2) a network of fine collagen strands appears in the parenchyma. In the higher stages (F3-F4) the pathological fibres form "bridges" connecting portal tracts and central veins, resulting in a cobweb-like macro-pattern. It should be noted that the highest stage (F4 - cirrhosis) is not present in the dataset.

5.2 Data

A set of 271 percutaneous liver biopsies from the PREV cohort of participants with known or suspected Metabolic Associated Fatty Liver Disease (MAFLD) was provided by Perspectum Ltd., Oxford, UK. The slides have been stained with PicroSirius Red (PSR), which dyes tissue collagen red. The images represented varying stages of fibrosis (Fig. 5.3). Each biopsy has been assigned a fibrosis stage by an expert pathologist in accordance to the METAVIR standard (F0-F4) [140]. The slides have been scanned with an Aperio scanner at 20x magnification. Here, the dataset has been divided into training and testing folds with a 80:20 ratio.

5.3 Methodology

The task of fibrosis detection can be modelled as a multi-class classification problem. Each slide is decomposed into a set of tiles and one of C classes is assigned to the whole sample. The classes are derived directly from the pathology staging system. In this work, we propose a novel method of tissue graph construction based on the identification of liver tissue landmarks (portal tracts, central veins and/or regions of fibrosis) and partitioning of the slide into regions centred around said landmarks.

Firstly, all slides are pre-processed by segmenting the tissue collagen at full image resolution. Next, tile clustering is performed and the results of this clustering are used to find fibrosis landmarks in the tissue (Fig. 5.4).

Collagen segmentation

A PSR stained tissue slide contains visual features which are not relevant to fibrosis pattern classification, such as empty fat bubbles (steatosis) and tissue edges. In order to remove the influence of such features on the output of slide classification, pixel-wise collagen segmentation was applied to the liver WSI at full imaging resolution. Here a U-net Mini model was used for collagen segmentation. The training and validation of the collagen segmentation model for the PREV cohort is described in detail in Chapter 4.

Landmark identification

After segmentation, the collagen maps of the full-sized tissue slides were divided into 256 by 256 px tiles at the highest magnification of the WSI (See Fig. 5.2). Such tiles correspond to 127 by 127 micron regions of tissue in the full 20x slide resolution with pixel x and y dimensions of 0.4954 microns. Such small size of tiles was selected to enable the characterisation of local textural patterns of collagen in the tissue slide via a heatmap (Fig. 5.4).

Features from small tiles (256 x 256 px) are extracted from the full resolution collagen segmentation maps using a ResNet18 model pretrained on the ImageNet dataset. The resulting feature vectors $\in \mathbb{R}^{512}$, with each vector representing a single collagen tile were clustered using a k-means clustering algorithm. About 10% of tiles randomly sampled from the whole dataset were used to learn the fibre cluster representation. Experimenting on different numbers of clusters, we found that setting $k = 4$ resulted in easily interpretable clusters containing tiles with increasing amounts of collagen. In particular, here tiles from cluster 3 contained regions of dense collagen: liver vasculature and regions of fibrosis.

Finally, the learned tile clusters were applied to perform a tile classification in the whole tissue slide (a collagen heatmap). The purpose of producing such a

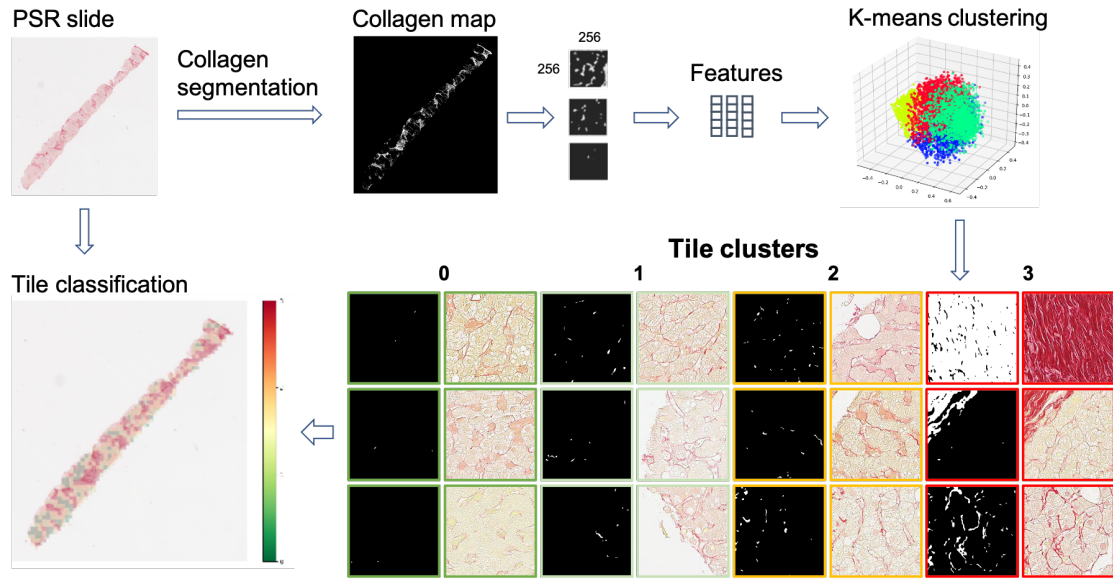


Figure 5.4: Tile subtyping pipeline. 4 clusters of collagen tiles have been identified, each representing a different level of collagen content. The classified tiles are coloured according to their predicted cluster. Regions with high collagen content (cluster 3) are later used as central nodes for Voronoi tessellation of tissue tiles. The tile size of 256x256 px was specifically chosen to capture the local pattern of the collagen fibres and at the same time to limit the field of view to only contain one type of collagen pattern at a time.

heat map of collagen tiles was the identification of collagen hot spots (cluster 3, red tiles). Such a hot spot can belong to either an anatomical structure of the liver: a portal tract, a central vein, the liver capsule; or alternatively, a region of pathological collagen deposition in the tissue. These architectural landmarks were consecutively the basis of tissue graph construction.

5.3.1 Graph construction

The graph construction pipeline is presented in Figure 5.5. Firstly, regions of high collagen content are identified using the tile labels generated by the tile subtyping pipeline (i.e. tiles with collagen in cluster 3 - see Fig. 5.4). Next, centroids of each dense collagen region are used as centre points for Voronoi tessellation. The final tissue graph is constructed from the whole slide using features extracted with an ImageNet pretrained ResNet18 model from each tile as node features X . The Euclidean distance D from each tile to its corresponding Voronoi centre is encoded as edge weight w . The resulting graph consists of star-like

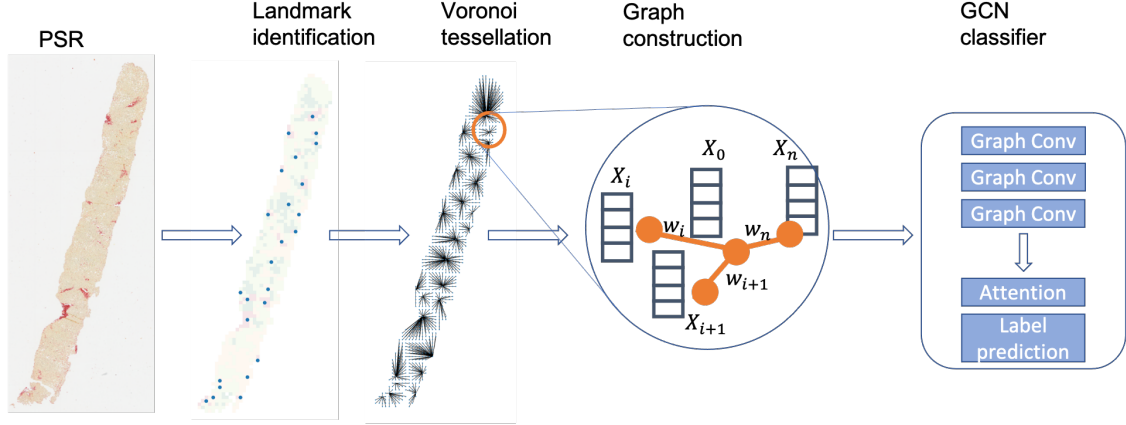


Figure 5.5: The slide classification pipeline. First, tissue landmarks are identified and are used as centres for a Voronoi tessellation of image tiles. Next, a tissue graph is constructed using features extracted from individual tiles and the tessellation weights. Finally, the graph representing the whole slide serves as an input to the GCN classifier.

structures, each one containing a biologically relevant region of the tissue (e.g. a single portal tract or region of fibrosis). The graph construction was implemented using NetworkX package [145].

5.3.2 Graph convolutional layers

Given the constructed graph $G(V, E)$ with N nodes, the associated features matrix $X \in \mathbb{R}^{N, 512}$ and adjacency matrix $A \in \mathbb{R}^{N, N}$ are used as an input to a 3-layer GNN. At each layer, a graph convolutional operator from GCN [146] is applied, followed by a non-linear activation function. The final nodes features matrix $H \in \mathbb{R}^{N, 64}$ contains nodes representations that encode both the local graph structure and the features of the nodes. Here G^* denotes each of the convolutional layers:

$$H' = \text{ReLu}(G^*(X, A)) \quad H'' = \text{ReLu}(G^*(H', A)) \quad H = \tanh(G^*(H'', A)). \quad (5.1)$$

As alternatives we also tested GAT [147] and GIN [148] and results can be found in Table 5.1

5.3.3 Attention layer

After the last graph layer, an attention layer is used to aggregate the node vectors into a set of 3 vectors, each vector being a representation of the input slide with

respect to one of the output classes. For each class, a set of attention scores is computed for the input vectors: each attention score represents the importance of an input vector for that class. The sum of the input vectors weighted by their corresponding attention scores generates the final vector representation of the slide with respect to the class.

Formally, attention for a class c is computed as a function of the input sequence H and a query vector u_c , jointly learned by the network during training. Similarity between each input element H_i and the query vector u_c is computed through dot product to obtain the corresponding attention weight $\hat{\alpha}_{ci}$. A softmax function is then applied to the vector of weights $\hat{\alpha}_c$ to obtain normalised weights α_c . Finally, the attention weights are used to compute a weighted sum of the input values and obtain the final slide representation v_c .

$$v_c = \sum_i \alpha_{ci} \mathbf{H}_i \quad \text{with} \quad \alpha_c = \text{Softmax}(\mathbf{H} u_c^\top). \quad (5.2)$$

Each slide vector v_c is passed to a fully connected layer for the final classification. The classifier has a separate set of weights to compute the raw score for each class $\hat{y}_c$. The class scores $\hat{y}_c$ are then concatenated and Softmax is used to obtain the likelihood vector $\hat{y}$.

$$\hat{y} = \text{Softmax}(\hat{y}_0, \dots, \hat{y}_C) \quad \text{with} \quad \hat{y}_c = \mathbf{W}_c^\top v_c + b_c. \quad (5.3)$$

The cross-entropy loss is computed from the likelihood vector $\hat{y}$ and the ground truth y . Training was performed to minimise the loss function by backpropagation.

As a baseline method, we used the approach from Yu et al. [75] with a ResNet18 applied to the whole pre-processed and downsampled slides [149]. In order to satisfy the memory limitations, the WSIs were here downsampled by a factor of 32.

5.4 Results

The proposed models were implemented using PyTorch [150] and PyTorch Geometric [151]. Training was performed using the Adam optimiser [152] with initial learning

Table 5.1: Mean F1 scores across fibrosis classes $\pm$ standard deviation. Results expressed in %.

Model	F0	F1	F2	F3
ResNet18	53.81 $\pm$ 6.36	21.40 $\pm$ 12.73	16.93 $\pm$ 15.00	24.44 $\pm$ 21.43
GCN	65.09 $\pm$ 5.50	19.75 $\pm$ 11.63	28.15 $\pm$ 24.48	70.71 $\pm$ 12.02
ATN-GCN	51.63 $\pm$ 2.83	27.27 $\pm$ 15.62	21.99 $\pm$ 6.72	75.56 $\pm$ 7.70
GAT	71.32 $\pm$ 8.56	34.56 $\pm$ 21.11	0.00 $\pm$ 0.00	44.44 $\pm$ 38.49
ATN-GAT	59.15 $\pm$ 9.09	33.23 $\pm$ 28.91	22.22 $\pm$ 38.49	19.05 $\pm$ 16.49
GIN	68.31 $\pm$ 0.23	13.33 $\pm$ 23.09	0.00 $\pm$ 0.00	13.33 $\pm$ 23.09
ATN-GIN	68.18 $\pm$ 0.00	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00	0.003 $\pm$ 0.00
	F0 - F1	-	F2	F3
ResNet18	67.98 $\pm$ 2.26		26.07 $\pm$ 7.74	0.00 $\pm$ 0.00
GCN	93.01 $\pm$ 0.53		18.79 $\pm$ 18.21	0.80 $\pm$ 0.00
ATN-GCN	89.26 $\pm$ 2.58		28.28 $\pm$ 11.00	78.57 $\pm$ 6.18
GAT	92.07 $\pm$ 1.40		27.78 $\pm$ 25.46	61.11 $\pm$ 9.62
ATN-GAT	86.22 $\pm$ 6.98		0.00 $\pm$ 0.00	0.00 $\pm$ 0.00
GIN	90.02 $\pm$ 0.00		0.00 $\pm$ 0.00	0.00 $\pm$ 0.00
ATN-GIN	90.02 $\pm$ 0.00		0.00 $\pm$ 0.00	0.00 $\pm$ 0.00
	F0 - F1 - F2	-	-	F3
ResNet18	74.19 $\pm$ 1.78			6.36 $\pm$ 5.53
GCN	99.03 $\pm$ 0.00			85.71 $\pm$ 0.00
ATN-GCN	99.36 $\pm$ 0.55			88.57 $\pm$ 10.30
GAT	96.85 $\pm$ 1.08			38.89 $\pm$ 34.69
ATN-GAT	96.90 $\pm$ 1.04			0.00 $\pm$ 0.00
GIN	96.30 $\pm$ 0.00			0.00 $\pm$ 0.00
ATN-GIN	96.30 $\pm$ 0.00			0.00 $\pm$ 0.00

rate of $1e-02$, $\beta_0 = 0.9$, $\beta_1 = 0.999$, $\epsilon = 1e-08$. For the GCN, we use a weight decay factor of $1e-5$. Models were trained for at least 300 epochs and training was stopped when the change of the loss function was negligible. When training the GCN, we applied dropout with probability 0.2 before the final classifiers. We use a batch size of 64 for the GCN and a batch size of 1 for the baseline ResNet18 due to memory limitations. All models have been trained on an NVIDIA Quadro RTX 6000 GPU. The source code of our analysis is available at <https://github.com/mkatw/gnn-fibrosis>.

The classification metrics (F1 scores) of the trained models are gathered in Table 5.1. The results reported in are averaged across three runs. Here, we have

tested three distinct classification setups. In the first, all fibrosis stages (F0-F3) were used directly as prediction classes. In the second and third setup, classes (F0 - F1) and (F0 - F2) were merged, respectively. This was done because stage F0, F1, and F2 cases have a similar macroscopic appearance and are often difficult to distinguish even for a trained pathologist. The experiments demonstrate that our GCN performs better than ResNet18 in each of the classification tasks.

Figure 5.6 shows examples of predicted tissue graphs overlaid with the original WSIs. It can be seen that the predicted Voronoi regions vary in shape and size and divide the tissue sections into meaningful regions which resemble liver lobules or acini.

The results of the quantitative analysis of slide fibre content using the identified tile subtypes are presented in Figure 5.7. Measuring the distribution of tile subtypes allows for a more fine-grained quantitative analysis of fibrosis in a slide. The results of tile subtyping for each case are summarised using a new collagen metric here called the weighted fibrosis score. The weighted fibrosis score is calculated using the formula from equation (5.4):

$$WFS = \frac{1}{C-1} \sum_{c=0}^{C-1} cp_c, \quad (5.4)$$

where C is the number of classes and p_c is the fraction of tiles of class c in a given slide.

A large variation in the overall fibrosis score can be seen within each of the pathology stages. The variation in the weighted score is expected, as a similar phenomenon is present in CPA measurements across the progressing stages of fibrosis [139]. Selected properties of the constructed tissue graphs are shown in Figure 5.8.

5.5 Conclusions

We have proposed a new graphical model that facilitates a quantitative analysis of fibrosis in liver biopsies. The proposed model, Attention-GCN has allowed to perform automatic staging of PSR-stained liver biopsies into cases with bridging fibrosis and with no bridging fibrosis. The proposed method has outperformed the baseline model in each of the tested classification tasks.

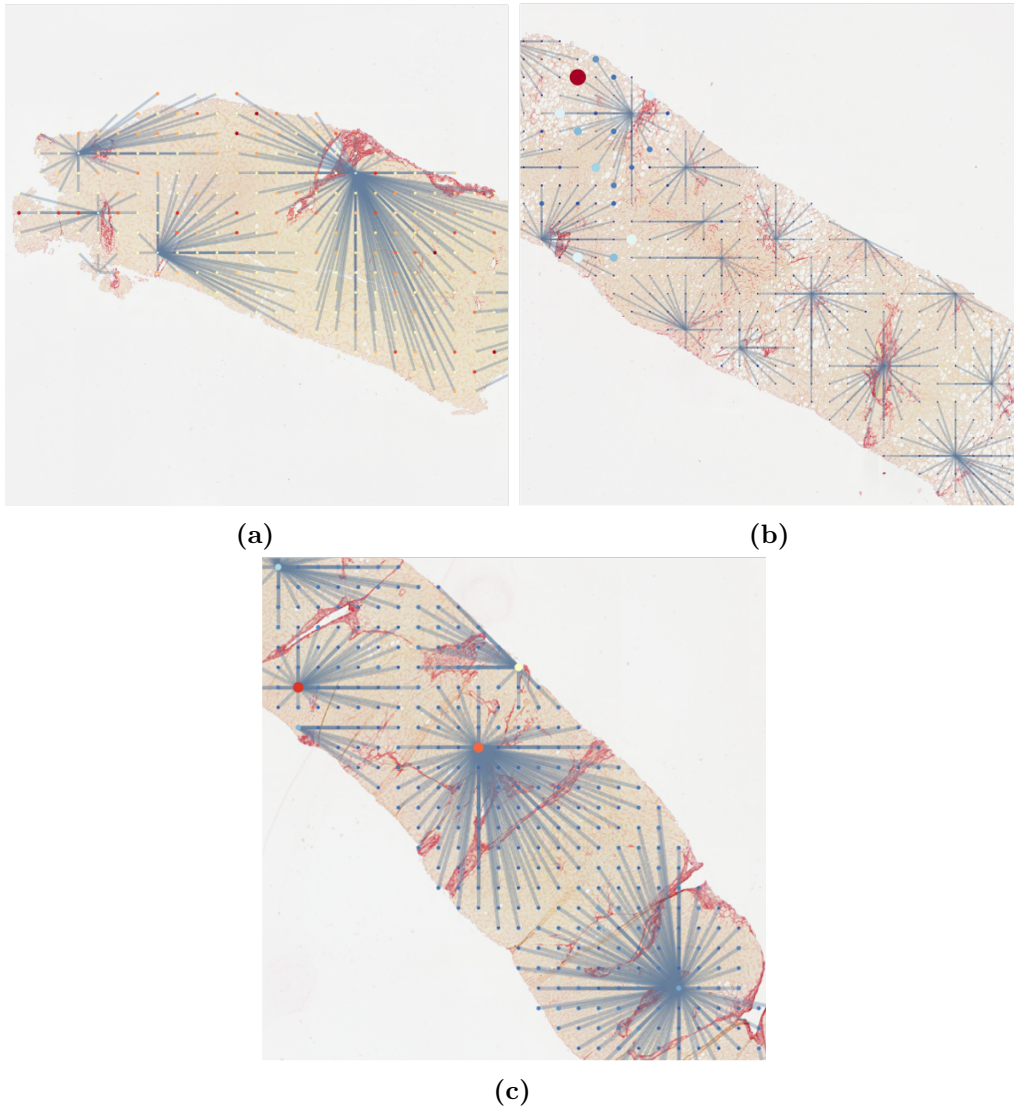


Figure 5.6: Examples of tissue graphs superimposed on the original images. Node size and colour reflect attention activation for the predicted slide class. Notice that the edge length in the individual sub-graphs corresponds to the underlying pattern of fibrosis and that the sub-graphs capture biologically meaningful regions, e.g. (5.6a) Stage F0 with no signs of fibrosis. (5.6b) Stage F2 with early fibrotic clusters in the parenchyma. (5.6c) Stage F3 with prominent bridging.

The developed pipeline allows for an explicit separation of fibres in the slide into localised fibrosis patterns (tile subtypes) and examination of those patterns. The arrangement of individual tiles into graphs allows to divide the tissue into biologically meaningful sub-regions, which can be inspected by a pathologist. This representation fully aligns with established fibrosis scoring such as METAVIR. Importantly, this approach can be integrated into an interactive scoring system

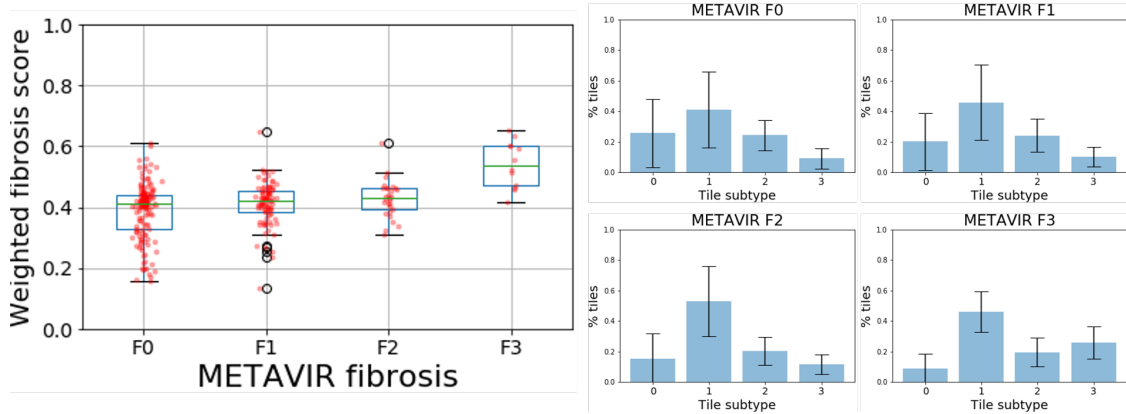


Figure 5.7: The weighted fibrosis score across fibrosis stages and the distribution of tile subtypes across each METAVIR stage. The weighted fibrosis score, which is not informed by tissue topology, does not separate between classes F0 and F1 (Wilcoxon signed-rank test coefficients: $W=-1.798$, $p=0.072$) and classes F1 and F2 ($W=-0.775$, $p=0.438$). However, class F3 can be differentiated from all the other classes: F0 ($W=-4.436$, $p<0.001$), F1 ($W=-4.181$, $p<0.001$), F2 ($W=-3.559$, $p<0.001$), and class F2 can be differentiated from F0 ($W=-2.062$, $p=0.039$).

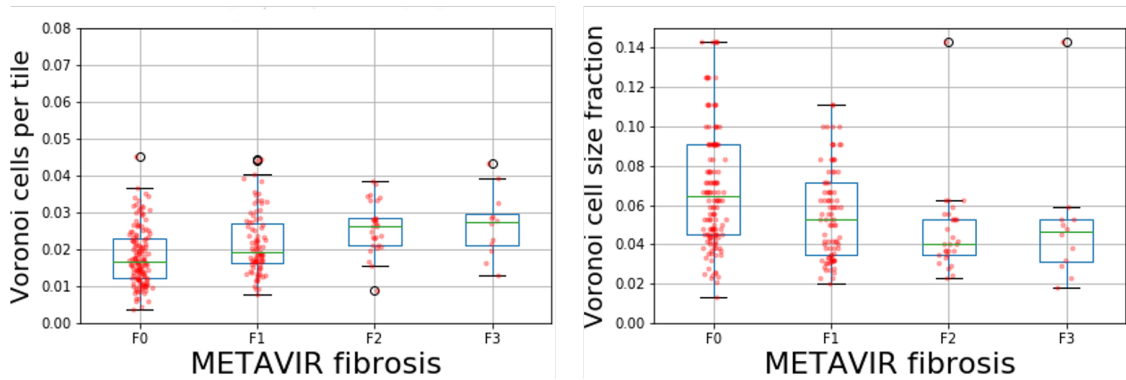


Figure 5.8: The number and size of the Voronoi cells across the fibrosis stages. It can be seen that the size of a Voronoi cell relative to the area of the biopsy can change by an order of magnitude, therefore allowing to capture biological objects of varying sizes.

where pathologists can eliminate specific graphs from the analysis. As the developed method relies on binary images of collagen in tissue slides, it could easily be extended to different fibrosis stainings and even different imaging modalities.

6

Beyond Staging: Localised Analysis of Disease Advancement

Contents

6.1 Characterisation of fibrosis sub-patterns	91
6.2 Methodology	93
6.2.1 Collagen tile subtyping	93
6.3 Results	98
6.3.1 CPA	101
6.3.2 Weighted-fibrosis score and tissue graph parameters	102
6.4 Visualisation of the learned disease space	105
6.5 Tracking disease progression	110
6.6 Conclusions	112

Fibrosis can be observed in the liver at multiple scales simultaneously. While CPA has become an established quantitative imaging marker of fibrosis, it does not allow to capture the topology of changes in liver fibrosis. Established fibrosis staging systems offer a semi-quantitative way of describing changes in tissue architecture, however they are in turn burdened by high inter-observer variability [1]. Because of these limitations, there now is a large interest in devising complementary fine-grained metrics to describe the quality of fibrotic patterns with higher accuracy. This interest is reflected in liver pathology literature with introduction of scores for chicken-wire fibrosis [4], progressive/regressive patterns [49] and substages of cirrhosis [42, 48].

6.1 Characterisation of fibrosis sub-patterns

The published clinical findings clearly demonstrate the utility of CPA in tracking disease progression and regression in cases with advanced bridging and cirrhosis, where it was shown to predict patient outcome [49]. However, it is also known that the inherent sampling variability of the needle biopsy introduces noise to CPA quantification. The primary limitation of CPA as an imaging biomarker of fibrosis is that collagen in the slide can belong to either normal anatomical structures (i.e. portal tracts, central veins and the liver capsule) or pathological collagen regions such as regions of pericellular fibrosis or fibrotic bridges.

For this reason, to counteract the influence of the presence of large portal tracts in the tissue slices, some groups have performed manual exclusion of portal tract walls in their collagen area quantitation, leading to a stronger overall correlation of CPA with fibrosis stage [74]. However, it can be argued that such a computation is really the computation of the area of a particular subtype of collagen pattern in the tissue rather than of collagen overall and therefore calling it "collagen area" is confusing. It is known that portal tracts also undergo pathological changes in the process of development of liver fibrosis. It is possible that excluding portal tract regions such an analysis misses the opportunity of quantifying the change in portal tract thickness with fibrosis progression.

Conversely, other groups have conducted studies into the influence of the collagen area of regions of portal tracts and portal bridging. A study by Sandrini et al. has found that so-called "portal-bridging fibrosis area" is a more accurate predictor of fibrosis stage than CPA [88]. Another example of such a pathologist-assisted quantitation is the quantitation of width of annotator-identified cirrhotic septa [89]. A more recent paper has demonstrated that the quantification of amount and quality of collagen fibres in the different parts of the liver lobule provides greater insights into regression of MAFLD induced fibrosis than the traditional metrics [82].

Performing manual annotations on tissue slides is a laborious process which is known to be burdened with high variability between annotators [79]. However, unsupervised machine learning techniques such as clustering methods can be utilised

to identify patterns in histological data without the need for human annotations. We propose that learning a representation of fibrosis subtypes directly from the data can offer fresh insights into the trajectory of development of liver fibrosis.

The hierarchical structure of fibrosis which can be examined using digital pathology methods is summarised in Figure 6.1. In this chapter we propose a novel way of describing the hierarchical structure of liver tissue from whole slide images and incorporating it into a joint representation. This is accomplished through a combination of descriptors acquired at different scales: the pixel level annotations, local fibre subtypes, anatomical regions and pathological stage for a final comprehensive description.

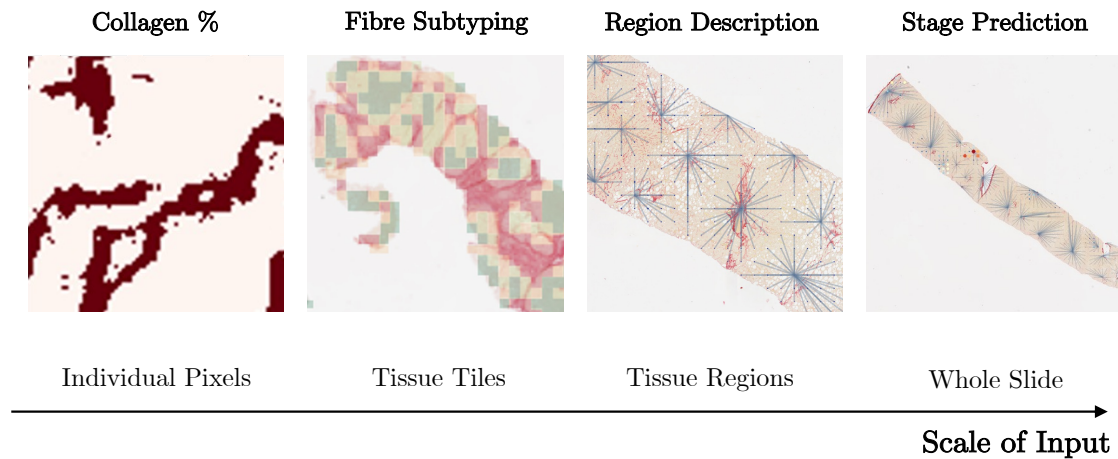


Figure 6.1: The hierarchical structure of liver fibrosis. Liver fibrosis can be observed at different levels ranging from individual pixels in microscopic images through local textural patterns, biological regions and whole slides. Advanced fibrosis can be further observed at the scale of the whole organ, using modalities such as CT or MRI.

In order to facilitate the analysis in the middle-range of tissue magnification, here we introduce a novel subtyping of liver micro-structural patterns learned directly from small tissue tiles. Similar methods have been recently successfully applied to description of fibrosis in bone marrow biopsies [153]. Our subtyping, learned in an unsupervised fashion is then applied to a large multi-centre cohort of tissue samples with different disease etiologies and the representation of all stages of fibrosis. Finally, the results of the analysis are shown for a small number of longitudinal samples to demonstrate the utility of this method in tracing fibrosis progression over time.

6.2 Methodology

Collagen proportionate area was computed for samples from all of the five studied cohorts. The training of the collagen segmentation pipeline is described in detail in Chapter 4. A variant of the U-net mini model was used to segment collagen in tissue tiles from all the cohorts. Slides scanned at 20x were segmented using a model trained on the combined cohort. A separate model for the segmentation of collagen in the FLINC cohort was developed following the exact same methodology. Here 512 by 512 px tiles of PSR stained tissue were annotated and processed at 40x magnification, and the images were processed at the highest magnification level. Tissue area was obtained for all slides at 20 x magnification. Holes (vessel lumens and empty fat vesicles) were excluded from tissue area computation.

The subsequent analysis was applied to whole slide images from the cohorts acquired at 20x magnification. Tissue graphs were computed for all the processed slides and the size parameters of the fitted Voronoi cells were extracted. The results of the computations are gathered in the results section below.

6.2.1 Collagen tile subtyping

The tile subtyping pipeline was developed using tiles from the lowest level of the WSI image pyramid at 20x. Example PSR tissue tiles used for learning the subtypes are shown in Figure 6.2. As can be seen, the tiles differ from each other not only by the amount and texture of collagen in the image, but also by features unrelated to fibrosis, such as slide colour, presence of background areas and emptied fat bubbles.

In order to learn the tissue tile subtypes along all stages of liver fibrosis a sub-cohort consisting of 84 cases with an equal distribution of each fibrosis stage was gathered. Here only cases from the CALM and PREV studies were used. As fibrosis was scored according to different pathological systems in each study, stages from the Ishak system were here cast onto the stages from the METAVIR system to offer an equalised representation of disease progression in the combined cohort [140]. The collagen subtype training cohort along with the mapping between the Ishak and METAVIR systems used is briefly summarised in Table 6.1.

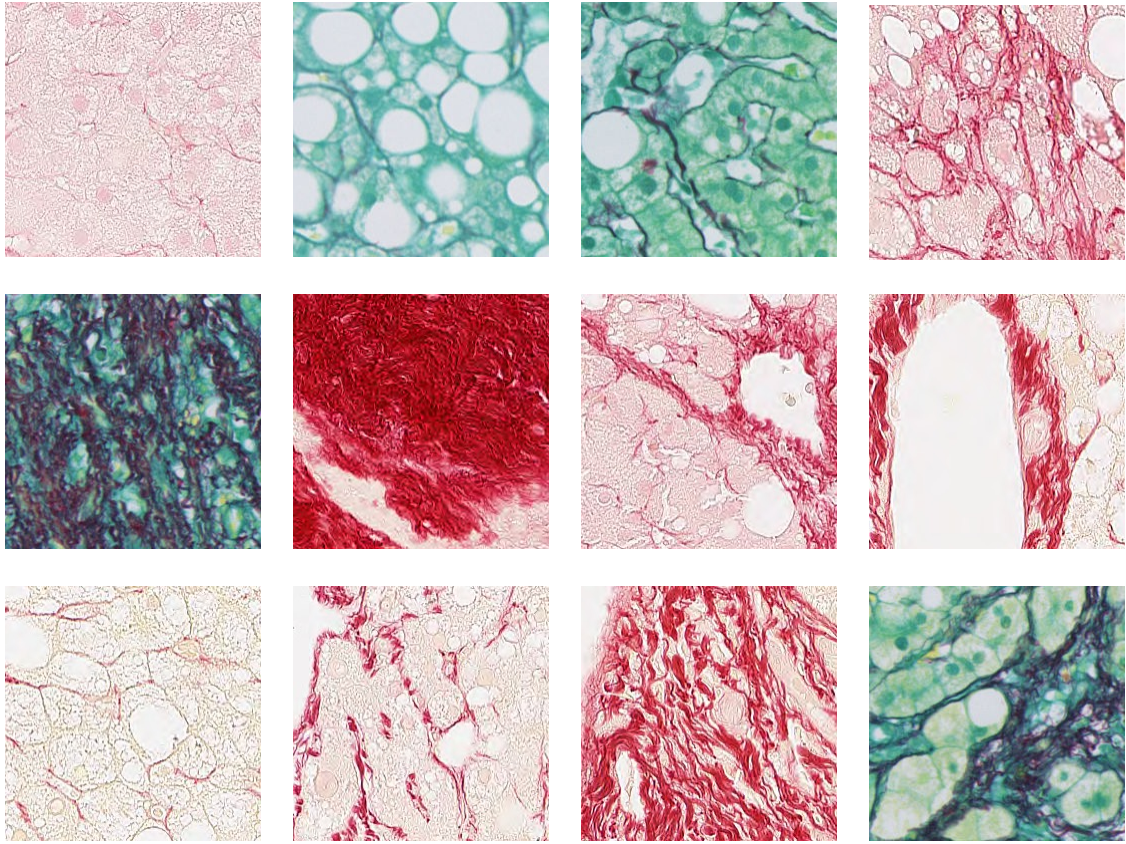


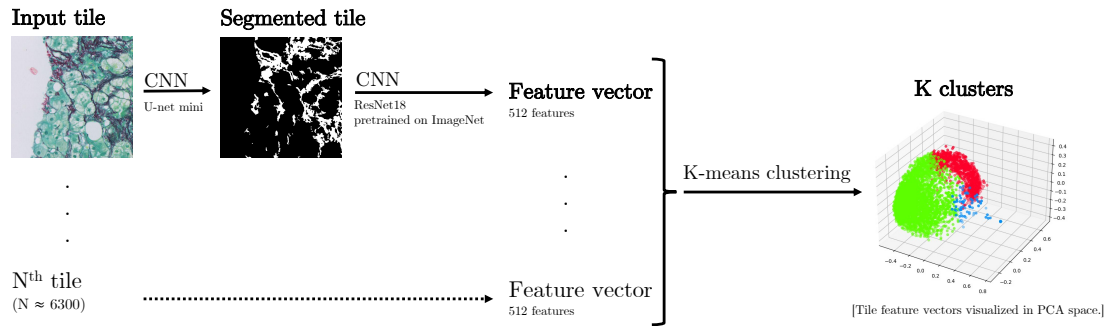
Figure 6.2: Examples tiles used for learning the subtypes of local collagen patterns. Each of the images shows a region of liver tissue stained with a variant of PSR staining for collagen. The size of each square is 256 by 256 pixels, corresponding to a tissue region of approximately 127 by 127 μm .

Overall about 63 000 PSR-stained tissue tiles were extracted from the 84 cases and a randomly selected subset of 10% of those tiles was used to learn the tissue clusters. The tile clustering pipeline is summarised in Figure 6.3. Firstly, each input tile was preprocessed by applying collagen segmentation in order to remove the influence of slide colour, tissue boundaries and fat from the learned collagen subtypes. Next, a ResNet18 model pre-trained on images from the ImageNet challenge was used to extract visual features from the binary collagen tiles [149]. The extracted feature vector contained 512 features for each image tile. Finally k-means clustering was applied to the collection of tile features to identify subtypes of tissue tiles.

K-means clustering is an unsupervised learning method which is often used for clustering data points into k groups. The number of clusters (k) is a parameter chosen by the user. The algorithm works by randomly selecting k data points as

Table 6.1: Characteristics of the training cohort for the collagen tile subtyping pipeline.

Ishak stage	n (CALM)	METAVIR stage	n (PREV)	Total
0	6	0	6	12
1	6	1	6	12
2	6	2	12	24
3	6			
4	8	3	10	24
5	8			
6	12	4	4	12
sum =	52		34	84

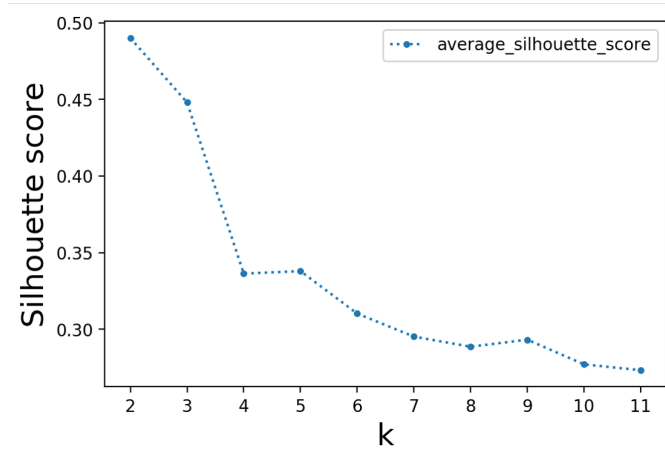
**Figure 6.3:** Tile clustering pipeline.

initial cluster centroids and assigning each of the data points to the nearest centroid, based on Euclidean distance. In each iteration of the algorithm the centroid of the cluster is updated as the mean of all the data points assigned to the cluster. The process is repeated until the cluster assignments no longer change. Here we've explored clustering of tissue tiles with different values of k to learn a subtyping which best represents the data. The characteristics of the learned clusters as a function of k are shown in Figure 6.4.

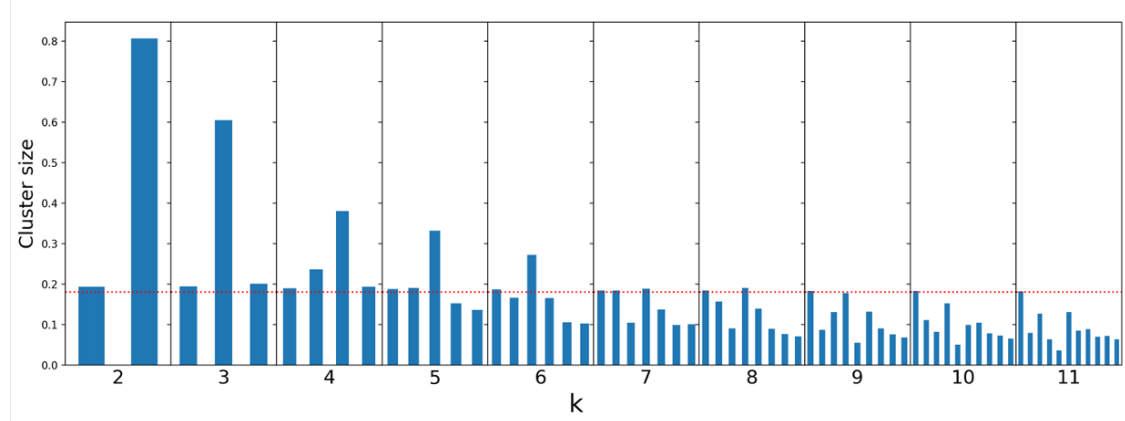
One of the measures which can be used to verify how well data points are separated into clusters is cluster silhouette score. The silhouette score is computed for each point as follows:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

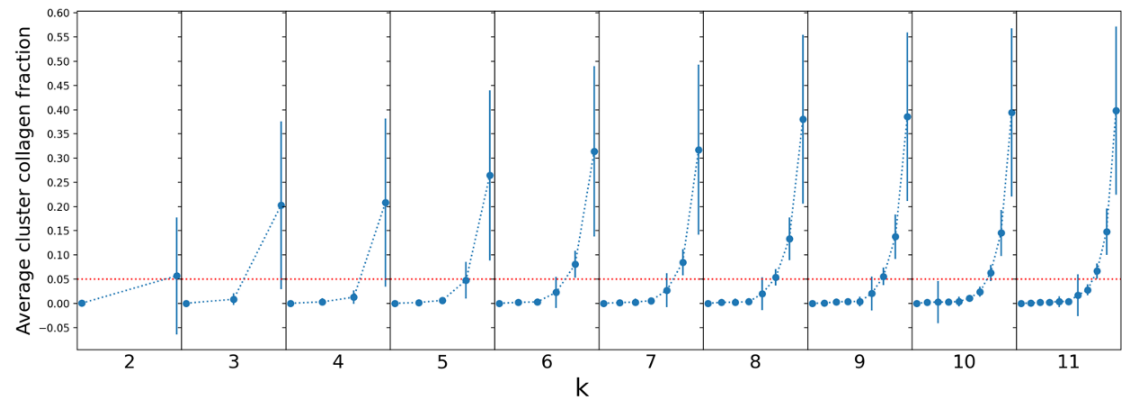
where: $s(i)$ is the silhouette score for data point i , $a(i)$ is the average distance between data point i and all other points in the same cluster, $b(i)$ is the average distance between data point i and all points in the nearest neighbouring cluster.



(a) The change in the mean cluster silhouette score depending on the number of clusters.



(b) The distribution of cluster sizes depending on k . Each bar represents a single tile cluster. For each k the bars are ordered by their mean collagen content. The dotted line represents the proportion of tiles with no collagen in the training cohort (18%).



(c) The distribution of mean cluster collagen content depending on k . The dotted line represents 5% of collagen in a tile.

Figure 6.4: Collagen tile cluster properties as a function of the number of clusters.

The overall silhouette score for the clustering is then given by the average of the silhouette scores for the individual data points:

$$\text{silhouette score} = \frac{1}{n} \sum_{i=1}^n s(i)$$

It was found that the values of cluster silhouette score were dropping with an increase of k . This can be explained by the fact that the visual features in the analysed collagen tiles could be overall considered highly similar.

Instead, in order to derive an interpretation of the consecutive tissue tile clusters, we calculated mean cluster collagen content. Figure 6.5 shows the distribution of tile collagen % in the whole population of the training cohort. It can be seen that the distribution of collagen percentage in the training tiles follows the pattern of exponential decay. The majority of the clustered tiles contain less than 5% of collagen.

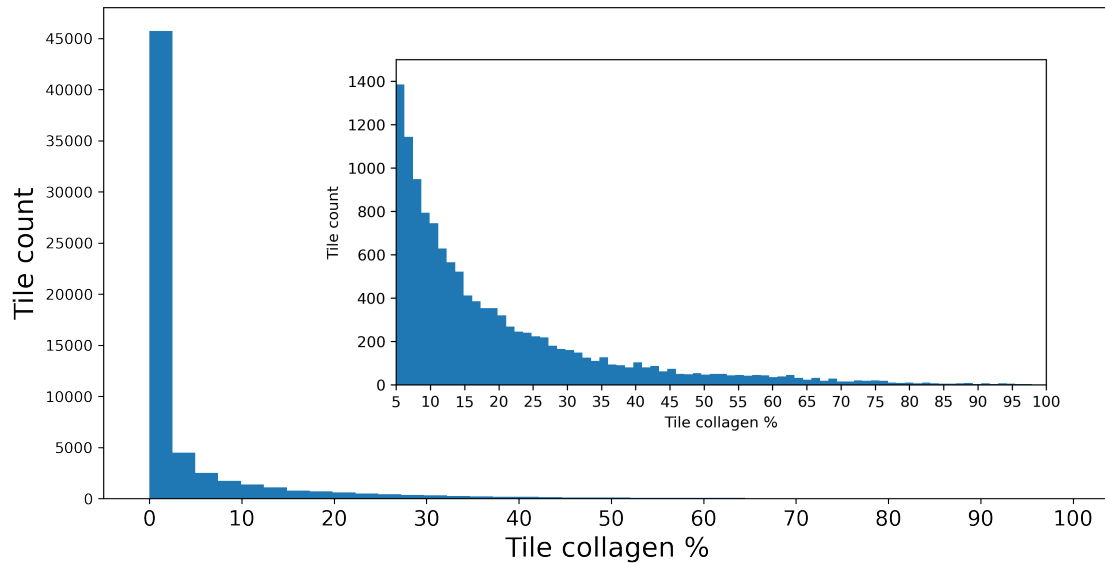


Figure 6.5: The distribution of tile collagen content across the tile population. The distribution of tile collagen in the cohort follows the pattern of exponential decay. Most of the tiles in the cohort contain very little collagen. The smaller figure is zoomed-in to the tail portion of the whole distribution.

The results of clustering at each k were compared in relation to mean cluster collagen size (6.4b), average tile collagen % in each of the clusters (6.4c) and the visual inspection of tiles in the obtained clusters. For each k , the clusters were

ordered according to their mean collagen content with cluster 0 containing the least collagen and cluster k containing the most. It was observed that the size of cluster 0 remained very similar irrespective of the value of k , meaning that empty tiles always formed a separate cluster.

K value of 4 returned one cluster of which the mean value of tile collagen was over 5%. $K=4$ was previously used to identify collagen hot-spots as potential Voronoi centres in the tissue graph building pipeline described in Chapter 5. It was observed that for k values equal to or higher than 8, there were consistently three clusters in the collagen area range above 5%. Increasing k above 8 resulted in creation of sub-clusters in the tiles with very little collagen. Because of this $k=8$ was chosen to perform tissue subtyping.

Figure 6.6 shows example collagen tiles from each of the 8 clusters along with the mean collagen content of tiles in each cluster. This learned collagen tile classification was applied to all the histology slides from the cohorts imaged at the resolution of 20x.

6.3 Results

The results of collagen tile subtyping are shown for four example cases from the analysed cohorts in Figure 6.7. Here the collagen subtypes are represented by tile colour overlaid over the original whole slide image. For each case, additionally the histogram of tissue tiles belonging to each of the clusters is displayed along with additional characteristics of the slide. Weighted fibrosis score (WFS) values were computed from the obtained cluster distributions using equation 5.4.

A quick look at the obtained collagen distributions reveals that the information they contain differs from both CPA values and pathologist assigned fibrosis stage. Case a) shows a needle biopsy which could be considered normal liver from the perspective of fibrosis. In this case all the obtained metrics are in concord. The pathologist-assigned fibrosis stage is F0 and the values of both CPA and WFS are close to 0. The tile histogram shows that most of the tiles in the slide contain little to no collagen.

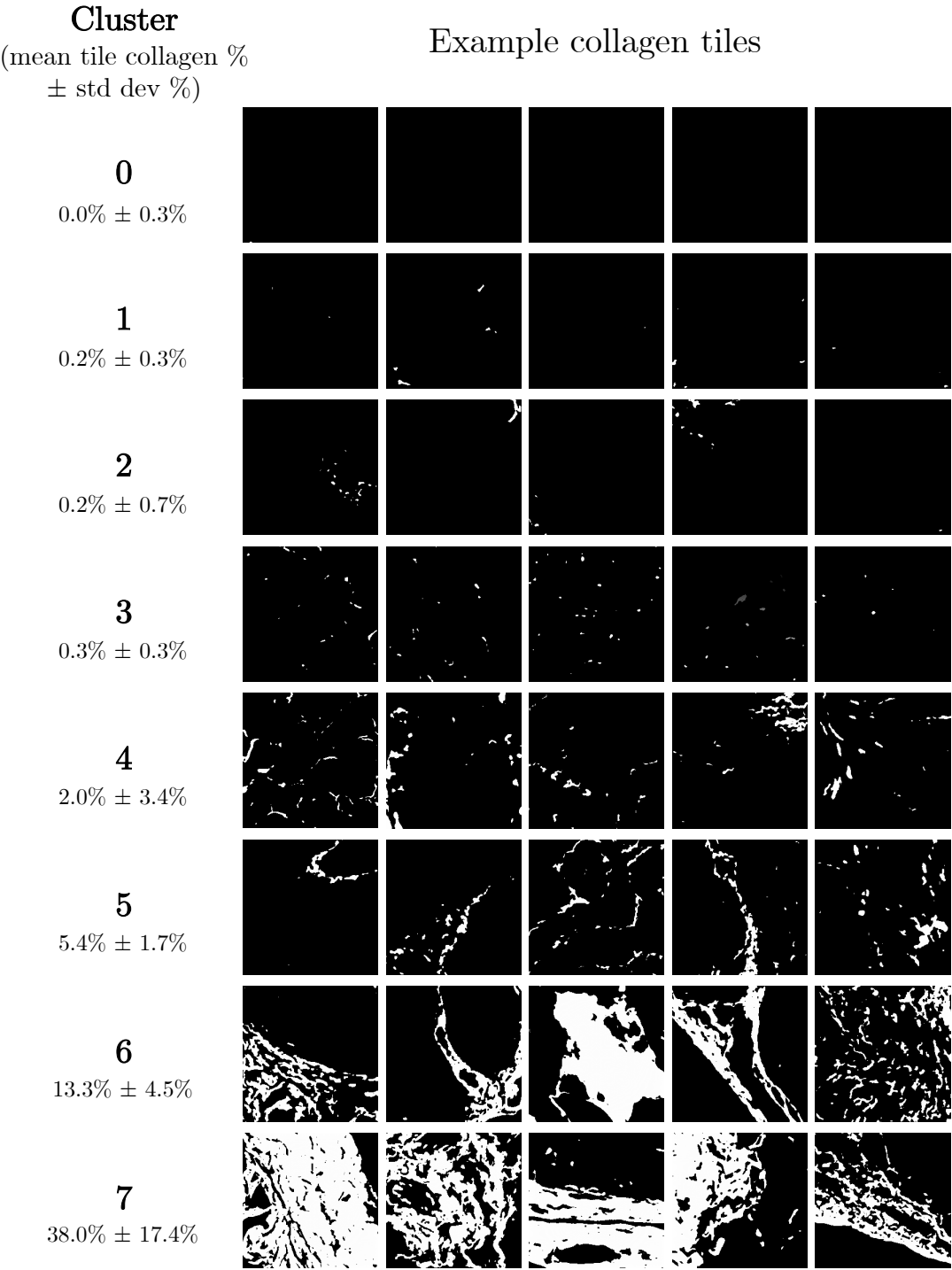


Figure 6.6: Example collagen tiles from k=8 clusters, ordered by mean collagen %.

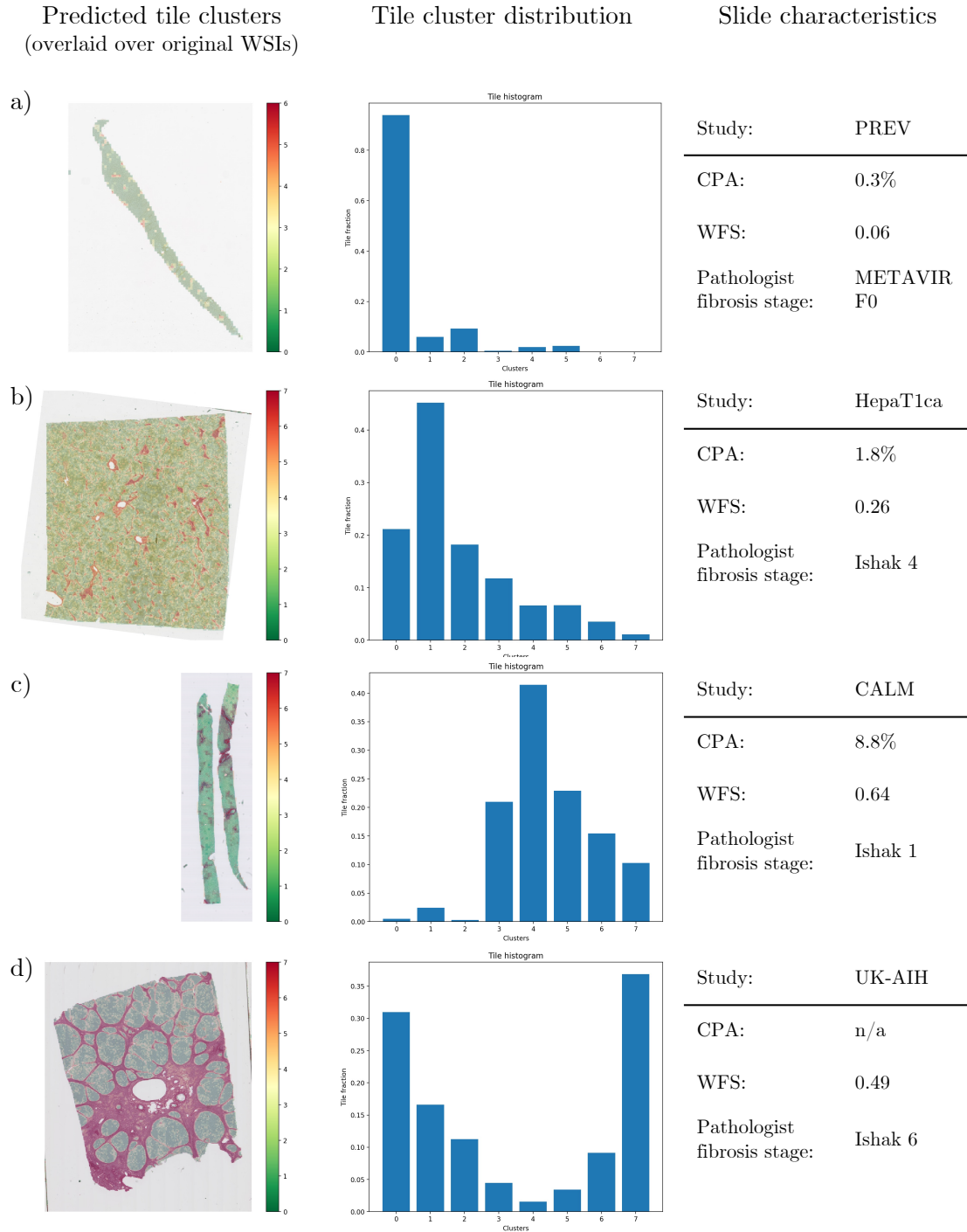


Figure 6.7: Examples of tissue samples with different tile cluster distributions. It can be seen that the computed tile clusters, as well as the weighted cluster score (WFS) provide different information about the image than both CPA and pathologist fibrosis stage. In particular, the tile clusters allow to observe the quality of collagen fibres in different parts of the tissue. The histograms allow to visualise the proportions of tiles from each of the clusters in an image.

Cases b) and c) present livers with fibrosis. Case c) was assigned Ishak stage 1 by the annotating pathologist. This means that some fibrotic expansion but no bridging was observed. The tile cluster histogram reveals that overall the parenchyma contains few tiles with no collagen and many tiles belonging to clusters 3 and 4 which contain small local collagen depositions. These depositions could be interpreted as regions of pericellular fibrosis infiltrating the tissue. Case b) shows a surgical liver resection which was classified as stage Ishak 4 (marked bridging). It can be seen that the visual context available in the examination of this case is much larger than in case c). Both CPA values and tile histograms suggest that case c) has progressed further in the development of fibrosis than case b). It is a well-known phenomenon in liver fibrosis assessment that the size of the sample heavily influences the outcome of pathological staging, with smaller samples often being assigned milder fibrosis stages [86]. It is possible that the narrow shape of the needle biopsy sample in case c) caused that no bridging was observed and that if a larger piece of tissue was available for examination, a different histological stage might have been assigned.

Finally case d) shows a liver with clear cirrhosis with regenerative nodules (stage Ishak 6). The tissue is characterised by high heterogeneity, containing both regions of dense collagen (fibrotic septa) and of relatively collagen-free parenchyma (regenerative nodules). This heterogeneity is captured by the bi-modal shape of tile cluster histogram in a way which cannot be expressed by a single number such as CPA¹ or WFS. It can be seen that while case d) is cirrhotic, the inferred WFS value is smaller than for case c). This shows a limitation of WFS as a numerical descriptor of fibrosis, as it is merely the average of the computed cluster distribution.

6.3.1 CPA

The results of CPA quantification for data from the combined study cohort are presented in Figure 6.8. The results of CPA quantification for each of the individual

¹The CPA value for this case was unavailable at the time of thesis submission due to extremely large size of the whole slide image, which couldn't be loaded into computer memory for CPA quantification.

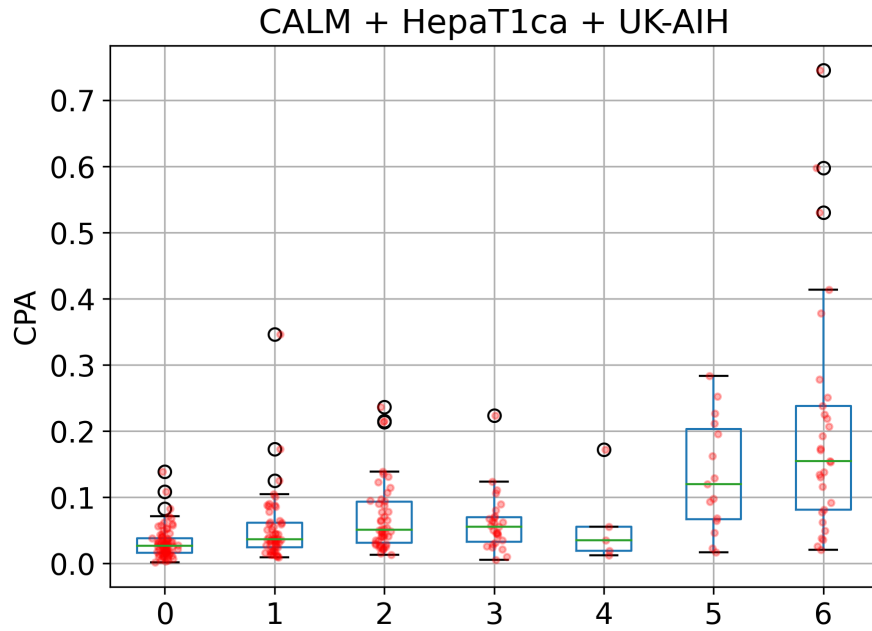
cohorts are further shown in Figure 6.9. The correlation statistics values for the individual and combined cohorts are gathered in Tables 6.2a and 6.2b, respectively. The obtained correlation values are in line with previously published data on fibrosis stage prediction from CPA [20, 70, 74]. It can be seen that CPA correlates with fibrosis stage in all cohorts except for the PREV cohort, which contains mostly samples with normal livers and early fibrosis. This showcases a known limitation of CPA as an imaging marker, that is its low predictive value in early fibrosis.

Table 6.2: Correlation statistics between CPA and pathologist-assigned fibrosis stage.

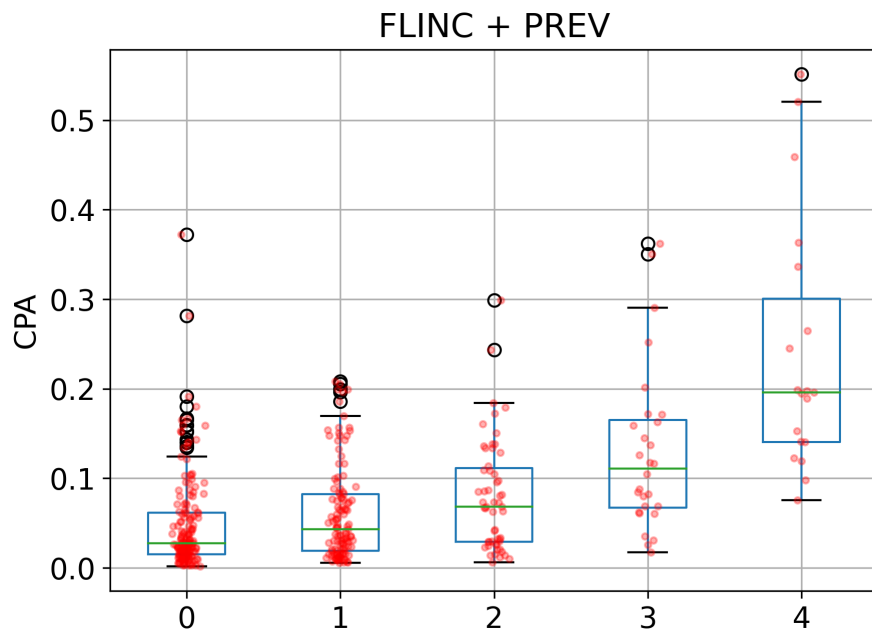
(a) CPA vs. Ishak fibrosis stage.		
Cohort	Correlation coefficient	p-value
CALM	0.39	$p < 0.001$
HepaT1ca	0.41	$p = 0.006$
UK-AIH	0.63	$p < 0.001$
Combined	0.55	$p < 0.001$
(b) CPA vs. Kleiner-Brunt/METAVIR fibrosis stage.		
Cohort	Correlation coefficient	p-value
FLINC	0.48	$p < 0.001$
PREV	0.07	$p = 0.22$
Combined	0.50	$p < 0.001$

6.3.2 Weighted-fibrosis score and tissue graph parameters

The results of WFS quantification for data from the combined study cohort are presented in Figure 6.11. Here, rather than predicting the overall fibrosis stage using WFS as an imaging marker, we compare it to CPA. Figure 6.10 demonstrates the relationship between CPA and WFS with a fitted regression curve. It can be seen that WFS creates a high spread between study points in the low CPA range, suggesting that it differentiates images with different collagen pattern subtype distribution. A case with no collagen infiltration in the parenchyma but with a large portal tract visible in the biopsy can have the same CPA as a case with regions



(a) Cohorts staged according to the Ishak system.



(b) Cohorts staged according to the Kleiner-Brunt system. The PREV cohort has been staged using the METAVIR staging system, which used the same fibrosis stage definitions as the Kleiner-Brunt system. Here the categories from both systems and studies were assumed to be equivalent.

Figure 6.8: CPA by stage of fibrosis in the combined cohort.

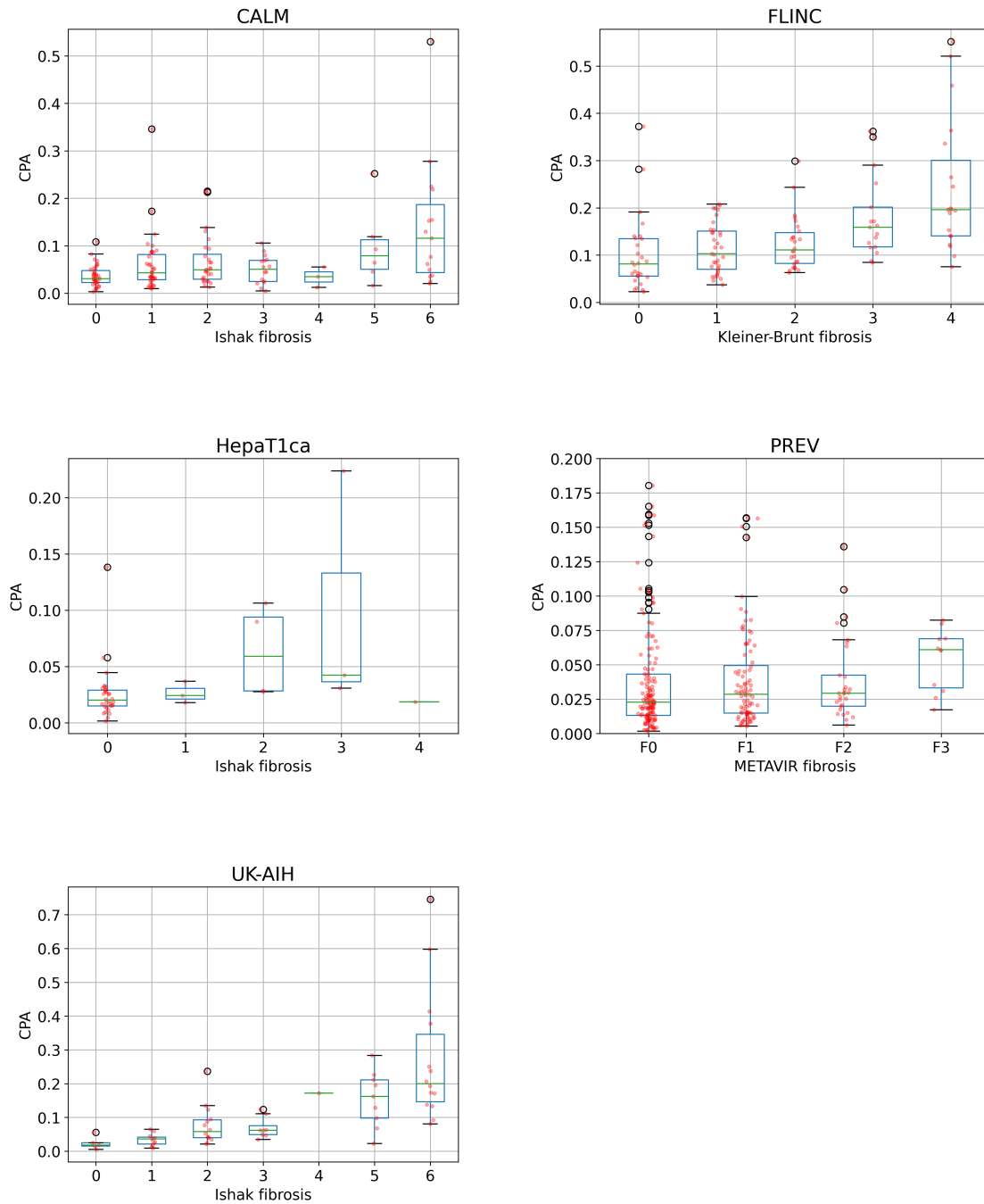


Figure 6.9: Mean CPA by fibrosis stage in each of the individual studied cohorts.

of peri-cellular fibrosis. However, as was shown before, the interpretation of WFS values without the corresponding collagen subtype histograms is aimless.

Similarly, Figure [6.12](#) shows the results of Voronoi cell size quantification in the studied cohorts.

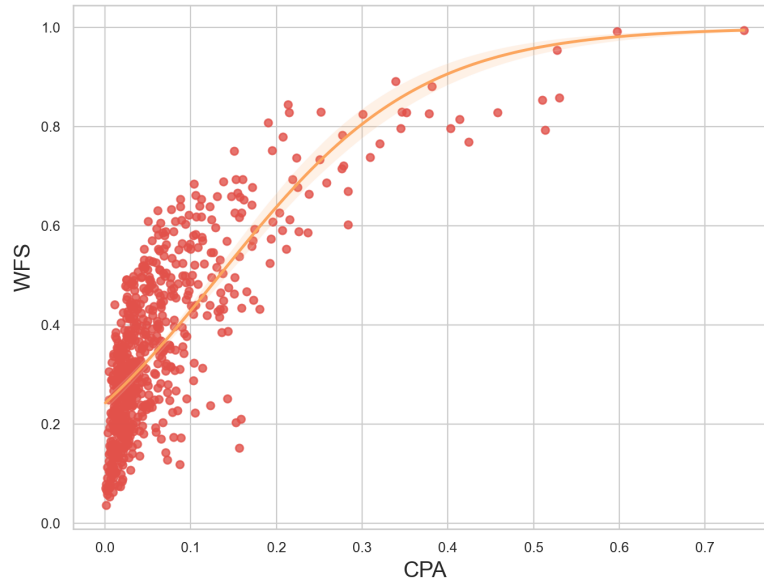


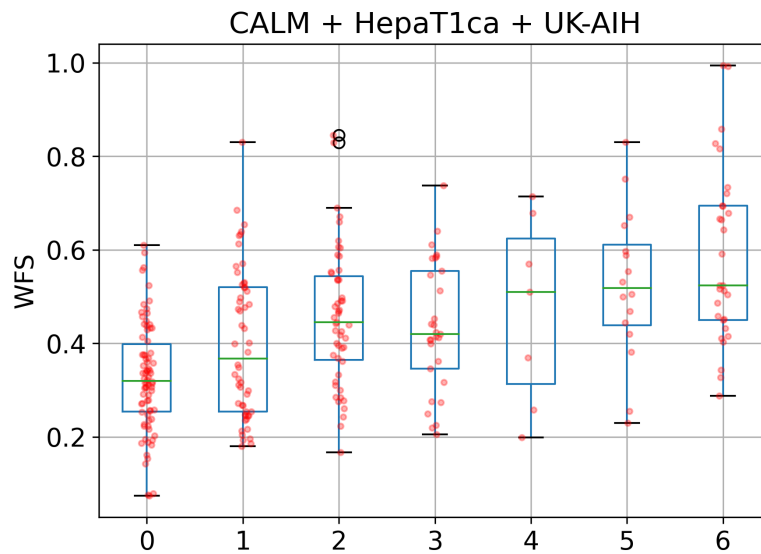
Figure 6.10: Logistic regression plot fitted to WFS as a function of CPA. The points represent individual biopsies from the CALM, HEpaT1ca, PREV and UK-AIH cohorts.

6.4 Visualisation of the learned disease space

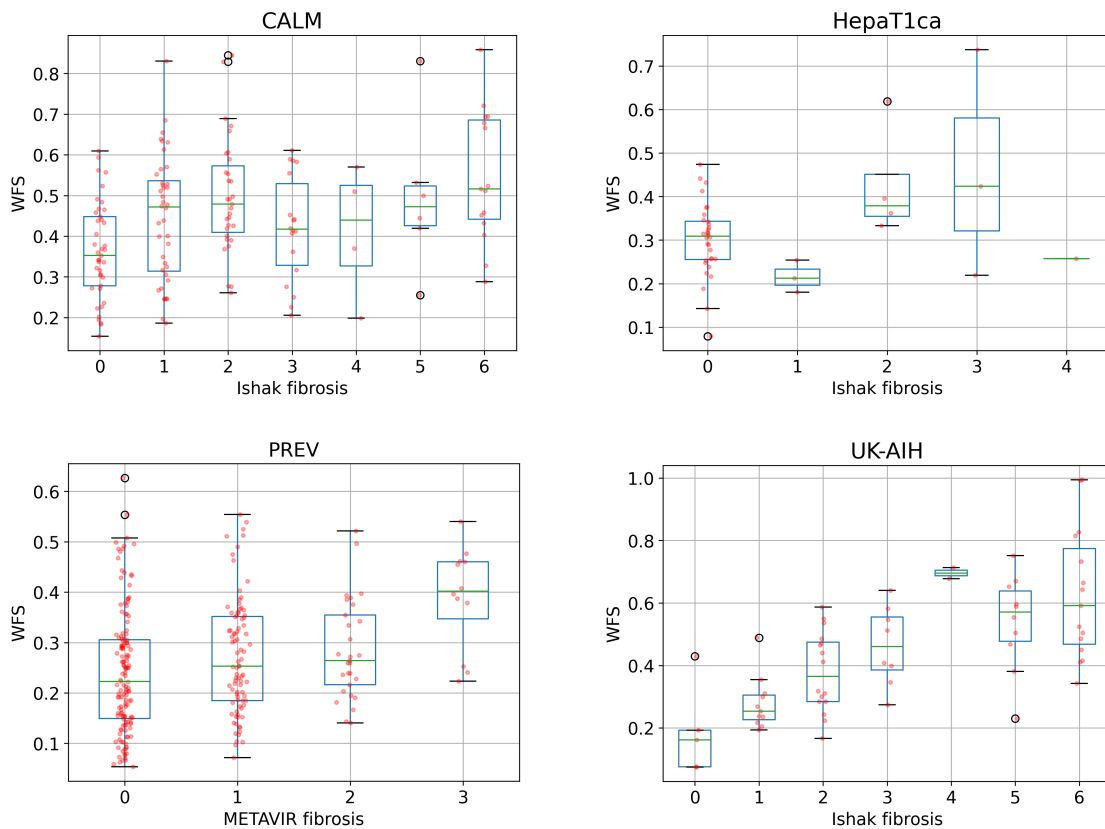
Principal component analysis (PCA) was performed to visualise the distribution of the collected biopsies in the studied disease space. PCA is a statistical technique for reducing the dimensionality of the dataset while capturing the maximum variance along the quantified features.

Here the inputs to PCA were: case CPA, weighted fibrosis score, the proportions of tiles of each fibrosis cluster in each case and the number of computed Voronoi cells normalised by the number of tiles in the image. Figure 6.13 shows the obtained distribution of the biopsies along the principal components.

It can be seen that samples from the different individual studies maintain a tendency towards different sides of the learned distribution. This is to be expected as it was previously shown in Chapter 3 that the respective studies represent different kinds of liver disease and different distributions of fibrosis stages. Figure C.1 shows mean values of CPA by study for comparison.



(a) Combined cohort staged according to the Ishak system.



(b) Mean WFS by fibrosis stage in the individual study cohorts.

Figure 6.11: Mean WFS by stage of fibrosis in the combined study cohort.

6.11a cohorts staged according to the Ishak system. **6.11b** individual study cohorts.

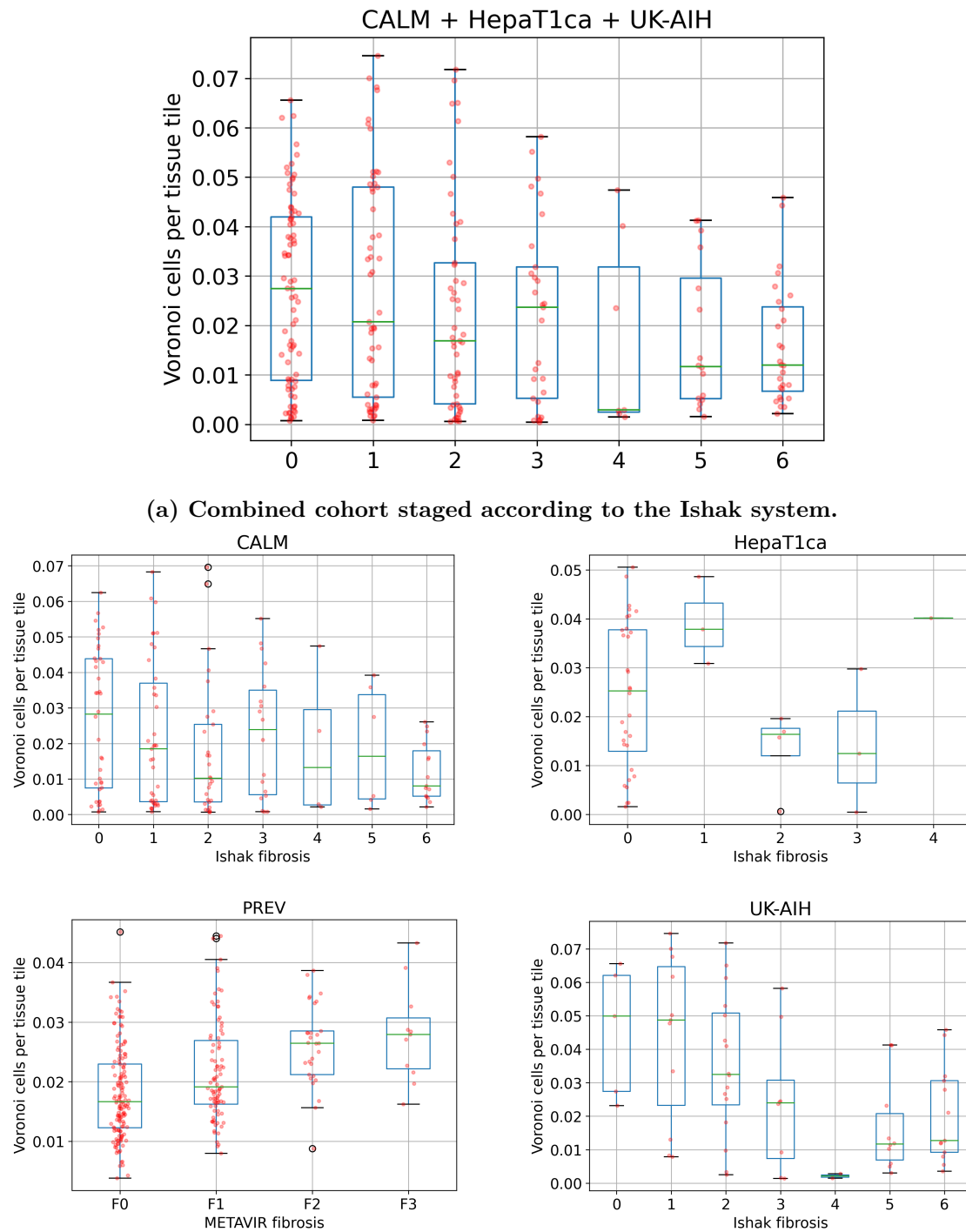


Figure 6.12: Mean fraction of Voronoi cells per tissue tile by stage of fibrosis in the combined study cohort. 6.12a cohorts staged according to the Ishak system. 6.12b individual study cohorts.

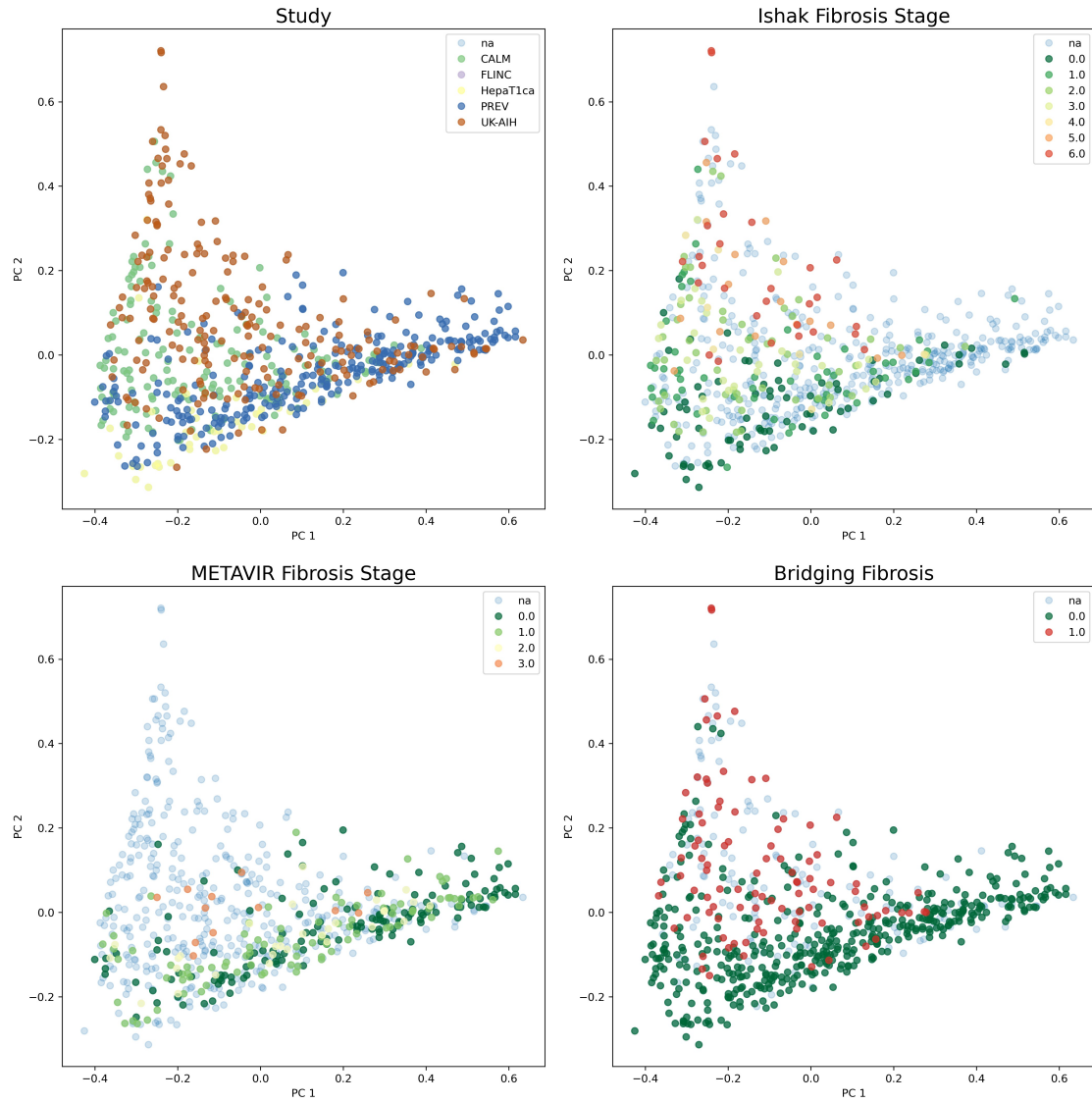


Figure 6.13: The results of principal component analysis of the analysed disease space. The datapoints are coloured according to: study, Ishak fibrosis stage, METAVIR fibrosis stage and the presence of bridging fibrosis in each of the respective plots.

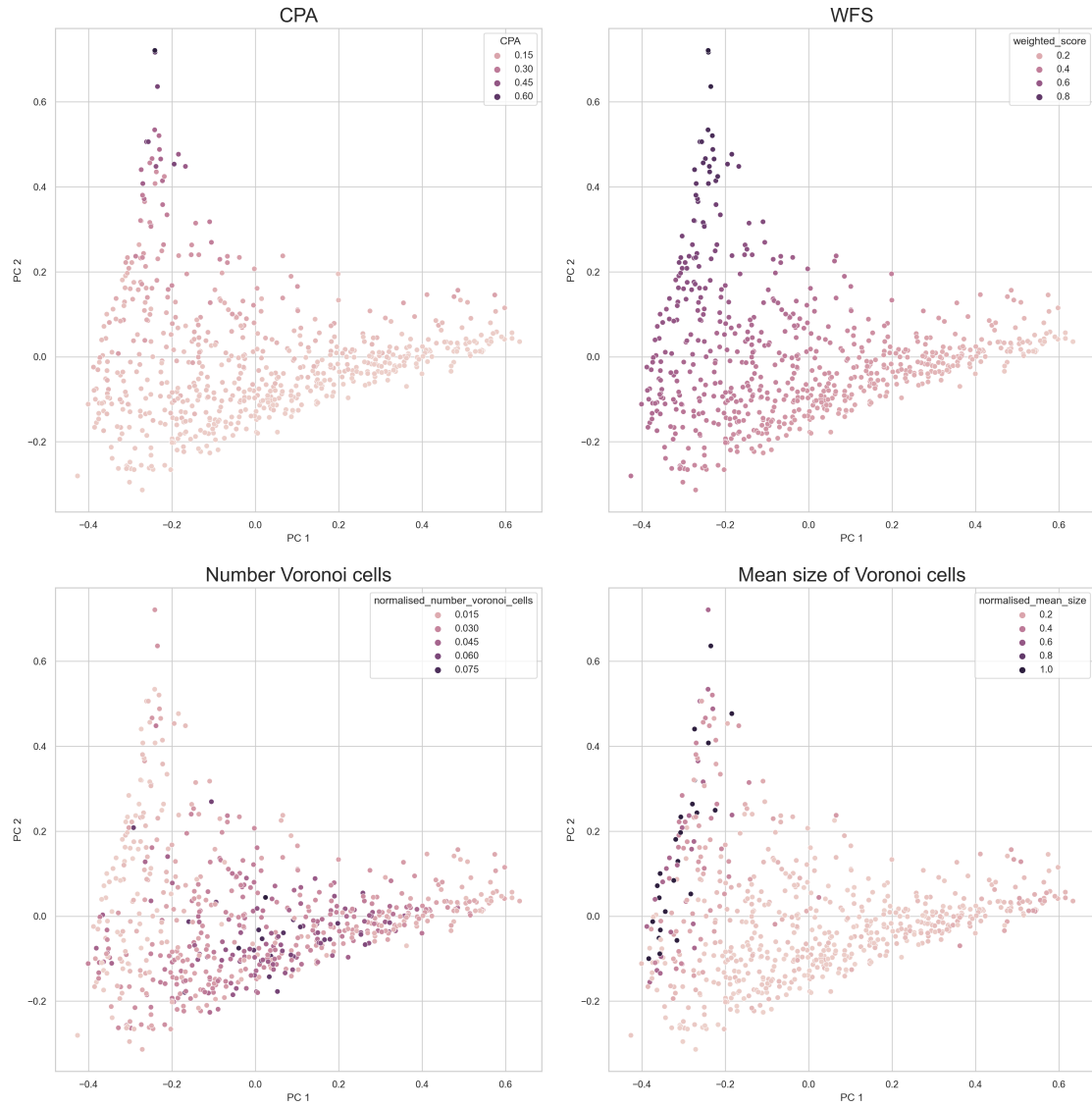


Figure 6.14: The results of principal component analysis of the analysed disease space coloured by the numerical variables: CPA, WFS and Voronoi cell mean number and size.

Next, the distribution of cases with different fibrosis stages is examined, with the different staging systems cast onto bridging/no-bridging fibrosis according to the conversion table from [40]. It can be seen that biopsies with bridging fibrosis occupy the top-centre of the disease space and cases with no bridging cluster along the lower edge of the PCA triangle.

Figure 6.14 shows the obtained distribution of the biopsies along the principal components coloured by the analysed numerical variables. It can be seen that cases with very high CPA are clustered in the top corner of the distribution space. The distribution of WFS values resembles that of the pathologist-assigned fibrosis stages shown in Fig 6.13. Interestingly, cases with a high number of Voronoi cells are gathered in the lower-centre of the distribution. Such cases contain many small disconnected regions of fibrosis which can be interpreted as early fibrous expansions of portal tracts and central veins or regions of pericellular (chicken-wire) fibrosis. An example of such case can be seen in Figure 5.6b. Cases with a high mean size of Voronoi cells are clustered on the left edge on the case distribution. Such cases contain either very few collagen hot-spots or demonstrate such a level of bridging that all the collagen regions in the image are connected into a small number of large septa.

6.5 Tracking disease progression

Figure 6.15 shows two example cases from the UK-AIH cohort for which improvement in Ishak fibrosis stage was observed. Figure 6.16 shows the progression of the respective timepoints along the learned fibrosis distribution.

It can be seen that both cases demonstrate advanced fibrosis, which has reversed upon treatment (decrease in Ishak fibrosis stage has been observed). The change is reflected in both CPA and WFS as well as the change of the collagen subtype histogram within each case. The change in biopsy location within the learned disease distribution reveals a difference in patient trajectory. Patient 1 has progressed towards the lower-right part of the plot and Patient 2 towards the left. The core difference between the two post-treatment biopsy samples is the high number of tiles from cluster 3 relative to cluster 1 in Patient 2. This indicates a higher level

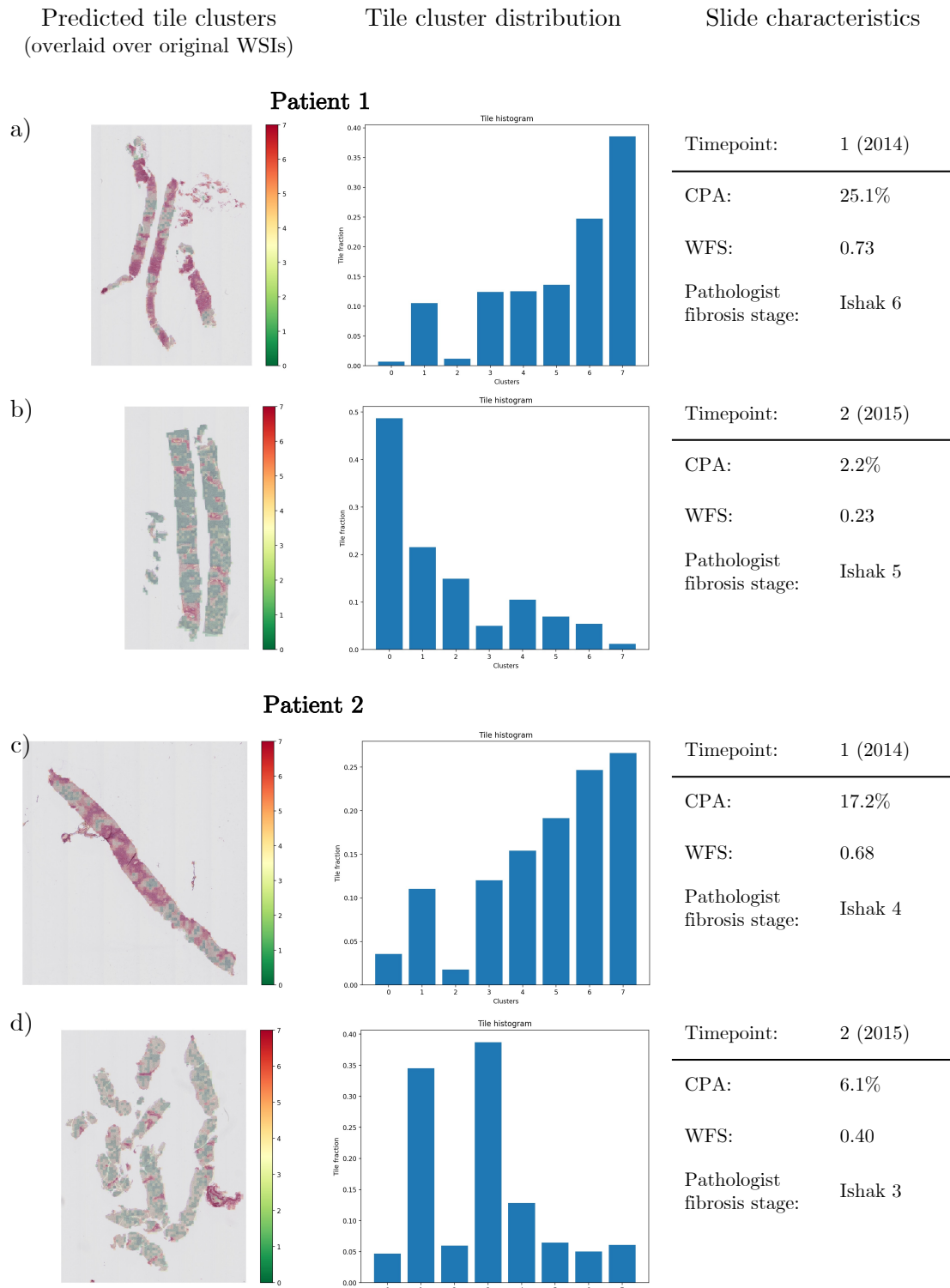


Figure 6.15: Example cases from the UK-AIH cohort for which improvement in Ishak fibrosis stage was observed.

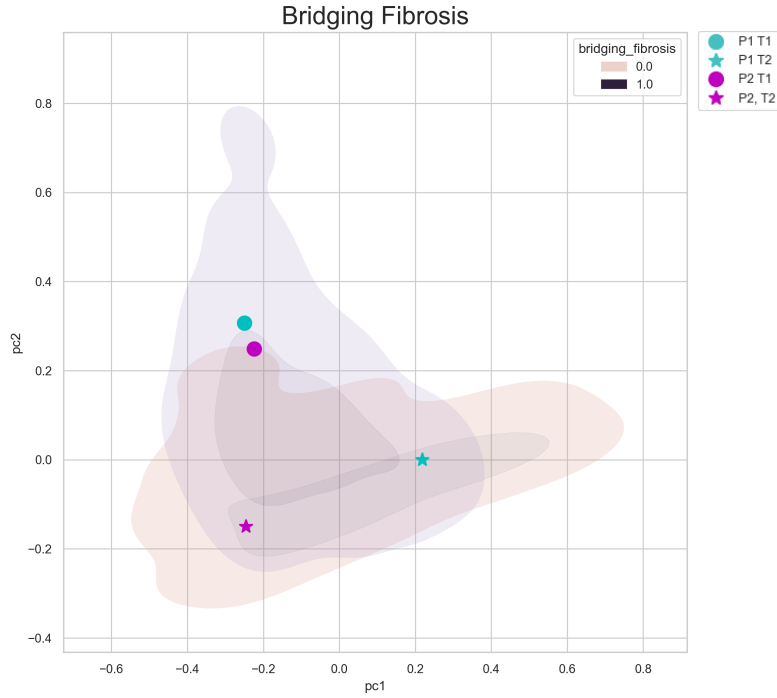


Figure 6.16: Progression of the cases depicted in Figure 6.15 along the learned disease space.

of fine collagen depositions within the parenchyma of this patient’s liver. Further study involving additional markers of disease (i.e. serum marker levels) needs to be conducted in order to test the clinical relevance of the findings in the lower WFS regions of the disease space.

6.6 Conclusions

This chapter presents a novel way of characterising patterns of liver fibrosis in whole-slide images of tissue sections. A new subtyping of local collagen micro-patterns was developed in an unsupervised manner. The learned tissue representation was combined with the previously established imaging biomarker of fibrosis - collagen proportionate area and graph-based descriptors of tissue topology. Together the metrics present a new description of disease state which integrates information obtained at different image levels, ranging from pixels, through local cell neighbourhoods to anatomical tissue regions. We argue that such a representation allows to capture a fuller snapshot of the hierarchical structure of fibrosis than CPA or

categorical fibrosis staging alone. By utilising quantitative imaging parameters such a representation is free the subjectivity of pathological scoring.

By performing the analysis in consecutive steps, we separate the uncertainties coming from different aspects of tissue processing. The results of collagen segmentation can be verified by the pathologist before progressing to the next steps of the pipeline. Similarly, the results of collagen pattern classification and tissue graph matching can also be examined separately.

Finally, we demonstrate that the developed methods can be used to track disease progression in time using examples from the longitudinal sub-cohort of the UK-AIH study. Future work is needed to verify the robustness of such analysis against variations in tissue staining.

7

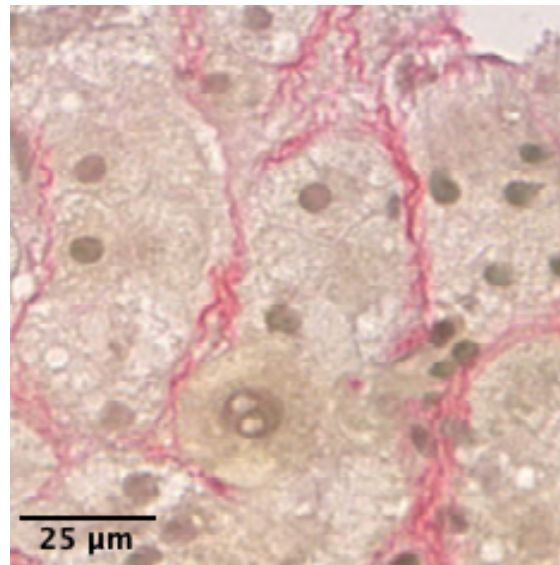
Future work

7.1 FLINC cohort - research collaboration with Novo Nordisk

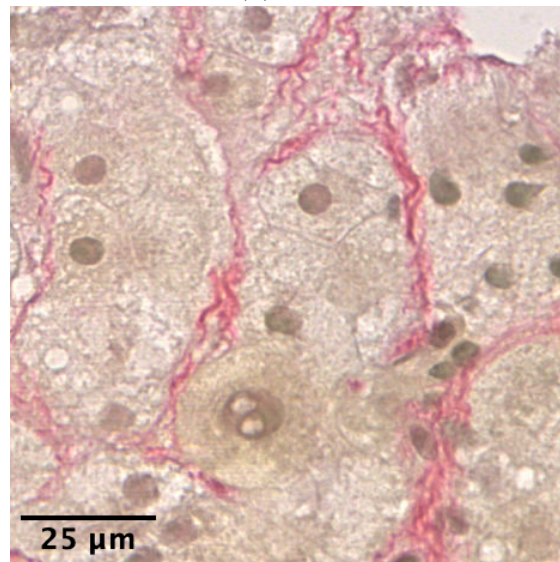
The recent collaboration between the Big Data Institute and Novo Nordisk, Denmark has enabled the analysis of data gathered in two large studies of liver disease: the FLINC cohort and a drug trial. Whole slide images of tissue sections stained with specialised histology stainings for features such as: inflammation, fibrosis and ballooning were made available to our research group. The results of some of the analysis of the fibrosis slides has been already included in this thesis manuscript. The work presented in this thesis is soon to be extended in the format of a post-doctoral research project which will be conducted by me (MW) in the Rittscher group in collaboration with Novo Nordisk.

In this project we will perform the analysis of collagen pattern subtypes in slides images at a higher resolution than those previously studied (see Fig. [7.1](#)). It is anticipated that the examination of collagen fibres at the level of individual cells will enable us to connect the imaging features with local gene expression.

The genetic information for the cohort has been gathered using two methods: bulk RNA sequencing and spatial transcriptomics [\[154\]](#). The combined data will



(a) 20x.



(b) 40x.

Figure 7.1: Example of a small tissue region from a WSI from the FLINC study. A comparison of the features visible at the previously used magnification - 20x and 40x.

be used to conduct a pioneering study into the genetic and molecular drivers of inflammation and fibrosis in MAFLD patients.

Appendices



Additional information concerning the clinical cohorts

Contents

A.1 Example written pathology report from UK-AIH study	117
A.2 Full clinical information overview	119
A.3 Imaging artefacts in the FLINC cohort	123
A.4 PSR staining protocol for the FLINC cohort	125
A.5 Colour comparison between cohorts	128

A.1 Example written pathology report from UK-AIH study

Below is an example of a written pathology report from the UK-AIH study (Fig. [A.1](#)). 237 individual reports were gathered within the study. Patient identifiable information was removed from the reports by persons responsible for curating the data within the UK-AIH consortium before the reports were shared with me.

The UK-AIH report pool contains a multitude of report formats and reporting styles used by pathologists in 20 UK hospitals, in the years between 1998 and 2017. The individual reports were originally written wholly for the purpose of medical examination. Because of this, they do not follow a standardised format

MDM / Teaching Reprints ONLY

Department of Cellular Pathology
The Royal Victoria Infirmary
Newcastle Upon Tyne Trust, NE1 4LP
Tel: 0191-233-6161 Ext 24445 Fax: 0191-282-5892

UKA14 01-0002

BIOPSY REPORT

[REDACTED] m PR002058Q/03
 Clinician: Prof M F Bassendine (Medi DoB: [REDACTED]/1979
 Received: 21/01/2003 Ward: Freeman wd 6 - Admission Ward
 NHS No: [REDACTED] Hosp No: [REDACTED]
 Issued: [REDACTED]

Note: The format of this report is not that of the original report.

Clinical History

Liver biopsy.

Macroscopic Report

Liver biopsy: Four cores of friable tan tissue, the longest measuring 5mm, the shortest 2mm. All embedded in 1A.

Microscopic Report

This comprises a small fragmented specimen. There are areas of fibrosis and these are surrounding nodules of parenchyma. It is difficult however to provide an accurate assessment of the architecture given the fragmented nature of the biopsy. Thus while there is undoubtedly some fibrosis it is not clear whether or not there is an established cirrhosis. Within the areas of fibrous tissue there is a mild ductular reaction and there is a chronic inflammatory cell infiltrate. In one residual portal tract there is an accumulation of ceroid-laden macrophages but there is no definite bile duct disease. Within the parenchyma there are a few foci of liver cell dropout but no evidence of steatosis or steatohepatitis or a significant lobular hepatitis. There is no evidence of alpha-1-antitrypsin deficiency but there is accumulation of copper associated protein. There is no evidence of iron overload.

In summary this is a poor specimen in which there is evidence of fibrosis and some features of chronic cholestasis but it is not possible to provide a more definitive diagnosis. Please note that no clinical information was provided with this request. The case will be discussed further at the forthcoming CPC.

Pathologists:
Prof AD Burt

Figure A.1: An example written pathology report from the UK-AIH study. The content of the report is divided into three major parts: 1) the clinical history of the patient, 2) observations from the macroscopic examination of the tissue and 3) the disease-related observations of the same specimen examined under the microscope, often accompanied by a diagnosis.

characteristic of prospective clinical studies. There isn't a common predefined question which the pathologists were aiming to answer, nor were any specific grading systems used commonly between reports.

While the reports are highly variable in both format and content, they all share the same essential parts. Firstly, they each contain a brief description of the specimen, with comments about its quality (e.g. size, number of tissue fragments). If the quality of the sample is perceived by the pathologist to be sub-optimal, the pathologist may also indicate how certain they were when drawing conclusions from the examination.

The middle part is typically a description of observations made by the pathologist from several tissue slices stained with different stains. In the text, they sometimes refer to specific regions within the slides on which their observations were based. They also sometimes refer to particular slices and stainings when describing morphological features of the tissue. The focus, however, remains on the conclusions from the combined multi-slice and multi-stain examination. Finally, depending on the quality of the sample, the amount of clinical information provided, and the pathologist's own certainty, they may or may not provide a diagnosis for the patient.

Because of the lack of standardisation of reporting, the written reports were not directly used in the project, either for quality assessment or to derive pathology stages. The UK-AIH reports did, however, provide me with valuable insights into the process of clinical tissue examination and the challenges associated with clinical reporting, which I have incorporated into the thesis.

A.2 Full clinical information overview

Below is a summary of the clinical and imaging data available in each of the studies included in this thesis (Tab. [A.1](#)). Table [A.2](#) summarises the blood tests accompanying the clinical data. Table [A.3](#) contains the comparison of pathology grading systems used by pathologists who scored the tissues for MAFLD-related features and subtypes of inflammatory patterns.

Table A.1: A summary table of the clinical information gathered within each of the study cohorts - general and imaging information.

		CALM	PREV	FLINC	HepaT1ca	UK-AIH
Patient information						
	Age at biopsy					
	Age at LMS scan					
	Sex					
	Race					
	Ethnicity					
Diagnoses						
	Pathologist NAFLD diagnosis					
	Pathologist NASH diagnosis					
	AIH					
	Hypertension					
	Type 2 diabetes					
	Cirrhosis					
	Transplant liver					
	Other					
Medication						
	Simvastatin					
	GLP1RA					
	Pioglitazone					
	DPP4I					
	SGLT2					
	Basal insulin					
	Metformin					
	Immunosuppressants					
	Chemotherapy					
BMI						
	Height					
	Weight					
	Waist circumference					
	Hip circumference					
Alcohol						
	Alcohol units per week					
Liver decompensation						
	Oesophageal varices					
	Ascites					
LMS						
	cT1					
	T2*					
	fat					
	MRI report					
Fibroscan						
	LSM					
	CAP					
MRE						
	MRE result					

Table A.2: A summary table of the clinical information gathered within each of the study cohorts - blood tests.

		CALM	PREV	FLINC	HepaT1ca	UK-AIH
Selected blood markers						
	Glucose					
	Creatinine					
	Bilirubin					
	ALT					
	AST					
	ALK p					
	Albumin					
	GGT					
	AFP					
	IgG					
	IgA					
	IgM					
	CRP					
	Transferrin					
	Transferrin saturation					
	Serum Iron					
	Ferritin					
	Total cholesterol					
	Triglyceride					
	HDL					
	LDL					
	vLDL					
	Haemoglobin					
	White cell count					
	Platelets					
	Prothrombin					
	ELF					
	SVAP1					
	APRI					
	NFS					
	FIB4					

Table A.3: A summary table of the clinical information gathered within each of the study cohorts - pathology scoring.

		CALM	PREV	FLINC	HepaT1ca	UK-AIH
Biopsy Info						
	Biopsy date					
	Written report					
Sample adequacy						
	Biopsy adequate					
	Longest core length					
	Total length					
	Number of portal tracts					
	Number of fragments					
Fibrosis						
	Ishak fibrosis (0-6)					
	NAS fibrosis (0-4)					
	METAVIR fibrosis (0-4)					
	CPA (pathologist)					
Inflammation						
	Portal inflammation (0-4)					
	Portal inflammation (0-2)					
	Interface inflammation (0-4)					
	Lobular inflammation (0-3)					
	Lobular focal necrosis (0-4)					
	Confluent necrosis (0-6)					
	Total inflammation (0-18)					
Steatosis						
	Steatosis grade (0-3)					
	Pathologist steatosis percentage					
	Steatosis type					
	Steatosis localisation					
	Microvacuolar steatosis (0-1)					
Clarification/ballooning						
	Ballooning (0-2)					
	Ballooning (0-3)					
	Apoptotic bodies (0-1)					
	Mallory-Denk bodies (0-1)					
NAS components						
	Kleiner-Brunt fibrosis (0-4)					
	Ballooning					
	Lobular inflammation					
	Brunt steatosis					
	NAS					
SAF components						
	SAF Lobular inflammation					
	SAF Ballooning					
	SAF Steatosis					
	SAF Activity Score					
	SAF Fibrosis					
	SAF Total					

A.3 Imaging artefacts in the FLINC cohort

A large number of PSR stained slides from the FLINC cohort was excluded from analysis in this thesis because of image quality issues. The figure below (Fig. [A.2](#)) shows common types of imaging artefacts present in WSIs from the FLINC study.

41 slides contained non-tissue objects in the image. Most commonly, these were the borders of the glass slide and slide labels. 9 slides contained regions of tissue from organs other than liver (most commonly skin) in addition to liver tissue. 7 slides labelled as "PSR" contained no collagen staining. 4 slides were stained with reticulin staining in addition to Sirius Red staining. Reticulin staining is a staining protocol which uses silver molecules to stain reticular fibres black. Reticular fibres are made of type III collagen. Both Sirius Red and silver stainings are used to create differential contrast between collagen and other tissue objects. Slides stained with reticulin staining cannot be directly compared to those stained with Sirius Red, therefore they were rejected from analysis. In total, 61 out of 203 FLINC PSR-stained samples were rejected from analysis in this thesis.

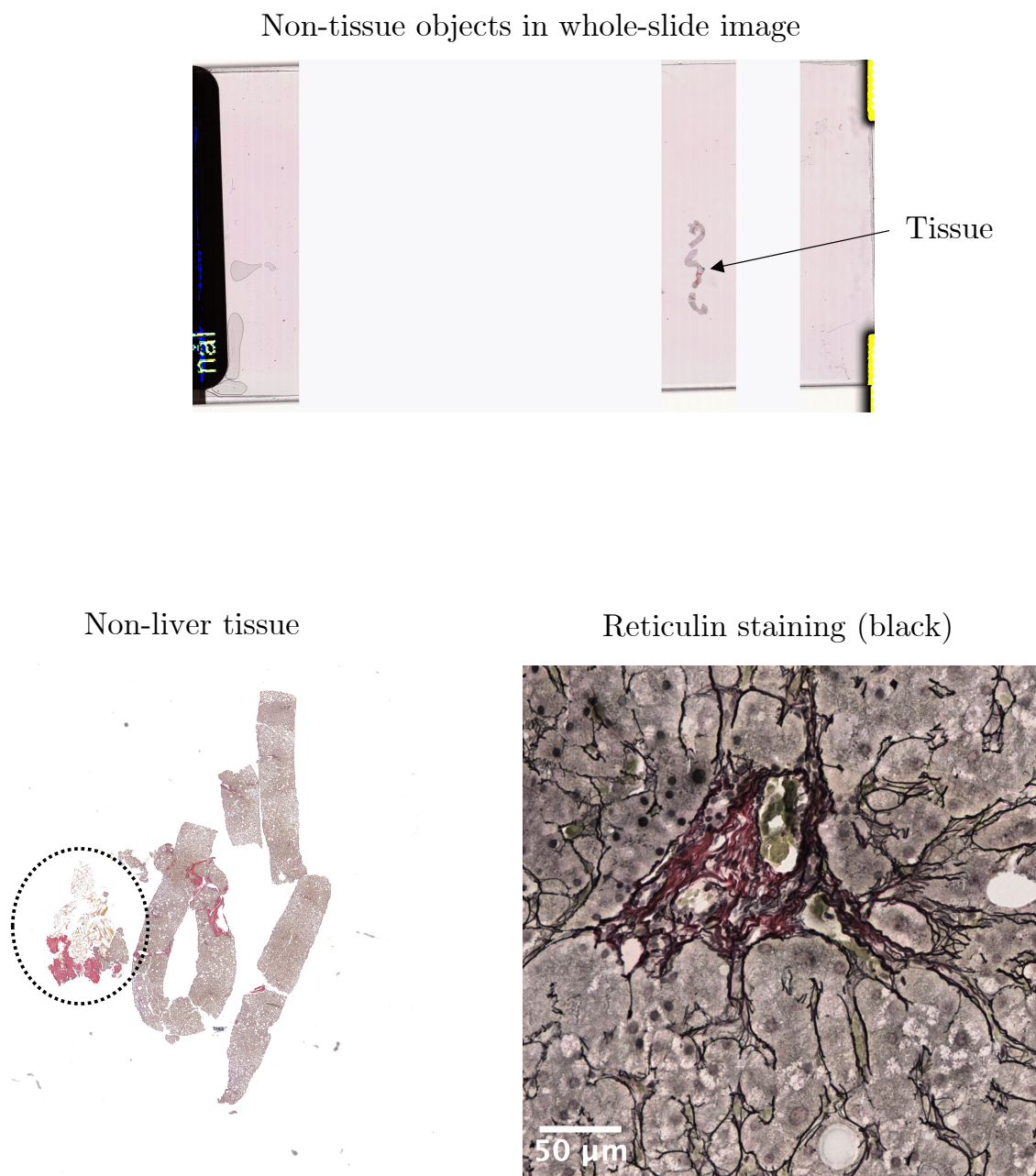


Figure A.2: Examples of imaging artefacts in WSIs from the FLINC cohort.

A.4 PSR staining protocol for the FLINC cohort

Below is the full protocol used for PicroSirius Red staining performed on the FLINC cohort, provided by Copenhagen University Hospital, Hvidovre and Novo Nordisk, Denmark. The staining is internally called "WAS" for the three component stains used in the protocol: Weigert's Haematoxylin, Alcian Blue and PicroSirius Red. Figure [A.3](#) presents the original protocol for manual staining, along with the expected results for the staining. The protocol is written in Danish. The nuclei are to be brown, the cytoplasm - yellow, and collagen (including reticulin) - red. Additionally mucin in bile duct epithelial cells and ground substance of the extracellular matrix are to be light blue. Figure [A.4](#) presents the protocol for equivalent staining using the Tissue-Tek Prisma Plus automated stainer.

Hvidovre Hospital
Patologifdelingen

Side 1 af 2

Was

<u>Fremgangsmåde:</u>	<u>Tid</u>	<u>Udsmidning</u>
1 Afparaffinering, snit til vand		
2 0, 6 % Alcian blue opl. pH 3	5 min	Affaldsgruppe H
3 3 % Eddikesyre	5 – 10 dyp	Affaldsgruppe H
4 Ionbyttet vand x 2		
5 0,3 % Natriumcarbonat	5 min	Kemikalieafløbet
6 Rindende vand		
7 Lillie Weigert 's jærnhematein	8 min	Kemikalieafløbet
8 Rindende vand	5 min	
9 Picrosiriusrød	5 min	Affaldsgruppe O
10 Et hurtigt dyp i rindende vand		
11 Hurtig dehydrering til 99%		Kemikalieafløbet
12 Montering		

Resultater

Kerner:	Brun-sorte
Cytoplasma:	Gult
Kollagen:	Rødt
Reticulin:	Rødt
Mucin i galdegangepithel:	Lyseblåt-blåt
Ungt bindevæv (grundsubstans):	Lyseblåt-blåt

Kontrolmateriale:

Titel: Was	Godkendt den: 03.01.2019
Ansvarlig: csta	Ændret den: 02.01.2019

Figure A.3: PSR staining protocol used in the FLINC cohort. Instructions for manual staining (Danish).

		Was protokol til prisma			
Step	Station	solution name	time	delay	mix
1	S*	Start station	-		-
2	D*	Drying station	00:15:00	**	OFF
3	9	Tissue Clear	00:05:00	**	OFF
4	1	Tissue Clear	00:05:00	**	OFF
5	2	Alcohol 99%	00:05:00	**	OFF
6	10	Alcohol 99%	00:05:00	**	OFF
7	11	Alcohol 96%	00:02:00	**	OFF
8	3	Alcohol 70%	00:02:00	**	OFF
9	W*	Wash Station	00:05:00	**	OFF
10	14	Alcian Blue	00:05:00	50%	ON
11	22	Eddikesyre 3%	00:00:10	**	ON
12	W*	Wash Station	00:01:00	**	OFF
13	29	Natriumcarbonat	00:05:00	**	OFF
14	W*	Wash Station	00:05:00	**	OFF
15	13	Weigert's	00:08:00	50%	ON
16	W*	Wash Station	00:05:00	**	OFF
17	20	Picrosiriusrød	00:05:00	50%	ON
18	W*	Wash Station	00:00:05	=	OFF
19	28	Alcohol 96%	00:00:20	=	OFF
20	19	Alcohol 99%	00:03:00	**	OFF
21	18	Alcohol 99%	00:05:00	**	OFF
22	17	Alcohol 99%	00:05:00	**	OFF
23	E*	End station	-	-	-

Figure A.4: PSR staining protocol used in the FLINC cohort. Instructions for staining using the Prisma automatic tissue stainer.

A.5 Colour comparison between cohorts

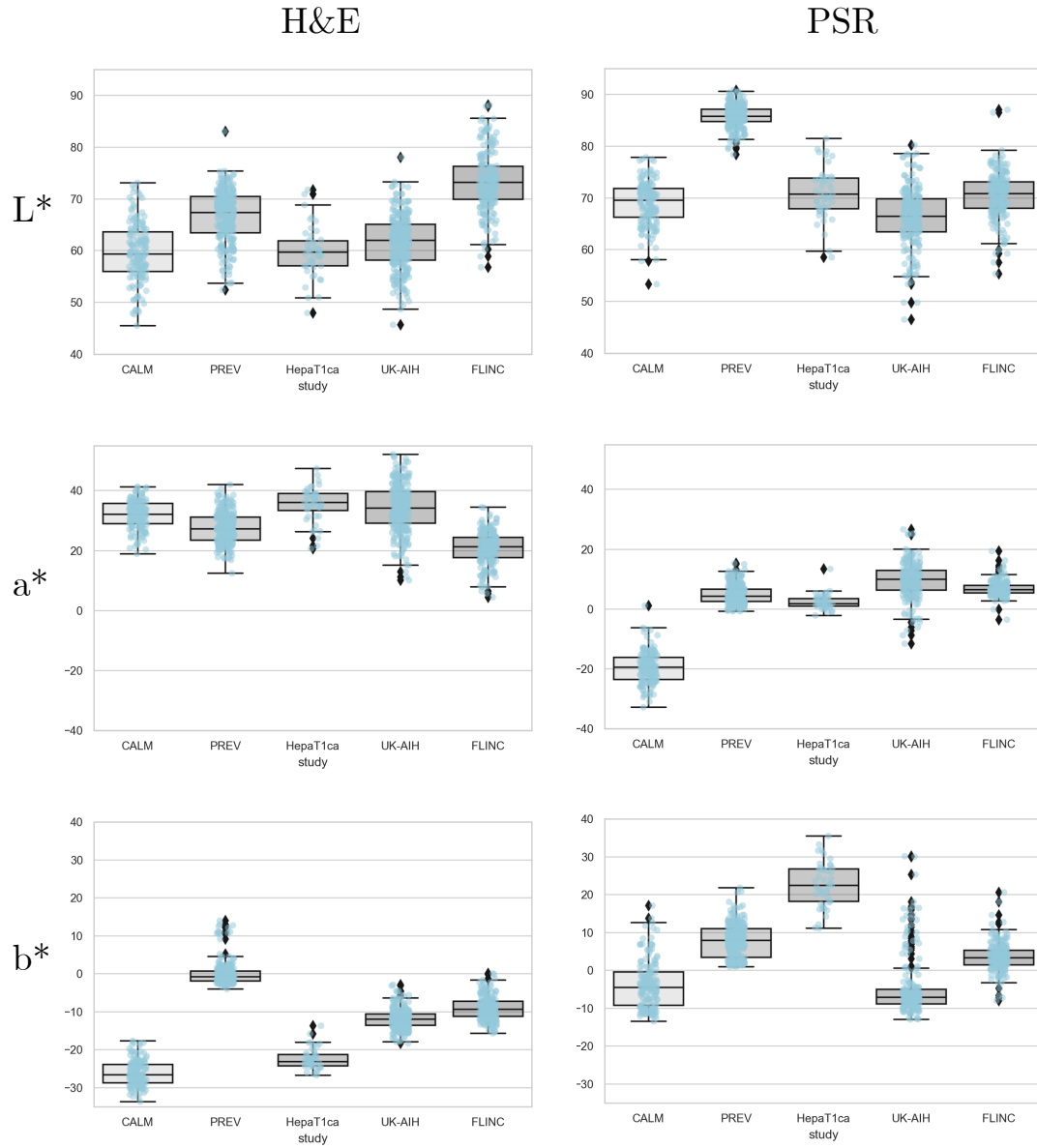


Figure A.5: Distribution of the primary colours of slides from all studies in the CIELAB colourspace.

B

Collagen Identification Supplementary Material

Contents

B.1 Training parameters of the colour transfer pipeline	. . 129
--	----------------

B.1 Training parameters of the colour transfer pipeline

The colour transfer pipeline was trained on the HepaT1ca cohort of whole slide images from surgical liver sections. The cohort contains digitised PSR and H&E slides for 44 patients. 5 cases were removed from the analysis either for slide quality reasons (tears in tissue) or because of regions of metastatic cancer being visible in the liver slice. Overall, slides from 39 patients were used in pipeline training.

The whole slide images of PSR slices were digitally registered to their corresponding H&E slices based on my own manual annotations. The code used for the manual slice registration process was developed within the Rittscher group by Ka Ho Tam.

Following slide registration, pairs of 512 x 512 px tiles showing corresponding regions in both stainings were exported for deep network training and validation.

36204 pairs of tiles were extracted from the pool of 39 analysed cases. The tile cohort was then divided into a training and validation set with a 80:20 ratio.

Overall 28966 and 7238 pairs of tiles were used for network training and testing, respectively.

A CycleGAN model^[1] from the original published paper was used for the analysis [135]. The model was trained with default training parameters until no further improvement in the quality of the generated images was perceived.

¹<https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix>

C

UK-AIH longitudinal study supplementary figures

Contents

C.1 Comparison of mean CPA values by study	131
C.2 Tracking change in collagen metrics over time	131

C.1 Comparison of mean CPA values by study

C.2 Tracking change in collagen metrics over time

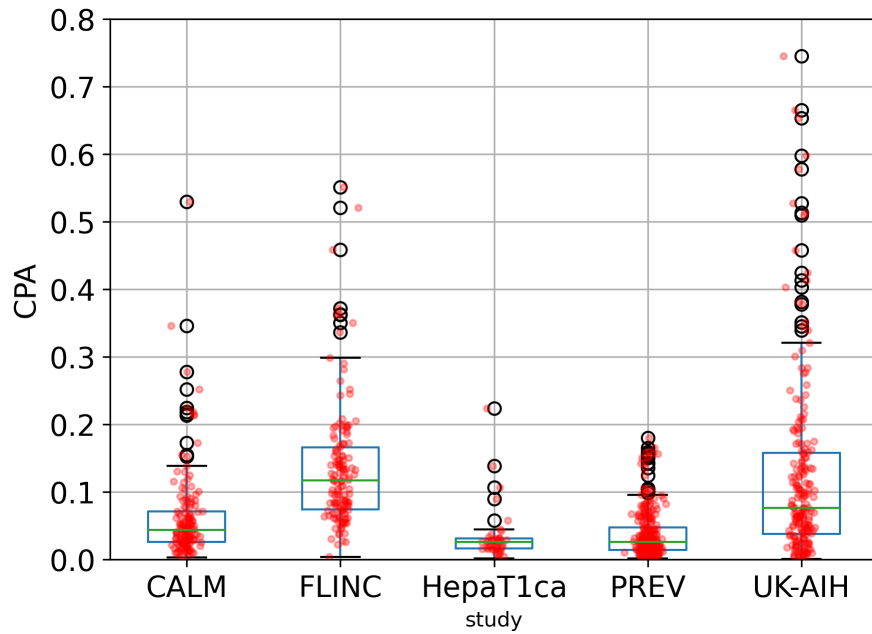


Figure C.1: Mean CPA by study.

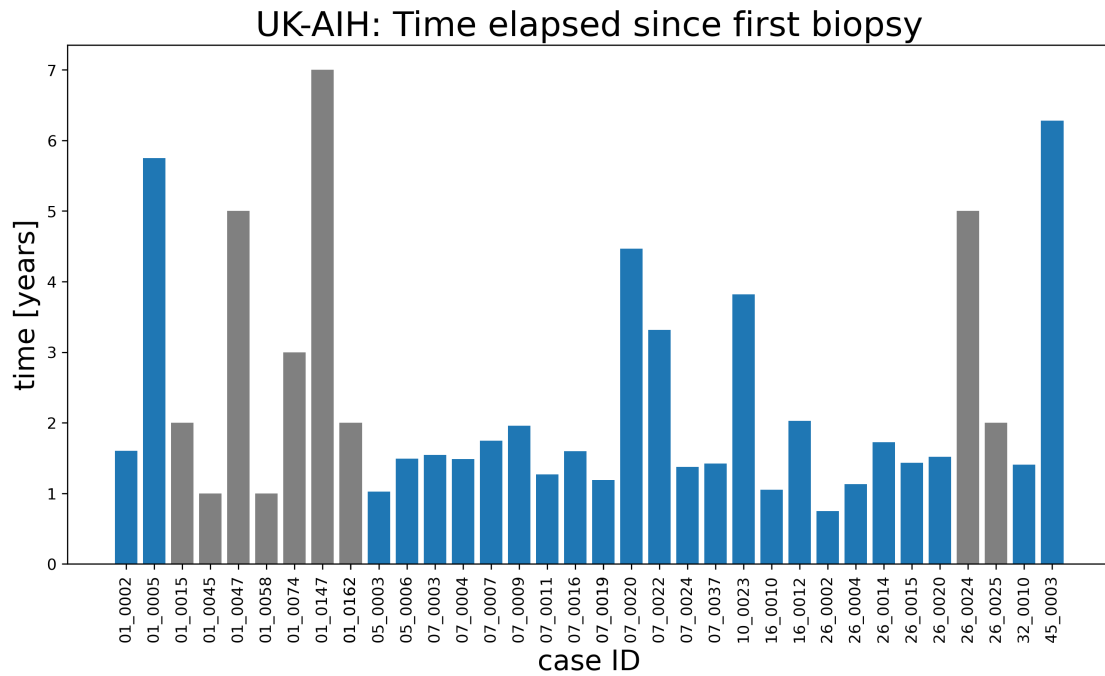


Figure C.2: Time difference between the follow-up and the baseline biopsy in the longitudinal sub-cohort of the UK-AIH study. Blue colour indicates that time was recorded in days. Grey colour indicates that time was recorded in years. The median time difference between time-points was 19 months. The shortest time difference recorded was 9 months. The longest was 7 years.

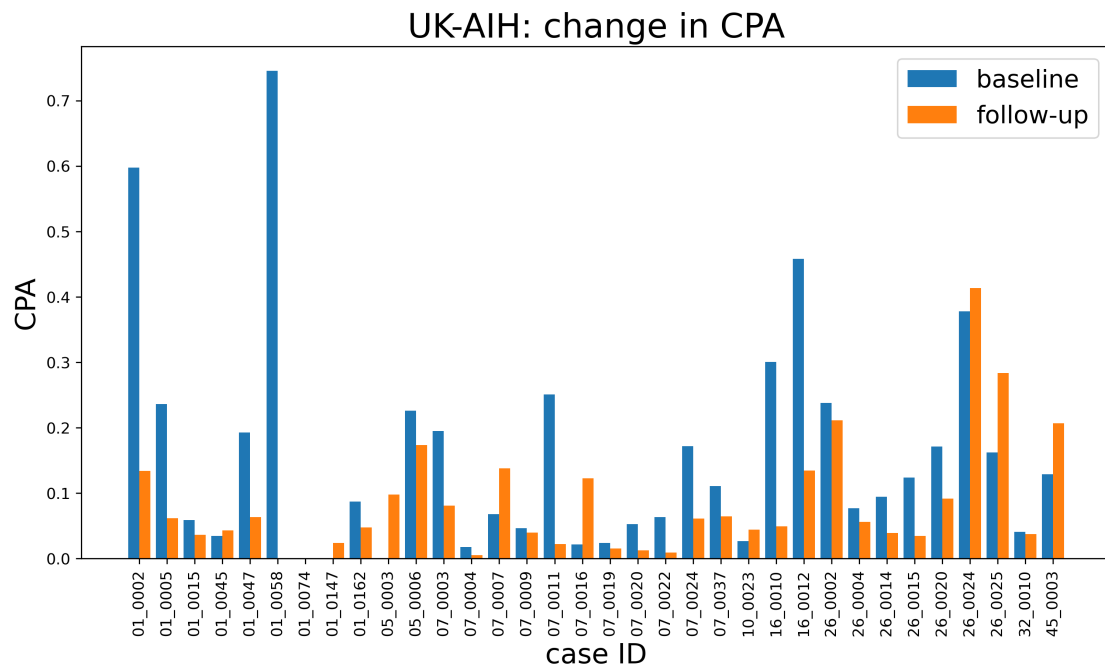


Figure C.3: Change in CPA in the longitudinal sub-cohort of the UK-AIH study. The participants were each following their individual treatments. Empty bars indicate excluded data.

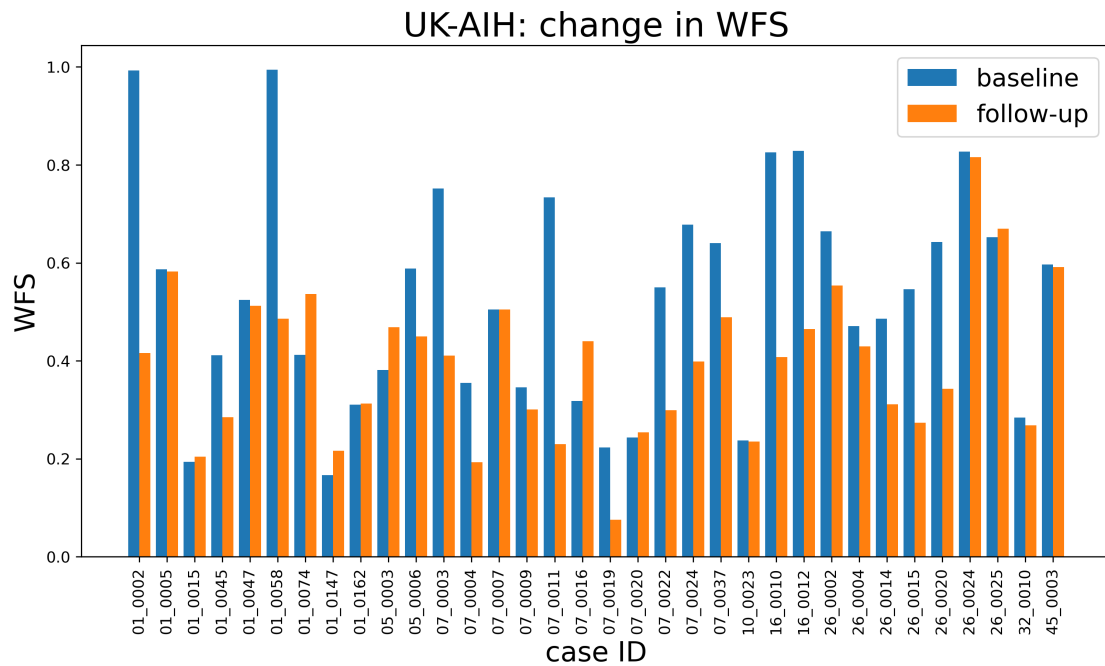


Figure C.4: Change in WFS in the longitudinal sub-cohort of the UK-AIH study. The participants were each following their individual treatments.

References

- [1] Michael Pavlides et al. “Interobserver Variability in Histologic Evaluation of Liver Fibrosis Using Categorical and Quantitative Scores”. In: *American Journal of Clinical Pathology* 147 (2017), pp. 364–369.
- [2] Zobair M. Younossi et al. “Global epidemiology of nonalcoholic fatty liver disease—Meta-analytic assessment of prevalence, incidence, and outcomes”. In: *Hepatology* 64.1 (2016), pp. 73–84.
- [3] Chris Estes et al. “Modeling the epidemic of nonalcoholic fatty liver disease demonstrates an exponential increase in burden of disease”. In: *Hepatology* 67.1 (2018), pp. 123–133.
- [4] David E. Kleiner et al. “Design and validation of a histological scoring system for nonalcoholic fatty liver disease”. In: *Hepatology* 41.6 (2005), pp. 1313–1321.
- [5] Stuart McPherson et al. “Evidence of NAFLD progression from steatosis to fibrosing-steatohepatitis using paired biopsies: Implications for prognosis and clinical management”. In: *Journal of Hepatology* 62.5 (2015), pp. 1148–1155.
- [6] Siddharth Singh et al. “Fibrosis Progression in Nonalcoholic Fatty Liver vs Nonalcoholic Steatohepatitis: A Systematic Review and Meta-analysis of Paired-Biopsy Studies”. In: *Clinical Gastroenterology and Hepatology* 13.4 (2015), pp. 643–654.
- [7] Mohammed Eslam et al. “MAFLD: A Consensus-Driven Proposed Nomenclature for Metabolic Associated Fatty Liver Disease”. In: *Gastroenterology* 158.7 (2020), 1999–2014.e1.
- [8] Yasser Fouad et al. “What’s in a name? Renaming ‘NAFLD’ to ‘MAFLD’”. In: *Liver International* 40.6 (2020), pp. 1254–1261.
- [9] K. George M.M. Alberti, Paul Zimmet, and Jonathan Shaw. “The metabolic syndrome - A new worldwide definition”. In: *Lancet* 366.9491 (2005), pp. 1059–1062.
- [10] Amanda Cheung et al. “Defining Improvement in Nonalcoholic Steatohepatitis for Treatment Trial Endpoints: Recommendations From the Liver Forum”. In: *Hepatology* 70.5 (2019), pp. 1841–1855.
- [11] Detlef Schuppan and Nezam H. Afdhal. “Liver cirrhosis”. In: *The Lancet* 371.9615 (2008), pp. 838–851.
- [12] Julius Balogh et al. “Hepatocellular carcinoma: a review”. In: *Journal of Hepatocellular Carcinoma* Volume 3 (2016), pp. 41–53.
- [13] Scott L. Friedman et al. “Mechanisms of NAFLD development and therapeutic strategies”. In: *Nature Medicine* 24.July (2018), p. 1.

- [14] Giulio Marchesini et al. “EASL-EASD-EASO Clinical Practice Guidelines for the management of non-alcoholic fatty liver disease”. In: *Journal of Hepatology* 64.6 (2016), pp. 1388–1402.
- [15] Bruno Märkl et al. “Number of pathologists in Germany: comparison with European countries, USA, and Canada”. In: *Virchows Archiv* 478.2 (2021), pp. 335–341.
- [16] Victor Mudenda et al. “Addressing the shortage of pathologists in Africa: Creation of a MMed Programme in Pathology in Zambia”. In: *African Journal of Laboratory Medicine* 9.1 (2020), pp. 1–7.
- [17] Lisa Browning et al. “Digital pathology and artificial intelligence will be key to supporting clinical and academic cellular pathology through COVID-19 and future crises: The PathLAKE consortium perspective”. In: *Journal of Clinical Pathology* 74.7 (2021), pp. 443–447.
- [18] Miguel Eugenio Zoubek, Christian Trautwein, and Pavel Strnad. “Reversal of liver fibrosis: From fiction to reality”. In: *Best Practice and Research: Clinical Gastroenterology* 31.2 (2017), pp. 129–141.
- [19] Yue Cheng Guo and Lun Gen Lu. “Antihepatic fibrosis drugs in clinical trials”. In: *Journal of Clinical and Translational Hepatology* 8.3 (2020), pp. 304–312.
- [20] Elena Buzzetti et al. “Collagen proportionate area is an independent predictor of long-term outcome in patients with non-alcoholic fatty liver disease”. In: *Alimentary Pharmacology and Therapeutics* 49.9 (2019), pp. 1214–1222.
- [21] Roberta Forlano et al. “NAFLD: Time to apply quantitation in liver biopsies as endpoints in clinical trials”. In: *Journal of Hepatology* 74.1 (2021), pp. 241–242.
- [22] Ender Gunes Yegin, Korkut Yegin, and Osman Cavit Ozdogan. “Digital image analysis in liver fibrosis : basic requirements and clinical implementation”. In: *Biotechnology & Biotechnological Equipment* 30.4 (2016), pp. 653–660.
- [23] Alexandros Arjmand et al. “Quantification of liver fibrosis-a comparative study”. In: *Applied Sciences (Switzerland)* 10.2 (2020), pp. 1–34.
- [24] Stuart Astbury et al. “Reliable computational quantification of liver fibrosis is compromised by inherent staining variation”. In: *Journal of Pathology: Clinical Research* 7.5 (2021), pp. 471–481.
- [25] Lihui Qin and James M. Crawford. “1 - Anatomy and Cellular Functions of the Liver”. In: *Zakim and Boyer’s Hepatology (Seventh Edition)*. Ed. by Arun J. Sanyal et al. Seventh Edition. Philadelphia: Elsevier, 2018, 2–19.e4.
- [26] American Cancer Society. *A schematic of liver anatomy*. URL: <https://www.cancer.org/cancer/liver-cancer/treating/embolization-therapy.html> (visited on 01/19/2023).
- [27] Pietro Majno et al. “Anatomy of the liver: An outline with three levels of complexity - A further step towards tailored territorial liver resections”. In: *Journal of Hepatology* 60.3 (2014), pp. 654–662.
- [28] Charlotte Debbaut et al. “Analyzing the human liver vascular architecture by combining vascular corrosion casting and micro-CT scanning: A feasibility study”. In: *Journal of Anatomy* 224.4 (2014), pp. 509–517.

- [29] Damian J. Mole et al. “Study protocol: HepaT1ca - An observational clinical cohort study to quantify liver health in surgical candidates for liver malignancies”. In: *BMC Cancer* 18.1 (2018), pp. 1–7.
- [30] *Hepatic structure*. URL: https://en.wikipedia.org/wiki/Liver#/media/File:Hepatic_structure2.svg (visited on 01/19/2023).
- [31] Rory P. Cunningham and Natalie Porat-Shliom. “Liver Zonation – Revisiting Old Questions With New Technologies”. In: *Frontiers in Physiology* 12.September (2021), pp. 1–17.
- [32] Kirstin Meyer et al. “A Predictive 3D Multi-Scale Model of Biliary Fluid Dynamics in the Liver Lobule”. In: *Cell Systems* 4.3 (2017), 277–290.e9.
- [33] Thomas Kietzmann. “Metabolic zonation of the liver: The oxygen gradient revisited”. In: *Redox Biology* 11.January (2017), pp. 622–630.
- [34] Shani Ben-Moshe and Shalev Itzkovitz. “Spatial heterogeneity in the mammalian liver”. In: *Nature Reviews Gastroenterology and Hepatology* 16.7 (2019), pp. 395–410.
- [35] Claire E. McQuitty et al. “Immunomodulatory Role of the Extracellular Matrix Within the Liver Disease Microenvironment”. In: *Frontiers in Immunology* 11.November (2020), pp. 1–33.
- [36] Ryuta Shigefuku et al. “Evaluation of hepatic tissue blood flow using xenon computed tomography with fibrosis progression in nonalcoholic fatty liver disease: Comparison with chronic hepatitis C”. In: *International Journal of Molecular Sciences* 15.1 (2014), pp. 1026–1039.
- [37] J. M. Duffin. “Why does cirrhosis belong to Laennec?” In: *Canadian Medical Association Journal* 137.5 (1987), pp. 393–396.
- [38] Yasuko Iwakiri, Vijay Shah, and Don C. Rockey. “Vascular pathobiology in chronic liver disease and cirrhosis - Current status and future directions”. In: *Journal of Hepatology* 61.4 (2014), pp. 912–924.
- [39] François Durand and Dominique Valla. “Assessment of the prognosis of cirrhosis: Child-Pugh versus MELD”. In: *Journal of Hepatology* 42.SUPPL. 1 (2005), pp. 100–107.
- [40] R. A. Standish et al. “An appraisal of the histopathological assessment of liver fibrosis”. In: *Gut* 55.4 (2006), pp. 569–578.
- [41] Robert G. Knodell et al. “Formulation and application of a numerical scoring system for assessing histological activity in asymptomatic chronic active hepatitis”. In: *Hepatology* 1.5 (1981), pp. 431–435.
- [42] Carolin Lackner et al. “Development and prognostic relevance of a histologic grading and staging system for alcohol-related liver disease”. In: *Journal of Hepatology* 75.4 (2021), pp. 810–819.
- [43] Peter J. Scheuer. “Classification of chronic viral hepatitis: a need for reassessment”. In: *Journal of Hepatology* 13.3 (1991), pp. 372–374.
- [44] K. P. Batts and J. Ludwig. *Chronic hepatitis: An update on terminology and reporting*. 1995.

- [45] Kamal Ishak et al. “Histological grading and staging of chronic hepatitis”. In: *Journal of Hepatology* 22.January (1995), pp. 696–699.
- [46] Pierre Bedossa and Thierry Poynard. “An algorithm for the grading of activity in chronic hepatitis C”. In: *Hepatology* 24.2 (1996), pp. 289–293.
- [47] E Brunt. “Nonalcoholic steatohepatitis: a proposal for grading and staging the histological lesions”. In: *The American Journal of Gastroenterology* 94.9 (1999), pp. 2467–2474.
- [48] Rumiko Kutami et al. “The Laennec grading system for assessment of hepatic fibrosis Validation by correlation with wedged hepatic vein pressure and clinical features”. In: 2000.
- [49] Yameng Sun et al. “New classification of liver biopsy assessment for fibrosis in chronic hepatitis B patients before and after treatment”. In: *Hepatology* 65.5 (2017), pp. 1438–1450.
- [50] Elizabeth M. Brunt and Dina G. Tiniakos. “Histopathology of nonalcoholic fatty liver disease”. In: *World Journal of Gastroenterology* 16.42 (2010), pp. 5286–5296.
- [51] James Neuberger et al. “Guidelines on the use of liver biopsy in clinical practice from the British Society of Gastroenterology, the Royal College of Radiologists and the Royal College of Pathology”. In: *Gut* 69.8 (2020), pp. 1382–1403.
- [52] Pierre Bedossa, Keyur Patel, and Laurent Castera. “Histologic and noninvasive estimates of liver fibrosis”. In: *Clinical Liver Disease* 6.1 (2015), pp. 5–8.
- [53] Rajarshi Banerjee et al. “Multiparametric magnetic resonance for the non-invasive diagnosis of liver disease”. In: *Journal of Hepatology* 60.1 (2014), pp. 69–77.
- [54] Sudhakar K. Venkatesh, Meng Yin, and Richard L. Ehman. “Magnetic resonance elastography of liver: Clinical applications”. In: *Journal of Computer Assisted Tomography* 37.6 (2013), pp. 887–896.
- [55] J. Foucher et al. “Diagnosis of cirrhosis by transient elastography (FibroScan): A prospective study”. In: *Gut* 55.3 (2006), pp. 403–408.
- [56] Rodrigo Escobar Díaz Guerrero et al. “Software tools and platforms in Digital Pathology: a review for clinicians and computer scientists”. In: *Journal of Pathology Informatics* 13.June (2022).
- [57] Feng Liu et al. “qFIBS: A Novel Automated Technique for Quantitative Evaluation of Fibrosis, Inflammation, Ballooning, and Steatosis in Patients With Nonalcoholic Steatohepatitis”. In: *Hepatology* 0.0 (2019), pp. 1–14.
- [58] Kaustav Bera et al. “Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology”. In: *Nature Reviews Clinical Oncology* 16.11 (2019), pp. 703–715.
- [59] Tung Hung Su, Chih Horng Wu, and Jia Horng Kao. “Artificial intelligence in precision medicine in hepatology”. In: *Journal of Gastroenterology and Hepatology (Australia)* 36.3 (2021), pp. 569–580.
- [60] Rish K. Pai et al. “Reliability of histologic assessment for NAFLD and development of an expanded NAFLD activity score”. In: *Hepatology* 76.4 (2022), pp. 1150–1163.

- [61] Nazre Batool. “Detection and spatial analysis of hepatic steatosis in histopathology images using sparse linear models”. In: *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)* (2016), pp. 1–6.
- [62] André Homeyer et al. “Automated quantification of steatosis: Agreement with stereological point counting”. In: *Diagnostic Pathology* 12.1 (2017), pp. 1–10.
- [63] Deepak Sethunath et al. “Automated assessment of steatosis in murine fatty liver”. In: *PLOS ONE* 13.5 (May 2018), pp. 1–16.
- [64] Scott Vanderbeck et al. “Automatic classification of white regions in liver biopsies by supervised machine learning”. In: *Human Pathology* 45.4 (2014), pp. 785–792.
- [65] Maria Tsiplakidou et al. “Automated detection of liver histopathological findings based on biopsy image processing”. In: *Information (Switzerland)* 8.1 (2017).
- [66] George Boon Bee Goh et al. “Quantification of hepatic steatosis in chronic liver disease using novel automated method of second harmonic generation and two-photon excited fluorescence”. In: *Scientific Reports* 9.1 (2019), pp. 1–8.
- [67] Mousumi Roy et al. “Segmentation of Overlapped Steatosis in Whole-Slide Liver Histopathology Microscopy Images”. In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. Vol. 2018-July. 2018, pp. 810–813.
- [68] Nir I. Nativ et al. “Automated image analysis method for detecting and quantifying macrovesicular steatosis in hematoxylin and eosin-stained histology images of human livers”. In: *Liver Transplantation* 20.2 (2014), pp. 228–236.
- [69] Vincent Roullier et al. “Fuzzy algorithms to extract vacuoles of steatosis on liver histological color images”. In: *Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings* 400 (2007), pp. 5575–5578.
- [70] David Marti-Aguado et al. “Digital pathology: accurate technique for quantitative assessment of histological features in metabolic-associated fatty liver disease”. In: *Alimentary Pharmacology and Therapeutics* 53.1 (2021), pp. 160–171.
- [71] Fabian Heinemann, Gerald Birk, and Birgit Stierstorfer. “Deep learning enables pathologist-like scoring of NASH models”. In: *Scientific Reports* 9.1 (2019), pp. 1–10.
- [72] Scott Vanderbeck et al. “Automatic quantification of lobular inflammation and hepatocyte ballooning in nonalcoholic fatty liver disease liver biopsies”. In: *Human Pathology* 46.5 (2015), pp. 767–775.
- [73] Roberta Forlano et al. “High-Throughput, Machine Learning-Based Quantification of Steatosis, Inflammation, Ballooning, and Fibrosis in Biopsies From Patients With Nonalcoholic Fatty Liver Disease”. In: *Clinical Gastroenterology and Hepatology* March (2019).
- [74] Vincenza Calvaruso et al. “Computer-assisted image analysis of liver collagen: Relationship to Ishak scoring and hepatic venous pressure gradient”. In: *Hepatology* 49.4 (2009), pp. 1236–1244.
- [75] Yang Yu et al. “Deep learning enables automated scoring of liver fibrosis stages”. In: *Scientific reports* 8.1 (2018), pp. 1–10.

- [76] Samer Gawrieh et al. “Automated quantification and architectural pattern detection of hepatic fibrosis in NAFLD”. In: *Annals of Diagnostic Pathology* 47 (2020), p. 151518.
- [77] Nicola Dioguardi et al. “Liver fibrosis and tissue architectural change measurement using fractal-rectified metrics and Hurst’s exponent”. In: *World Journal of Gastroenterology* 12.14 (2006), pp. 2187–2194.
- [78] Shuoyu Xu et al. “QFibrosis: A fully-quantitative innovative method incorporating histological features to facilitate accurate fibrosis scoring in animal model and chronic hepatitis B patients”. In: *Journal of Hepatology* 61.2 (2014), pp. 260–269.
- [79] Elizabeth M. Brunt et al. “Complexity of ballooned hepatocyte feature recognition: Defining a training atlas for artificial intelligence-based imaging in NAFLD”. In: *Journal of Hepatology* 76.5 (2022), pp. 1030–1041.
- [80] Andreas Holzinger et al. “Towards the Augmented Pathologist: Challenges of Explainable-AI in Digital Pathology”. In: (2017), pp. 1–34. eprint: [1712.06657](https://arxiv.org/abs/1712.06657).
- [81] Amaro Taylor-Weiner et al. “A Machine Learning Approach Enables Quantitative Measurement of Liver Histology and Disease Monitoring in NASH”. In: *Hepatology* 74.1 (2021), pp. 133–147.
- [82] Nikolai V. Naoumov et al. “Digital pathology with artificial intelligence analyses provides greater insights into treatment-induced fibrosis regression in NASH”. In: *Journal of Hepatology* 77.5 (2022), pp. 1399–1409.
- [83] Wanxin Sun et al. “Nonlinear optical microscopy: use of second harmonic generation and two-photon microscopy for automated quantitative liver fibrosis studies”. In: *Journal of Biomedical Optics* 13.6 (2008), p. 064010.
- [84] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. Ed. by Nassir Navab et al. Vol. 9351. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015, pp. 234–241.
- [85] M. Tarek Shaban et al. “Staingan: Stain style transfer for digital histological images”. In: *Proceedings - International Symposium on Biomedical Imaging* 2019-April (2019), pp. 953–956.
- [86] Guido Colloredo et al. “Impact of liver biopsy size on histological evaluation of chronic viral hepatitis: The smaller the sample, the milder the disease”. In: *Journal of Hepatology* 39.2 (2003), pp. 239–244.
- [87] Marta Wojciechowska et al. *Early Detection of Liver Fibrosis Using Graph Convolutional Networks*. Vol. 12908 LNCS. Springer International Publishing, 2021, pp. 217–226.
- [88] Jérémy Sandrini et al. “Quantification of portal-bridging fibrosis area more accurately reflects fibrosis stage and liver stiffness than whole fibrosis or perisinusoidal fibrosis areas in chronic hepatitis C”. In: *Modern Pathology* 27.7 (2014), pp. 1035–1045.
- [89] Bingqiong Wang et al. “Advanced septa size quantitation determines the evaluation of histological fibrosis outcome in chronic hepatitis B patients”. In: *Modern Pathology* (2018), pp. 1567–1577.

- [90] Wes McKinney. “Data Structures for Statistical Computing in Python”. In: *Proceedings of the 9th Python in Science Conference*. Ed. by Stéfan van der Walt and Jarrod Millman. 2010, pp. 56–61.
- [91] Michael Waskom. “Seaborn: Statistical Data Visualization”. In: *Journal of Open Source Software* 6.60 (2021), p. 3021.
- [92] Natasha Mcdonald et al. “Multiparametric magnetic resonance imaging for quantitation of liver disease : a two-centre cross- sectional observational study”. In: *Scientific Reports* February (2018), pp. 1–10.
- [93] Andrea Dennis et al. “Correlations Between MRI Biomarkers PDFFF and cT1 With Histopathological Features of Non-Alcoholic Steatohepatitis”. In: *Frontiers in Endocrinology* 11.January (2021), pp. 1–10.
- [94] Damian J. Mole et al. “Quantitative magnetic resonance imaging predicts individual future liver performance after liver resection for cancer”. In: *PLoS ONE* 15.12 December (2020), pp. 1–15.
- [95] Stephen A. Harrison et al. “Prospective evaluation of the prevalence of non-alcoholic fatty liver disease and steatohepatitis in a large middle-aged US cohort”. In: *Journal of Hepatology* 75.2 (2021), pp. 284–291.
- [96] Jessica K. Dyson et al. “Inequity of care provision and outcome disparity in autoimmune hepatitis in the United Kingdom”. In: *Alimentary Pharmacology and Therapeutics* 48.9 (2018), pp. 951–960.
- [97] Philippe Lefebvre and Bart Staels. “Hepatic sexual dimorphism — implications for non-alcoholic fatty liver disease”. In: *Nature Reviews Endocrinology* 17.11 (2021), pp. 662–670.
- [98] Valerio Nobili et al. “Comparison of the Phenotype and Approach to Pediatric vs Adult Patients with Nonalcoholic Fatty Liver Disease”. In: *Gastroenterology* 150.8 (2016), pp. 1798–1810.
- [99] Maurizio Bonacini et al. “Racial differences in prevalence and severity of non-alcoholic fatty liver disease”. In: *World Journal of Hepatology* 13.7 (2021), pp. 763–773.
- [100] *Molecular in vitro diagnostic examinations. Specifications for preexamination processes for formalin-fixed and paraffin-embedded (FFPE) tissue. Part 4, In situ detection techniques [electronic resource]*. eng. London, 2021.
- [101] Kelly B. Engel and Helen M. Moore. “Effects of preanalytical variables on the detection of proteins by immunohistochemistry in formalin-fixed, paraffin-embedded tissue”. In: *Archives of Pathology and Laboratory Medicine* 135.5 (2011), pp. 537–543.
- [102] D. Kimberley Molina and Vincent J.M. Dimaio. “Normal organ weights in men: Part II-the brain, lungs, liver, spleen, and kidneys”. In: *American Journal of Forensic Medicine and Pathology* 33.4 (2012), pp. 368–372.
- [103] Don C. Rockey et al. “Liver biopsy”. In: *Hepatology* 49.3 (2009), pp. 1017–1044.
- [104] Haydee Lara et al. “Quantitative Image Analysis for Tissue Biomarker Use: A White Paper from the Digital Pathology Association”. In: *Applied Immunohistochemistry and Molecular Morphology* 29.7 (2021), pp. 479–493.

- [105] Timothy J. Kendall et al. “Hepatic elastin content is predictive of adverse outcome in advanced fibrotic liver disease”. In: *Histopathology* 73.1 (2018), pp. 90–100.
- [106] Holde Puchtler and Faye Sweat. “Histochemical specificity of staining methods for connective tissue fibers: Resorcin-fuchsin and van Gieson’s picro-fuchsin”. In: *Histochemie* 4 (1964), pp. 24–34.
- [107] R. W. Dapson et al. “Certification procedures for sirius red F3B (CI 35780, Direct red 80)”. In: *Biotechnic and Histochemistry* 86.3 (2011), pp. 133–139.
- [108] A. C. Ruifrok and D. A. Johnston. “Quantification of histochemical staining by color deconvolution”. In: *Analytical and Quantitative Cytology and Histology* 23.4 (2001), pp. 291–299.
- [109] Peter Bankhead et al. “QuPath: Open source software for digital pathology image analysis”. In: *Scientific Reports* 7.1 (2017), p. 16878.
- [110] Abhishek Vahadane et al. “Structure-preserved color normalization for histological images”. In: *Proceedings - International Symposium on Biomedical Imaging* 2015-July (2015), pp. 1012–1015.
- [111] Adam Goode et al. “OpenSlide: A vendor-neutral software foundation for digital pathology”. In: *Journal of Pathology Informatics* 4.1 (2013), p. 27.
- [112] Paul Heckbert. “Color Image Quantization for Frame Buffer Display”. In: *Computer Graphics* 16.3 (1982), pp. 157–166.
- [113] Nobuyuki Otsu. “A Threshold Selection Method from Gray-Level Histograms”. In: *IEEE Transactions on Systems, Man, and Cybernetics* C.1 (1979), pp. 62–66.
- [114] Stéfan Van Der Walt et al. “Scikit-image: Image processing in python”. In: *PeerJ* 2014.1 (2014), pp. 1–18.
- [115] Lester J. Layfield et al. “Impact of Consensus Conference Review on Diagnostic Disagreements in the Evaluation of Cervical Biopsy Specimens”. In: *American Journal of Clinical Pathology* 147.5 (2017), pp. 473–476.
- [116] Mads Israelsen et al. “Collagen proportionate area predicts clinical outcomes in patients with alcohol-related liver disease”. In: *Alimentary Pharmacology and Therapeutics* 52.11-12 (2020), pp. 1728–1739.
- [117] Stéphane Bancelin et al. “Determination of collagen fibril size via absolute measurements of second-harmonic generation signals”. In: *Nature Communications* 5 (2014), pp. 1–8.
- [118] Seyed Mohammad Siadat et al. “Measuring collagen fibril diameter with differential interference contrast microscopy”. In: *Journal of Structural Biology* 213.1 (2021), p. 107697.
- [119] Benoit B Mandelbrot. *The fractal geometry of nature*. eng. [Rev. ed.] New York: W.H. Freeman, 1983.
- [120] Cyrill Matenaers et al. “Practicable methods for histological section thickness measurement in quantitative stereological analyses”. In: *PLoS ONE* 13.2 (2018), pp. 1–21.
- [121] Ozan Oktay et al. “Attention U-Net: Learning Where to Look for the Pancreas”. In: Midl (2018). arXiv: [1804.03999](https://arxiv.org/abs/1804.03999).

- [122] Yarín Gal and Zoubin Ghahramani. “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning”. In: *Proceedings of The 33rd International Conference on Machine Learning*. Ed. by Maria Florina Balcan and Kilian Q. Weinberger. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, 2016, pp. 1050–1059.
- [123] Alex Kendall and Yarín Gal. “What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017.
- [124] Radford M Neal. *Bayesian Learning for Neural Networks*. 1996.
- [125] Alex Graves. “Practical Variational Inference for Neural Networks”. In: *Advances in Neural Information Processing Systems*. Ed. by J. Shawe-Taylor et al. Vol. 24. Curran Associates, Inc., 2011.
- [126] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. “Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles”. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS’17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 6405–6416.
- [127] Armen Der Kiureghian and Ove Ditlevsen. “Aleatory or epistemic? Does it matter?”. In: *Structural Safety* 31.2 (2009). Risk Acceptance and Risk Communication, pp. 105–112.
- [128] Yongchan Kwon et al. “Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation”. In: *Computational Statistics and Data Analysis* 142 (2020), p. 106816.
- [129] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. “Deep Ensembles: A Loss Landscape Perspective”. In: (2019), pp. 1–15. arXiv: [1912.02757](https://arxiv.org/abs/1912.02757).
- [130] Marc Macenko et al. “A method for normalizing histology slides for quantitative analysis”. In: *Proceedings - 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2009* (2009), pp. 1107–1110.
- [131] Adnan Mujahid Khan et al. “A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution”. In: *IEEE Transactions on Biomedical Engineering* 61.6 (2014), pp. 1729–1738.
- [132] Peter Haub and Tobias Meckel. “A model based survey of colour deconvolution in diagnostic brightfield microscopy: Error estimation and spectral consideration”. In: *Scientific Reports* 5. February (2015), pp. 1–15.
- [133] Erik Reinhard et al. “Color transfer between images”. In: *IEEE Computer Graphics and Applications* 21.5 (2001), pp. 34–41.
- [134] Andrew Janowczyk, Ajay Basavanahally, and Anant Madabhushi. “Stain Normalization using Sparse AutoEncoders (StaNoSA): Application to digital pathology”. In: *Computerized Medical Imaging and Graphics* 57 (2017), pp. 50–61.
- [135] Jun-yan Zhu et al. “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks”. In: (2017). arXiv: [1703.10593](https://arxiv.org/abs/1703.10593).

- [136] Yair Rivenson et al. “PhaseStain: the digital staining of label-free quantitative phase microscopy images using deep learning”. In: *Light: Science and Applications* 8.1 (2019). arXiv: [1807.07701](#).
- [137] Kevin de Haan et al. “Deep learning-based transformation of H&E stained tissues into special stains”. In: *Nature Communications* 12.1 (2021), pp. 1–13.
- [138] Nicolas Wagner et al. *Federated Stain Normalization for Computational Pathology*. Vol. 1. Springer Nature Switzerland, 2022, pp. 14–23.
- [139] Zachary D Goodman et al. “Progression of fibrosis in advanced chronic hepatitis C: evaluation by morphometric image analysis”. In: *Hepatology* 45.4 (2007), pp. 886–894.
- [140] Zachary D. Goodman. “Grading and staging systems for inflammation and fibrosis in chronic liver diseases”. In: *Journal of Hepatology* 47.4 (2007), pp. 598–607.
- [141] Harshita Sharma et al. “A review of graph-based methods for image analysis in digital histopathology”. In: *Diagnostic Pathology* 1.1 (2015), pp. 1–51.
- [142] Yushan Zheng et al. “Encoding histopathological WSIs using GNN for scalable diagnostically relevant regions retrieval”. In: *Lecture Notes in Computer Science* 11764 LNCS (2019), pp. 550–558.
- [143] Yanning Zhou et al. “CGC-net: Cell graph convolutional network for grading of colorectal cancer histology images”. In: *Proceedings - 2019 International Conference on Computer Vision Workshop, ICCVW 2019* (2019), pp. 388–398. arXiv: [1909.01068](#).
- [144] Pushpak Pati et al. “HACT-Net: A Hierarchical Cell-to-Tissue Graph Neural Network for Histopathological Image Classification”. In: *Lecture Notes in Computer Science* 12443 LNCS (2020), pp. 208–219. eprint: [2007.00584](#).
- [145] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. “Exploring Network Structure, Dynamics, and Function using NetworkX”. In: *Proceedings of the 7th Python in Science Conference*. Ed. by Gaël Varoquaux, Travis Vaught, and Jarrod Millman. Pasadena, CA USA, 2008, pp. 11–15.
- [146] Thomas N. Kipf and Max Welling. “Semi-Supervised Classification with Graph Convolutional Networks”. In: *International Conference on Learning Representations (ICLR)* (2017).
- [147] Keyulu Xu et al. “How powerful are graph neural networks?” In: *7th International Conference on Learning Representations, ICLR 2019* (2019), pp. 1–17. arXiv: [1810.00826](#).
- [148] Petar Veličković et al. “Graph attention networks”. In: *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings* (2018), pp. 1–12. arXiv: [1710.10903](#).
- [149] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2016-December (2016), pp. 770–778. arXiv: [1512.03385](#).
- [150] Adam Paszke et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems* 32. Curran Associates, Inc., 2019, pp. 8024–8035.

- [151] Matthias Fey and Jan Eric Lenssen. *Fast Graph Representation Learning with PyTorch Geometric*. 2019. arXiv: [1903.02428 \[cs.LG\]](#).
- [152] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *3rd International Conference for Learning Representations, San Diego, 2015* (2017). arXiv: [1412.6980 \[cs.LG\]](#).
- [153] Hosuk Ryou et al. “Continuous Indexing of Fibrosis (CIF): Improving the Assessment and Classification of MPN Patients”. In: (2022), pp. 1–31.
- [154] Zhenqin Wu et al. “Graph deep learning for the characterization of tumour microenvironments from spatial protein profiles in tissue specimens”. In: *Nature Biomedical Engineering* 6.December (2022).